



Dielectric Phenomena in Solids

Kwan Chi Kao

DIELECTRIC PHENOMENA IN SOLIDS

With Emphasis on Physical Concepts of Electronic Processes

DIELECTRIC PHENOMENA IN SOLIDS

With Emphasis on Physical Concepts of
Electronic Processes

Kwan Chi Kao

Professor Emeritus of Electrical and
Computer Engineering University of Manitoba



ELSEVIER
ACADEMIC
PRESS

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Elsevier Academic Press
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
84 Theobald's Road, London WC1X 8RR, UK

This book is printed on acid-free paper. 

Copyright 2004, Elsevier, Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone: (+44) 1865 843830, fax: (+44) 1865 853333, e-mail: permissions@elsevier.com.uk. You may also complete your request on-line via the Elsevier Science homepage (<http://elsevier.com>), by selecting "Customer Support" and then "Obtaining Permissions."

Library of Congress Cataloging-in-Publication Data

Kao, Kwan Chi.

Dielectric phenomena in solids: with emphasis on physical concepts of electronic processes / Kwan Chi Kao.

p. cm.

Includes bibliographical references and index.

ISBN 0-12-396561-6 (alk. paper)

1. Dielectrics. I. Title.

QC585.K36 2004

537'.24—dc22

2003070901

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN: 0-12-396561-6

For all information on all Academic Press publications
visit our website at www.academicpress.com

Printed in the United States of America

04 05 06 07 08 9 8 7 6 5 4 3 2 1

Dedicated to my aunt, Hui Jien, my sister, Chih Hai, and my daughter, Hung Jien, whose love has contributed so much to my strength and faith.

Contents

Preface	xiv	1.4.7 The Unit of Debye	38
		1.4.8 The Chemical Unit of Mole	38
1 INTRODUCTION	1	References	38
1.1 Maxwell's Equations	2	2 ELECTRIC POLARIZATION AND RELAXATION	41
1.1.1 Ampere's Law	3	2.1 Fundamental Concepts	41
1.1.2 Faraday's Law	5	2.1.1 Electric Charge Carriers and Their Motion	41
1.1.3 Inseparable Magnetic Poles	8	2.1.2 Electromechanical Effects	44
1.1.4 Gauss's Law	9	The Force Acting on the Boundary between Two Different Dielectric Materials	47
1.2 Magnetization	9	The Force Acting on Conductor Surfaces	47
The Orbital Motion of Electrons around the Nucleus of an Atom	11	The Force Elongating A Bubble or A Globule in A Dielectric Fluid	48
The Rotation of Electrons about Their Own Axis in an Atom	12	The Dielectrophoretic Force	49
The Rotation of Protons and Neutrons inside the Nucleus	12	The Electrostriction Force	49
1.2.1 Diamagnetism	13	The Torque Orientating A Solid Body	50
1.2.2 Paramagnetism	15	2.1.3 Electrostatic Induction	51
1.2.3 Ferromagnetism	18	2.2 Electric Polarization and Relaxation in Static Electric Fields	52
1.2.4 Magnetostriction	23	2.2.1 Vacuum Space	53
1.2.5 Magnetic Materials	23	2.2.2 Conducting Materials	53
1.2.6 Magnetic Resonance	26	2.2.3 Dielectric Materials	54
Nuclear Magnetic Resonance (NMR)	27	2.3 The Mechanisms of Electric Polarization	58
Electron Paramagnetic Resonance (EPR)	28	2.3.1 Electronic Polarization (Also Called Optical Polarization)	59
Electron Spin Resonance (ESR)	30	Classical Approach	59
1.2.7 Permanent Magnets	30	Quantum Mechanical Approach	64
1.2.8 Magnetic Memories	30	2.3.2 Atomic or Ionic Polarization (Vibrational Polarization)	65
1.3 Electromagnetic Waves and Fields	32	2.3.3 Orientational Polarization	67
1.4 Dimensions and Units	34		
1.4.1 Length	34		
1.4.2 Mass	35		
1.4.3 Time	35		
1.4.4 Electric Charge	35		
1.4.5 Derived Units	35		
1.4.6 Cgs System of Units and Cgs/SI Conversion	36		

2.3.4	Spontaneous Polarization	74	2.6.6	The Cole–Cole Plot	95
2.3.5	Space Charge Polarization	75	2.6.7	Temperature Dependence of Complex Permittivity	97
	Hopping Polarization	75	2.6.8	Field Dependence of Complex Permittivity	100
	Interfacial Polarization	77		Semiconducting Materials	100
2.4	Classification of Dielectric Materials	78		Ferroelectric Materials	101
2.4.1	Nonferroelectric Materials (Normal Dielectric or Paraelectric Materials)	78		Insulating Materials	102
	Nonpolar Materials	78	2.7	Dielectric Relaxation Phenomena	105
	Polar Materials	78	2.7.1	The Hamon Approximation	105
	Dipolar Materials	78	2.7.2	Distribution of Relaxation Times	108
2.4.2	Ferroelectric Materials	79	2.7.3	The Relation between Dielectric Relaxation and Chemical Structure	110
2.5	Internal Fields	79	References		112
2.5.1	Local Fields for Nondipolar Materials	79	3	OPTICAL AND ELECTRO-OPTIC PROCESSES	115
	The Clausius–Mosotti Equation	80	3.1	Nature of Light	115
	The Lorentz–Lorenz Equation	80	3.1.1	Corpuscular Theory	115
2.5.2	The Reaction Field for Dipolar Materials	81		The Outstanding Difference between Photons and Electrons or Protons	115
	The Debye Equation	82		Frequency, Energy and Momentum of Photons	115
	The Onsager Equation	83	3.1.2	Wave Theory	116
	Statistical-Mechanical Approaches and the Kirkwood Equation	84		Principle of Superposition	117
	The Frohlich Equation	85		Interference and Interferometry	118
2.6	Electric Polarization and Relaxation in Time-Varying Electric Fields	86		Diffraction	121
2.6.1	The Time-Domain Approach and the Frequency-Domain Approach	86		Polarization, Reflection and Refraction	123
2.6.2	Complex Permittivity	86	3.2	Modulation of Light	128
2.6.3	Time Dependent Electric Polarization	87	3.2.1	Double Refraction and Birefringence	128
	Case A	89		Quarter-Wave Plate (QWP)	130
	Case B	89		Optical Activity	131
2.6.4	Kramers–Kronig Equations	91	3.2.2	Electro-Optic Effects	131
2.6.5	Debye Equations, Absorption, and Dispersion for Dynamic Polarizations	92		The Pockels Effect or Linear Electro-Optic Effect	131
	The Effects of the Local Field	94		The Kerr Effect or Quadratic Electro-Optic Effect	133
	The Effects of DC Conductivity	94	3.2.3	The Photorefractive Effect	138
			3.2.4	The Magneto-Optic Effect	142
				The Faraday Effect	142
				The Voigt Effect	142

3.2.5	The Acousto-Optic Effect	143	3.5.2	Photoemission from Crystalline Solids	185
3.3	Interaction between Radiation and Matter	144		Effects of Surface Conditions	188
3.3.1	Generation of Radiation	144		Effects of Defects in Crystals	188
	Thermal Radiation	144		Energy Distribution of Emitted Electrons	189
	Luminescent Radiation	145		Multiquantum Processes	189
	Units of Light	145	3.6	Photovoltaic Effects	191
3.3.2	The Franck–Condon Principle	146	3.6.1	Bulk Photovoltaic Effects	191
3.3.3	Radiative and Nonradiative Transition Processes	146		The Dember Effect	192
	Multiple-Phonon Transition	150		The Photoelectro-Magnetic (PEM) Effect	194
	Auger Recombination Processes	150	3.6.2	Contact Potential	
	Nonradiative Transition due to Defects	152		Photovoltaic Effects	194
	Nonradiative Transition in Indirect Bandgap Materials	152		Schottky Barrier Photovoltages	194
3.3.4	Absorption and Dispersion	154		MIS Solar Cells	196
3.3.5	The Franz–Keldysh Effect	157		Photovoltaic Behavior of P-N Junctions	196
3.3.6	Formation and Behavior of Excitons	158		PIN Structures for Amorphous Si-based Photovoltaic Devices	201
	Exciton Transport Processes	161	3.6.3	Photosynthesis Photovoltaic Effects	203
	Exciton Interactions	162	3.6.4	Anomalous Photovoltaic Effects	206
3.4	Luminescence	164	References		207
3.4.1	Photoluminescence	165			
	Fluorescence	165	4 FERROELECTRICS, PIEZOELECTRICS, AND PYROELECTRICS		213
	Phosphorescence and Phosphors	169	4.1	Introductory Remarks	213
3.4.2	Electroluminescence	171	4.2	Ferroelectric Phenomena	216
	Classical Electroluminescence	171	4.2.1	General Features	216
	Injection Electroluminescence	173	4.2.2	Phenomenological Properties and Mechanisms	221
3.5	Photoemission	181		BaTiO ₃ -Type Ferroelectrics	221
3.5.1	Photoemission from Electrical Contacts	181		KH ₂ PO ₄ -Type Ferroelectrics	227
	From a Metal into Wide Bandwidth Semiconductors	182		KNaC ₄ H ₄ O ₆ • 4H ₂ O (Rochelle Salt)-Type Ferroelectrics	228
	From a Narrow Bandwidth Emitter into Wide Bandwidth Semiconductors	184		(NH ₂ CH ₂ COOH) ₃ • H ₂ SO ₄ -Type Ferroelectrics	229
	From a Metal into Narrow Bandwidth Crystals	184		Alloys of PbO, ZrO ₂ , and TiO ₂ (PZT alloys)-Type Ferroelectric Ceramics	230

	PVDF $[(\text{CH}_2\text{-CF}_2)_n]$ -Type Ferroelectric Polymers	234	5 ELECTRETS	283
4.2.3	Thermodynamic Theory	236	5.1	Introductory Remarks
	Ferroelectric Transition	238	5.2	Formation of Electrets
	Antiferroelectric Transition	241	5.2.1	Thermo-Electrical Method
4.2.4	Formation and Dynamics of Domains	242	5.2.2	Liquid-Contact Method
4.2.5	Ferroelectric Materials	246	5.2.3	Corona Discharge Method
4.2.6	Applications of Ferroelectrics	249	5.2.4	Electron-Beam Method
	Capacitors	249	5.2.5	Electromagnetic Radiation Method
	Thermo-Autostabilization Nonlinear Dielectric Elements (TANDEL)	252		290
	High-Energy Electrical Pulse Generators	252	5.3	Charges, Electric Fields and Currents in Electrets
	Memories	255		Case I: The Electret Has Only Surface Charges
4.3	Piezoelectric Phenomena	257		Case II: The Electret Has Both Surface Charges $\sigma(d)$ and Volume Charges $\rho(x)$
4.3.1	Phenomenological Approach to Piezoelectric Effects	257		293
4.3.2	Piezoelectric Parameters and Their Measurements	262	5.4	Measurements of Total Surface Charge Density and Total Charges
4.3.3	Piezoelectric Materials	264	5.4.1	Total Surface Charge Density
4.3.4	Applications of Piezoelectrics	266		Compensation Method
	Gas Igniters	266		Capacitive Probe Method
	Delay Lines	267	5.4.2	Total Charges per Unit Planar Area Q_T
	Piezoelectric Positioners and Actuators	267		Electrostatic Induction or Faraday Pail Method
	Piezoelectric Transformers	268		Thermal Pulse Method
4.4	Pyroelectric Phenomena	269	5.5	Charge Storage Involving Dipolar Charges
4.4.1	Phenomenological Approach to Pyroelectric Effects	270	5.5.1	Basic Poling Processes
4.4.2	Pyroelectric Parameters and Their Measurements	271	5.5.2	Relaxation Times of Dipoles and the Thermally Stimulated Discharge Currents (TSDC) Technique
4.4.3	Pyroelectric and Thermally Sensitive Materials	272		298
	NTC Materials	273	5.5.3	Spatial Distribution of Dipolar Polarization
	PTC Materials	274	5.5.4	Isothermal Polarization Decay Processes
4.4.4	Applications of Pyroelectrics	275		302
	Pyroelectric Radiation Detectors	275	5.6	Charge Storage Involving Real Charges
	Pyroelectric Burglar Alarm Systems	278	5.6.1	Spatial Distributions of Trapped Real Charges
	Pyroelectric Thermometry	278		Sectioning Method
	Pyroelectric Energy Conversion	279		Electron Beam Sampling Method
References		279		Pulsed Electro-Acoustic Method
				305

Other Methods	308	Correction due to Drift and Diffusion of Carriers in the Depletion Region	350
5.6.2 Energy Distribution of Trapped Real Charges	308	Effects of Phonon Scattering and Quantum- Mechanical Reflection	352
5.6.3 Isothermal Real Charge Decay Processes	310	Other Factors	352
Stage 1: For $t \leq t_0$	311	6.2.3 Field Emission	354
Stage 2: for $t \geq t_0$	311	Without the Effects of Defects in Solids	354
5.6.4 Distinction between Dipolar Charges and Real Charges	312	With the Effects of Defects in Solids	358
5.7 Basic Effects of Electrets	313	6.2.4 Thermionic-Field Emission	363
5.8 Materials for Electrets	316	6.3 Tunneling through Thin Dielectric Films between Electrical Contacts	364
5.8.1 General Remarks	316	6.3.1 Analysis Based on a Generalized Potential Barrier	364
5.8.2 Physical Properties	316	For $T = 0$	366
5.9 Applications of Electrets	321	For $T > 0$	366
5.9.1 Electret Microphones	321	6.3.2 Elastic and Inelastic Tunneling	367
5.9.2 Electromechanical Transducers	322	6.3.3 Tunneling Through a Metal-Thin Insulation Film-Semiconductor (MIS) System	371
5.9.3 Pyroelectric Detectors	322	6.3.4 Effects of Space Charges and Traps on Tunneling Efficiency and Impurity Conduction	374
5.9.4 Other Applications	323	6.4 Charge Transfer at the Metal-Polymer Interface	376
References	323	References	378
6 CHARGE CARRIER INJECTION FROM ELECTRICAL CONTACTS	327	7 ELECTRICAL CONDUCTION AND PHOTOCONDUCTION	381
6.1 Concepts of Electrical Contacts and Potential Barriers	327	PART I: ELECTRICAL CONDUCTION	381
6.1.1 Electrical Contacts, Work Functions, and Contact Potentials	328	7.1 Introductory Remarks	381
Contact Potential Measurements	330	7.2 Ionic Conduction	382
Photoelectric Emission Method	330	7.2.1 Intrinsic Ionic Conduction	383
Thermionic Emission Method	330	7.2.2 Extrinsic Ionic Conduction	385
6.1.2 Types of Electrical Contacts	334	7.2.3 Effects of Ionic Conduction	385
Neutral Contacts	334	7.3 Electronic Conduction	387
Blocking Contacts	336	7.3.1 Electrical Transport	389
Ohmic Contacts	336	Band Conduction	389
6.1.3 Surface States	341	Defect-Controlled Conduction	398
6.2 Charge Carrier Injection through Potential Barriers from Contacts	345	Electrical Transport by a Tunneling Process	399
6.2.1 Potential Barrier Height and the Schottky Effect	345		
Neutral Contacts	345		
Blocking Contacts	347		
6.2.2 Thermionic Emission	350		
Effect of Effective Mass	350		

	Electrical Transport by a Hopping Process	401			
	Polaron Conduction	402			
7.3.2	Lifetime and Relaxation				
	Electrical Conduction	403			
	Lifetime Regime	403			
	Relaxation Regime	404			
7.4	Bulk Limited Electrical Conduction	406			
7.4.1	Basic Concepts Relevant to Space Charge Limited Electrical Conduction	406			
7.4.2	SCL Electrical Conduction: One Carrier (Single) Planar Injection	408			
	Theoretical Analyses	408			
	The Scaling Rule	420			
	The Effect of Carrier Diffusion	422			
7.5	Bulk Limited Electrical Conduction Involving Two Types of Carriers	423			
7.5.1	Physical Concepts of Carrier Trapping and Recombination	423			
	Capture Rates and Capture Cross-Sections	425			
	Recombination Rates and Recombination Cross-Sections	426			
	Demarcation Levels	427			
	Coulombic Traps	430			
	Characteristic Times	432			
7.5.2	Kinetics of Recombination Processes	433			
	Band-to-Band Recombination without Involving Recombination Centers or Traps	434			
	With a Single Set of Recombination Centers but without Traps	435			
7.5.3	Space-Charge Limited Electrical Conduction: Two-Carrier (Double) Planar Injection	437			
	Without Recombination Centers and without Traps	438			
	With Recombination Centers but without Traps	440			
7.6	High-Field Effects	443			
			7.6.1	Filamentary Charge-Carrier Injection in Solids	443
				Filamentary One-Carrier (Single) Injection	444
				Filamentary Two-Carrier (Double) Injection	446
			7.6.2	The Poole–Frenkel Detrapping Model	447
			7.6.3	The Onsager Detrapping Model	449
			7.6.4	Field-Dependent Carrier Mobilities	451
			7.7	Transitions between Electrical Conduction Processes	455
			7.7.1	Basic Transition Processes	455
				A Solid between Similar Contacts	455
				A Solid between Dissimilar Contacts	459
			7.7.2	Transition from Bulk-Limited to Electrode-Limited Conduction Process	460
			7.7.3	Transition from Electrode-Limited to Bulk-Limited Conduction Process	461
			7.7.4	Transition from Single-Injection to Double-Injection Conduction Process	462
			7.8	Current Transient Phenomena	463
			7.8.1	Space-Charge Free (SCF) Transient	466
				Case 1: In the Absence of Traps and Diffusion	466
				Case 2: In the Absence of Diffusion but with Traps	466
			7.8.2	Space-Charge Limited (SCL) Transient	468
				Case 1: In the Absence of Traps and Diffusion	468
				Case 2: In the Absence of Diffusion but with Traps	470
			7.8.3	Space-Charge Perturbed (SCP) Transient	470
				Carrier Generation by a Strong Light Pulse	470
				Carrier Generation by a Weak Light Pulse	471

7.9	Experimental Methodology and Characterization	472	7.13	Photosensitization	503
7.9.1	Simultaneous Measurements of Charging Current and Photocurrent Transients	472	7.14	Transient Photoconduction	505
7.9.2	Alternating Measurements of the I–V and C–V Characteristics	472		References	509
7.9.3	Measurements of Surface Potentials	477	8	ELECTRICAL AGING, DISCHARGE, AND BREAKDOWN PHENOMENA	515
7.9.4	Capacitance Transient Spectroscopy	478	8.1	Electrical Aging	515
PART II: PHOTOCONDUCTION		480	8.1.1	Theory	515
7.10	Quantum Yield and Quantum Efficiency for Photoconduction	480	8.1.2	Measurements of Electrical Aging	520
7.11	Generation of Nonequilibrium Charge Carriers	482		Electron Paramagnetic Resonance (EPR) Spectroscopy	521
7.11.1	Energy Distribution of Nonequilibrium Charge Carriers	484		Infrared (IR) Absorption Spectroscopy	522
7.11.2	Spatial Distribution of Nonequilibrium Charge Carriers	484		Surface Potential Measurements	523
7.11.3	Lifetimes of Nonequilibrium Charge Carriers	486		Small Angle Scattering of X-Rays (SASX) Spectroscopy	523
	Linear Recombination	486		Lifetime Evaluation	524
	Quadratic Recombination	488		Other Measurements	525
	Instantaneous Lifetimes	489	8.1.3	Remedy for Electrical Aging	525
7.12	Photoconduction Processes	490		Emission Shields	525
7.12.1	Intrinsic Photoconduction	490		Radical Scavengers	526
7.12.2	Extrinsic Photoconduction	490	8.2	Electrical Discharges	528
	Photocurrent—Voltage Characteristics	491	8.2.1	Internal Discharges	528
	Light Intensity Dependence	494		Fundamental Features of Internal Discharges in a Cavity	529
	Light Wavelength Dependence	495		Discharge Current Impulses, Recurrence, Discharge Magnitudes, Discharge Energy, and Power Losses	530
	Temperature Dependence	496	8.2.2	Electrical Treeing	540
	Effects of Surface Conditions and Ambient Atmosphere	497		Prebreakdown Disturbances and Light Emission	540
7.12.3	Homogeneous and Nonhomogeneous Photoconduction	497		Mechanisms and Characterization of Electrical Treeing	542
	Homogeneous Photoconduction	497		Some Special Features of Electrical Treeing	543
	Nonhomogeneous Photoconduction	499		Inhibition of Electrical Treeing	545
7.12.4	Photoresponse Times	500		Electrical Treeing Initiated by Water Trees	545

8.2.3	Surface Discharges and Corona Discharges	546	8.3.3	Electrical Breakdown in Solids	559
8.2.4	Detection of Partial Discharges	547		Thermal Breakdown	559
8.3	Electrical Breakdown	549		Electrical Breakdown	560
8.3.1	Electrical Breakdown in Gases	549	8.3.4	Similarity in Breakdown Mechanisms for Gas, Liquid, and Solid Dielectrics	567
	Townsend Discharges	551	References		569
	Streamer Discharges	553			
	Discharges in Electronegative Gases	555	<i>Index</i>		<i>573</i>
	Paschen's Law	555			
8.3.2	Electrical Breakdown in Liquids	557			

Preface

The word *dielectric* is derived from the prefix *dia*, originally from Greek, which means “through” or “across”; thus, the *dielectric* is referred to as a material that permits the passage of the electric field or electric flux, but not particles. This implies that the dielectric does not permit the passage of *any* kind of particles, including electrons. Thus, it should not conduct the electric current. However, a dielectric is generally considered a nonconducting or an insulating material. There is no ideal dielectric in this planet. The perfect vacuum may be considered to be close to the ideal dielectric, but a perfect vacuum cannot be obtained on Earth. A vacuum of 10^{-14} torr still consists of about 300 particles per cubic centimeter. All real dielectric materials are imperfect, and thus permit, to a certain degree, the passage of particles. We have to coexist with the imperfections.

This book deals mainly with the phenomena resulting from the responses of the solid dielectric materials to external applied forces such as electromagnetic fields, mechanical stress, and temperature. The materials considered are mainly nonmetallic and nonmagnetic materials, which are generally not considered as dielectric materials. There is no clear demarcation between dielectrics and semiconductors. We can say that the major difference lies in their conductivity and that the dominant charge carriers in semiconductors are generated by the thermal excitation in the bulk, while those in dielectric materials come from sources other than the thermal excitation, including carrier injection from the electrical contacts, optical excitations, etc. This book will not include semiconductors as such, but certain dielectric phenomena related to semiconductors will be briefly discussed.

Dielectric phenomena include induced and spontaneous electric polarizations, relaxation processes, and the behavior of charge carriers

responsible for the macroscopic electrical and optical properties of the materials. In dielectric materials, carrier traps arising from various structural and chemical defects and their interactions with charge carriers injected from electrical contacts or other excitation sources always play major roles in dielectric phenomena. In today’s high technology era, the trend of electronics has been directed to the use for some solid-state devices of some dielectric materials, such as ceramics, which have good insulating properties as well as some special features such as spontaneous polarization. Therefore, a chapter dealing with ferroelectric, piezoelectric, pyroelectric, and electro-optic phenomena, as well as a chapter dealing with electrets, are also included in this book.

The theoretical analyses are general. We have endeavored throughout this book to keep the mathematics as simple as possible, and emphasized the physical insight of the mechanisms responsible for the phenomena. We use the international MKSC system (also called the *SI system*, the International System of Units [Système Internationale]) in which the unit of length is meter (m), the unit of mass is kilogram (kg), the unit of time is second (s), and the fourth unit of electrical charge is Coulomb (C), because all other units for physical parameters can easily be derived from these four basic units.

It should be noted that to deal with such vast subjects, although confined to the area of dielectric phenomena in only nonmetallic and nonmagnetic materials, it is almost unavoidable that some topics are deliberately overemphasized and others discussed minimally or even excluded. This is not because they are of less importance, but rather, it is because of the limited size of this book. However, an understanding of dielectric phenomena requires knowledge of the basic physics of Maxwell’s equations and the general electromagnetic

theory. Thus, in the first chapter, Introduction, we have described rather briefly the basic concepts of the magnetization and the ferromagnetism that are analogous to the counterpart of the electric polarization and the ferroelectricity, and also the electromagnetic waves and fields. The reader may be familiar with much of the content of that chapter, but may still find it useful as background knowledge for the main topics of dielectric phenomena in the following chapters.

I would like, first, to thank gratefully my former colleagues and students for their cooperation and encouragement, and particularly, Professor Demin Tu of Xi'an Jiaotong University for his long-term collaboration with my research team on many projects related to pre-breakdown and breakdown phenomena. I would like to express my appreciation to the Faculty of Engineering at the University of Manitoba for the provision of all facilities in the course of the writing of this book.

I wish especially to thank my sons Hung Teh and Hung Pin, and my daughters Hung Mei and Hung Hsueh, and their families for all the love

and support; Hung Pin and his wife Jennifer Chan for their weekly "good wishes" by telephone from California. I am also very thankful to Hung Hsueh and the staff of the Sciences and Technology Library of the University of Manitoba for obtaining all the references I needed for the writing.

I also want to thank some of my close friends, and in particular, Chang Chan, Siu-Yi Kwok, Hung-Kang Hu, Pi-Chieh Lin, Jia-Yu Yang, and Feng-Lian Wong for their unfailing constant care and support, which have contributed so much to my patience and confidence in dealing with the writing.

Finally, I am deeply indebted to Ms. Karin Kroeker for her skillful typing and her patience in typing the mathematical expressions, and to the editors, Mr. Charles B. Glaser and Ms. Christine Kloiber of Academic Press for their helpful cooperation and prompt actions.

Last but not least, I would like to acknowledge gratefully the publishers and societies for permission to reproduce illustrations. The publishers and societies are listed in the following table.

<i>Publisher or Society</i>	<i>Book or Journal</i>	<i>Figure</i>
American Institute of Physics	J. Appl. Phys.	3-22, 3-51, 6-16, 6-19, 6-23, 6-24, 6-32, 7-25, 7-26, 7-43, 7-45, 7-47, 7-48, 7-59(b), 8-6, 8-7, 8-8, 8-10, 8-11, 8-12
American Institute of Physics	Appl. Phys. Lett.	4-23
American Institute of Physics	J. Chem. Phys.	3-46, 7-37, 7-54
	Soviet Phys.—Solid State	3-69
American Physical Society	Phys. Rev	3-37, 3-63, 4-11, 4-12, 4-17, 4-18, 4-25, 6-28, 6-30, 6-35, 7-18, 7-19, 7-59(a)
American Chemical Society	J. Amer. Chem. Soc.	2-18
American Vacuum Society	J. Vac. Sci. Tech	8-40
American Ceramic Society	Amer. Ceram. Soc. Bull.	3-20
	J. Amer. Ceram. Soc.	4-20
IEEE Dielectrics and Electrical Insulation Society	IEEE Trans. Electr. Insul. IEEE Electrical Insulation Magazine	5-20, 8-22, 8-23 5-19
John Wiley and Sons	J. Appl. Polymer Sci. J. Polymer Sci.—Physics	7-38, 8-5, 8-9, 8-13 5-16

<i>Publisher or Society</i>	<i>Book or Journal</i>	<i>Figure</i>
	J. Polymer Science—Polym. Phys. Ed.	4-42
Institute of Physics (London)	J. Phys. D: Appl. Phys.	7-5, 7-6
Physical Society of Japan	J. Phys. Soc. Japan	4-21
Australian Chemical Society	Aust. J. Chem.	7-34
Physica Status Solidi	Phys. Stat. Sol.	7-36
Steinkopff, Darmstadt	Kolloid Z.	2-42
Electrochemical Society	J. Electrochem. Soc.	8-34
Pergamon Press	Solid State Commun. Solid State Electronics J. Phys. Chem. Solids	3-52, 8-25 3-45, 6-25, 6-26, 7-27 3-47, 3-48
MOSES King (Cambridge)	Science	4-22
Materials Research Society	MRS Bulletin	3-68
IEEE Dielectric and Electrical Insulation Society	2000 IEEE International Symposium on Electrical Insulation	4-35
American Institute of Physics	Proc. 1965 International Symposium on Thin Film Physics	6-15
Clarendon (Oxford)	Theory of Dielectrics (H. Frohlich) High-Field Transport In Semiconductors (E.M. Conwell)	2-14 2-37
North Holland (Amsterdam)	Ferroelectricity (E. Fatuzzo and W.J. Merz)	2-38
Springer-Verlag (New York)	Electrets (G.M. Sessler)	5-14
Elsevier	Thermally Stimulated Discharge of Polymer Electrets (J. Van Turnhout)	5-15, 5-23
Plenum (New York)	Principles of Fluorescence Spectroscopy (J.R. Lakowicz)	3-40
Springer Verlag (New York)	Electroluminescence (J.I. Pankove)	3-44
Dover	Optical Processes in Semiconductors (J.I. Pankove)	3-70
Academic Press	Piezoelectric Ceramics (B. Jaffe, W.R. Cook and H. Jaffe)	4-19, 4-41
MacMillan (New York)	Ferroelectric Crystals (F. Jona and G. Shirane)	4-13, 4-14, 4-15, 4-16
Clarendon (Oxford)	Electronic Process in Non-Crystalline Materials (N.F. Mott and E.A. Davis)	6-34
Clarendon (Oxford)	Ionized Gases (A. Von Engel)	8-29

Kwan Chi Kao
Winnipeg, Manitoba, Canada

1 Introduction

Learning without thought is labor forgone; thought without learning is perilous.
Confucius (600 BC)

The basic distinction between a semiconductor and a dielectric (or insulator) lies in the difference in the energy band gap. At the normal ranges of temperatures and pressures, the dominant charge carriers in a semiconductor are generated mainly by thermal excitation in the bulk because the semiconductor has a small energy band gap; hence, a small amount of energy is sufficient to excite electrons from full valence band to an upper empty conduction band. In a dielectric, charge carriers are mainly injected from the electrical contacts or other external sources simply because a dielectric's energy band gap is relatively large, so a higher amount of energy is required for such band-to-band transitions. A material consists mainly of atoms or molecules, which comprise electrons and nuclei. The electrons in the outermost shell of atoms, bound to the atoms or molecules coupled with the free charges, interact with external forces, such as electric fields, magnetic fields, electromagnetic waves, mechanical stress, or temperature, resulting in the occurrence of all dielectric phenomena. For non-magnetic dielectric materials, the dielectric phenomena include mainly electric polarization; resonance; relaxation; energy storage; energy dissipation; thermal, mechanical, and optical effects and their interrelations; and electrical aging and destructive breakdown. The discussion of these phenomena is the scope of this book.

Dielectric phenomena, like other natural phenomena, were noticed long before the time of Christ. As early as 600 BC, the Greek philosopher Thales discovered that amber, when rubbed with cloth, attracted light objects

such as bits of chaff. In Greek, “amber” was referred to as *electricity*. However, it is now well known that many substances possess this property to some extent. A glass or metal rod, after being rubbed with a polyester sheet, will attract a light piece of paper. This attraction phenomenon may be considered due to the charge on the rod tip polarizing the paper nearby. The electric polarization produces an opposite charge on the paper surface close to the charged rod tip, resulting in this attraction. Any electromagnetic wave will induce polarization in dielectric materials and magnetization in magnetic materials. Both the polarization and the magnetization also produce their own fields, which interact with the external fields, resulting in a vast scope of dielectric and magnetic phenomena.

However, dielectric phenomena did not receive much attention until the middle of the 18th century, although the Leyden jar condenser, which could store charges, was discovered in 1745 by the Dutch physicist van Musschenbrack, of the University of Leyden.¹ About 90 years later (in 1837) Faraday, in England, was the first to report² that the capacitance of a condenser was dependent on the material inside the condenser. At that time, he called the ratio of the capacitance of the condenser filled with a dielectric material to that of the same condenser, empty inside (free space), the specific inductive capacity, which is now called the permittivity. In 1873, following the discovery of Coulomb's law on forces between charges, Ohm's law on electrical conductivity, Faraday's law, and Ampère's law on magnetic and electric induction, Maxwell³ welded these

discoveries together to formulate a unified approach. He developed four equations, known as Maxwell's equations, to govern all the macroscopic electromagnetic phenomena. Obviously, dielectric phenomena are part of the electromagnetic phenomena, which result from the interaction of the material with electromagnetic fields. Therefore, it is very important to understand the meaning of these four equations.

1.1 Maxwell's Equations

Maxwell's four equations are

$$\nabla \times H = J + \frac{\partial D}{\partial t} \quad (1-1)$$

$$\nabla \times F = -\frac{\partial B}{\partial t} \quad (1-2)$$

$$\nabla \cdot B = 0 \quad (1-3)$$

$$\nabla \cdot D = \rho \quad (1-4)$$

where F , D , H , and B are four vectors denoting, respectively, the electric field, the electric flux density (or electric displacement), the magnetic field, and the magnetic flux density (or magnetic induction); J is also a vector denoting the electric current density; and ρ is a scalar quantity denoting a net charge density. These equations imply that at any point inside a material there exist four vectors— F , D , H , and B —when that material is subjected to an external electromagnetic field, and that the distribution of the electric current may be considered the conservation of the electric charges, which give rise to the electromagnetic field. In other words, Maxwell's equations describe the coupling between the electric field and the magnetic field and their interaction with the material, resulting in all electromagnetic phenomena.

The parameter B may be linked to parameter H , and so D to F and J to F , by the following relations:

$$B = \mu H \quad (1-5)$$

$$D = \epsilon F \quad (1-6)$$

$$J = \sigma F \quad (1-7)$$

where μ , ϵ and σ are, respectively, the permeability, the permittivity, and the conductivity of the material (medium). Microscopic theory may deduce the physical properties of a material from its atomic structure, which may be generally represented by these three parameters: μ , ϵ , and σ . The nature of these parameters is directly associated with the aggregate effect of the deformation of the atomic structure and the movement of electric charges caused by the electromagnetic field, which is mainly due to magnetization, polarization, and electrical conduction. We shall discuss the polarization and electrical conduction in some detail in later chapters. As we shall confine ourselves to dealing only with nonmagnetic materials, we will discuss magnetization only briefly in this chapter, with the aim of clarifying the difference between magnetic and nonmagnetic materials.

In this world, there is no lossless material as such. All materials are lossy. Only in a perfect vacuum can there be no loss and, hence, no dispersion in the presence of an electromagnetic field. In practice, we cannot achieve a perfect vacuum on earth. Even at the pressure of 10^{-14} torr, the lowest pressure (or the best vacuum) that today's technology can achieve, there are still about 300 particles per cubic centimeter. However, at the normal range of temperatures and pressures, the gas media can be considered to be very close to the so-called free space. In free space, we have

$$\begin{aligned} \mu_0 &= 4\pi \times 10^{-7} \text{ henry } m^{-1} \\ &= 1.257 \times 10^{-6} \text{ henry } m^{-1} \end{aligned}$$

According to wave theory, the velocity of electromagnetic waves (light) in free space is

$$\begin{aligned} c &= (\mu_0 \epsilon_0)^{-1/2} = 2.998 \times 10^8 \text{ ms}^{-1} \\ &= 3 \times 10^8 \text{ ms}^{-1} \end{aligned}$$

From this, ϵ_0 in free space is

$$\begin{aligned} \epsilon_0 &= 8.854 \times 10^{-12} \text{ farad } m^{-1} \\ &= (36\pi \times 10^9)^{-1} \text{ farad } m^{-1} \end{aligned}$$

Obviously, in free space $\sigma = 0$ and $J = 0$.

In isotropic media, we would expect that at any point D and J are parallel to F , and B is

parallel to H . The beauty of the Maxwell's equations is that a vast scope of electromagnetic phenomena can be described with only a few variables. Usually, we use the relative values of μ and ϵ , which are expressed as

$$\mu_r = \mu/\mu_0 \quad (1-8)$$

$$\epsilon_r = \epsilon/\epsilon_0 \quad (1-9)$$

μ_r and ϵ_r are called, respectively, the relative permeability and the relative permittivity (or simple dielectric constant). The parameters μ_r , ϵ_r , and σ generally characterize the electromagnetic properties of materials. Thus, information about the dependence of these parameters on some physical variables, such as density, temperature, field intensity, and frequency, would shed much light on the internal structure of matter.

As Maxwell's equations govern both the electrical and the magnetic properties of matter, as well as all electromagnetic phenomena in any medium, it is important to understand the physics and the historical development behind these equations.

Faraday discovered not only the dependence of the capacitance of a condenser on the material filled between the two metallic plates, but also the induction law, which involves a voltage induced in a coil when a time-varying magnetic field is in the region surrounded by the coil. At about the same time, Henry, in the United States, discovered the self-induction of the electric current, which led to the development of electromagnets. In fact, about 10 years before this discovery, in 1820, the Danish scientist Oersted observed the magnetic effect of an electric current.¹ After the announcement of Oersted's discovery, the French scientist Ampère discovered, in about 1824, the circuital law: the line integral of the magnetic field intensity around any closed path is equal to the total current linked with that path.

1.1.1 Ampère's Law

Suppose we have an iron core with a gap filled with air or a nonmagnetic material, with a uniform cross-section area A , and a mean magnetic flux path length ℓ_m around the iron core

and a length ℓ_a across the gap, as shown in Figure 1-1(a). When the coil of N turns around the core carries a current i , then (ignoring any flux leakage traversing the path for mathematical simplicity) the magnetomotive force (usually called the magnetomotive force, *mmf*) U_m can be expressed by

$$U_m = H_m \ell_m + H_a \ell_a = Ni \quad (1-10)$$

where H_m and H_a are, respectively, the magnetic field intensities in the iron core and the gap. Since the magnetic flux ϕ is continuous, as is the magnetic flux density B , ϕ and B are the same in the iron core and in the gap. Thus,

$$B = B_m = B_a = \frac{\phi}{A} \quad (1-11)$$

and Equation 1-10 may be written in the form

$$Ni = B \left(\frac{\ell_m}{\mu_m} \right) \left(1 + \frac{\mu_m}{\mu_a} \frac{\ell_a}{\ell_m} \right) \quad (1-12)$$

This shows that for a fixed magnetizing current i , the smaller the gap length ℓ_a , the larger the magnetic flux density B . This is the general feature of electromagnets.

We can write Equation 1-10 in general form as

$$\sum_i H_i \ell_i = Ni \quad (1-13)$$

A coil with N turns carrying a current i is the same as a coil with only one turn but carrying a current $I = Ni$. Since the north and south magnetic poles are inseparable, the magnetic flux lines must close on themselves. Considering these facts, Equation 1-13 can be written as

$$\oint_c H d\ell = I = \int_a J \cdot \vec{n} ds \quad (1-14)$$

where $\vec{n}$ is the unit vector normal to the surface. According to Stokes's theorem, the line integral of a vector around the boundary of a surface with an area S is equal to the surface integral of the curl of the same vector over the area bound by the path of this vector, as illustrated in Figure 1-1(b).

So we can write

$$\int_s (\nabla \times H) \cdot \vec{n} ds = \oint_c H d\ell \quad (1-15)$$

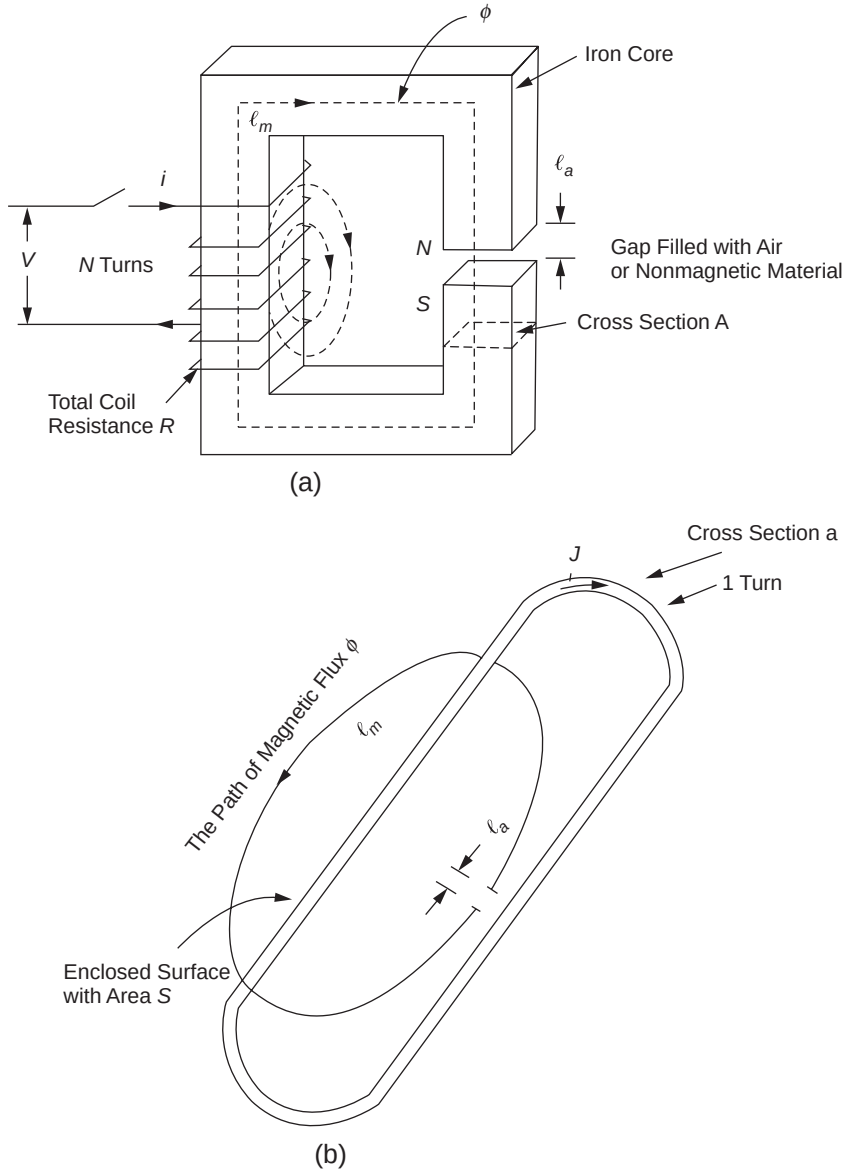


Figure 1-1 Schematic diagrams illustrating (a) Ampère's circuital law and an electromagnet and (b) Stokes's theorem based on the flux path in (a).

The surface can be an arbitrary surface, provided that it includes the current passing through the surface. This implies that $\int J \cdot \vec{n} ds = \int J \cdot \vec{n} ds$. Thus, from Equations 1-14 and 1-15 we obtain

$$\int_s (\nabla \times H) \cdot \vec{n} ds = \int_s J \cdot \vec{n} ds \quad (1-16)$$

These two surface integrals are equal; thus, we have

$$\nabla \times H = J \quad (1-17)$$

This equation is valid for both the steady current (DC) and the time-varying current (AC, or pulse or transient current). For the steady current (DC), ϕ becomes constant, and hence,

$\frac{d\phi}{dt} = 0$, but for the time-varying current (such as AC), the total current may also involve the displacement current $\frac{\partial D}{\partial t}$. If this is the case, Equation 1-17 must be written as

$$\nabla \times H = J + \frac{\partial D}{\partial t} \quad (1-18)$$

This is the Ampère's circuital law, which forms the first equation of Maxwell's equations. Depending on the situation, Equation 1-18 may involve only conduction current, with $\frac{\partial D}{\partial t} = 0$ (such as the magnetic fields in electrical machines), or it may involve only displacement current with $J = 0$ (such as the electromagnetic waves in free space). It should be noted that $\nabla \times H = 0$ does not happen, because this means mathematically⁴⁻⁶ that there exists a scalar field such that H is equal to ∇ (scalar field), but such a scalar field does not exist in magnetic circuits. This also implies that the north and south magnetic poles are inseparable.

1.1.2 Faraday's Law

Electromagnetic induction is a very interesting and important phenomenon, so we include here a brief discussion about the physics behind it. Let us return to the electromagnet system shown in Figure 1-1(a). Suppose the coil of copper wire, with a total resistance R , is connected to a DC voltage V through a switch. As soon as the switch is turned on, a magnetic flux will be induced in the magnetic circuit. During the time interval dt , energy equal to $Vidt$ will be supplied to the system from the DC source, of which $i^2 R dt$ energy (as heat loss) will be consumed in the coil, leaving $(Vi - i^2 R)dt$ in energy stored in the magnetic field

$$dW_m = (V - iR)idt \quad (1-19)$$

During the time interval, the magnetic flux changes with time: $d\phi/dt$. This change in magnetic flux, according to Faraday's induction law, will produce voltage equal to $XN d\phi/dt$. Thus, during dt , the stored energy dW_m should be

$$dW_m = i \left(N \frac{d\phi}{dt} \right) dt \quad (1-20)$$

From Equations 1-19 and 1-20, we obtain

$$i = \frac{V - N \frac{d\phi}{dt}}{R} = \frac{V + V_i}{R} \quad (1-21)$$

where V_i is the induced voltage, which is

$$V_i = -N \frac{d\phi}{dt} \quad (1-22)$$

Equation 1-22 means that whenever the current or its accompanying magnetic flux changes with time, a voltage V_i will be generated in the circuit, called the induced voltage, with the polarity opposite to the source voltage and the magnitude equal to the time rate of change of the flux. This is Lenz's law, which is the consequence of the principle of conservation of energy. V_i can also be written in terms of time rate of change of the current as

$$V_i = -\frac{d}{dt}(Li) = -L \frac{di}{dt} \quad (1-23)$$

where L is a constant generally called the self-inductance of the circuit. From Equations 1-22 and 1-23, we have

$$L = N \frac{d\phi}{di} = -V_i \bigg/ \frac{di}{dt} \quad (1-24)$$

The self-inductance L in the electromagnetic system is analogous to the capacitance C in the dielectric system in which $i = C dV/dt$. The latter will be discussed in Chapter 2.

Because of the parameters R and L in the magnetic circuit, when the switch is turned on, it takes some time for the current to establish its final, steady value. Based on the simple equivalent circuit, the system shown in Figure 1-1(a), we can write

$$V = Ri + L \frac{di}{dt} \quad (1-25)$$

For magnetic materials, L is not always a constant, so ϕ is not a linear function of i based on Equation 1-24. However, L can be considered a constant for nonmagnetic materials and can be assumed to be approximately constant for

magnetic materials if the magnetic flux density is sufficiently low or the magnetizing current is sufficiently small. When L is constant, the solution of Equation 1-25, using the initial boundary condition, when $t = 0$, $i = 0$, gives

$$i = \frac{V}{R} \left[1 - \exp\left(-\frac{R}{L}t\right) \right] \quad (1-26)$$

The variation of i with t is shown in Figure 1-2.

Theoretically, i never rises to its final value $I = V/R$, but practically this is usually accomplished in a rather short time. From Equations 1-24–1-26, $\tau = L/R$ is a time constant. Initially, i rises linearly with time, but the rate of the current rise gradually decreases because when i rises, there is an increasing storage of energy in the magnetic field. By multiplying Equation 1-25 by idt , we have

$$Vidt = Ri^2dt + Lidi \quad (1-27)$$

$\swarrow$ $\uparrow$ $\swarrow$
 Energy supplied Energy dissipated Energy stored
 to the magnetic as heat in the magnetic
 system loss in field
 the coil

When $t = \tau$, the current reaches about 63% of its final value $I = V/R$. The rate di/dt gradually decreases, implying that the induced voltage V_i decreases gradually, but the current and hence ϕ gradually increase to their final values. When i reaches its final value I , the induced voltage

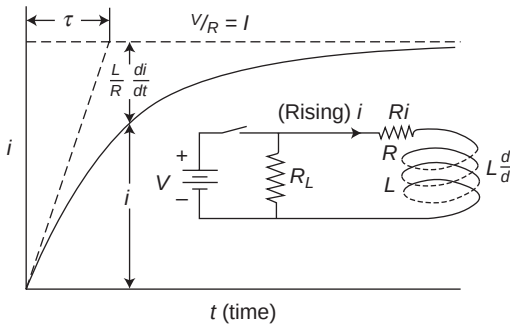


Figure 1-2 The current rise in an inductive circuit—equivalent circuit of the electromagnet system shown in Figure 1-1(a)

approaches zero, and the total energy stored in the magnetic field will be $\int_0^I Lidi = \frac{1}{2}LI^2$. This

phenomenon is analogous to the response of a C and R series circuit representing a dielectric system. In this case, the time constant is CR . When the voltage across the capacitor approaches the value of the applied DC voltage V , the charging current will approach zero. By this time, the energy stored in the capacitor is $\frac{1}{2}CV^2$.

For a large electromagnet, the amount of energy stored in the system could be quite large. If the switch is suddenly opened to disconnect the system from the source, there is an induced voltage, which is theoretically infinite because $di/dt \rightarrow \infty$ if there is no protective resistor R_L across the coil. In this case, the huge induced voltage may cause a spark across the switch or damage in the coil insulation. To protect the system, a protective resistor R_L is usually connected across the coil, as shown in the inserted circuit in Figure 1-2. This resistor, in series with the coil resistance R , will absorb the energy released from the system when and after the switch is opened.

Having discussed the concept of the electromagnetic behavior under a DC condition, we now turn to what is different under a time-varying current condition. Let us use the system shown in Figure 1-3(a), in which we have a closed iron core with a cross section area A and mean flux path length ℓ_m , a coil of N_1 turns on one side, and a coil of N_2 turns on the other side. This system is similar to that shown in Figure 1-1(a), except there is no gap in the core ($\ell_a = 0$). Note that the following analysis is valid for the core with or without a gap. We chose the core without a gap in order to use this system to illustrate the principle of transformers.

If the coil with N_1 turns is connected to an AC voltage source, the + sign means the positive polarity, i.e., the positive half-cycle of the AC voltage at the coil terminal at a particular moment as a reference. At that particular moment, the current i_1 will flow in coil 1 (primary coil) and induce a flux ϕ circulating the iron core, as shown in Figure 1-3(a).

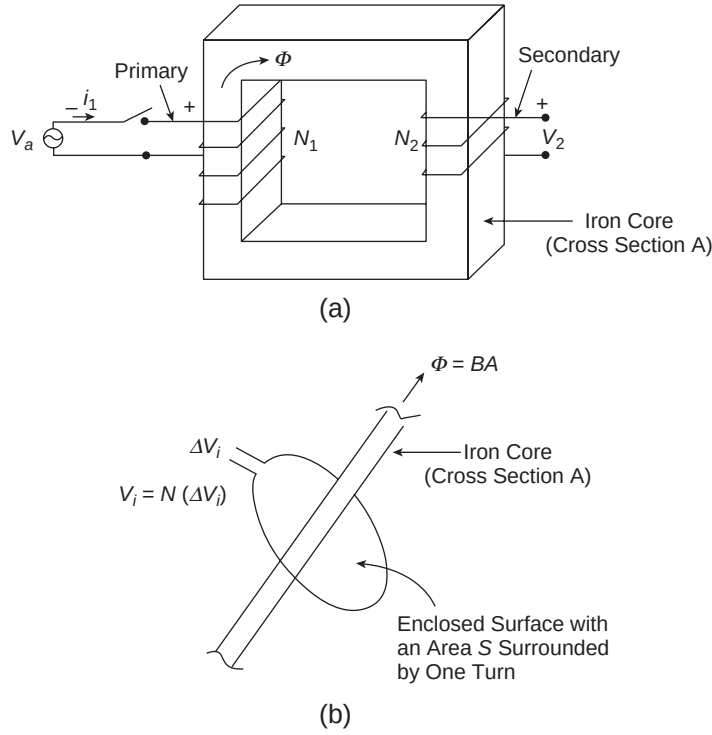


Figure 1-3 Schematic diagrams illustrating (a) Faraday's law and the transformer principle, and (b) Stokes's theorem based on the flux path in (a) for one turn.

According to Faraday's law and Lenz's law, this induced flux ϕ will also induce a voltage V_i opposite to the applied voltage V_1 and equal in magnitude to $V_a - R_1 i_1$, where R_1 is the resistance of coil 1 in order to oppose any change of the flux. V_i is given by

$$|V_i| = V_a - R_1 i_1 = V_1 = N_1 \frac{d\phi}{dt} \quad (1-28)$$

The same flux is also linked with coil 2 (secondary coil) of N_2 turns. This flux will induce a voltage in coil 2, which is given by

$$V_2 = -N_2 \frac{d\phi}{dt} \quad (1-29)$$

It is desirable to give a careful definition of the polarity of the coil terminals in relation to the flux. A positive direction of the flux is arbitrarily chosen as a reference; the terminals are then labeled in such a way that the winding of the coil running from the positive to the negative terminal constitutes a right-handed rotation

about the positive direction of the flux, as shown in Figure 1-3(a).

Now let us consider only one turn. The voltage induced in one turn will be

$$\nabla V_i = -\frac{d\phi}{dt} = -\frac{d}{dt} \int_s \vec{B} \cdot \vec{n} ds \quad (1-30)$$

The line integral of the electric field F around the coil of only one turn is equal to ΔV_i . Thus,

$$\nabla V_i = \oint_c F d\ell = -\int_s \frac{\partial \vec{B}}{\partial t} \cdot \vec{n} ds \quad (1-31)$$

This is an expression of Faraday's law of magnetic induction. Applying Stokes's theorem [see Figure 1-3(b)], the left side of Equation 1-31 can be expressed as

$$\oint_c F d\ell = \int_s \nabla \times \vec{F} \cdot \vec{n} ds \quad (1-32)$$

Substitution of Equation 1-32 into Equation 1-31 gives

$$\int_S \nabla \times \mathbf{F} \cdot \vec{n} ds = - \int_S \frac{\partial \mathbf{B}}{\partial t} \cdot \vec{n} ds \quad (1-33)$$

because the surface S is an arbitrary surface bounded by the loop C . Therefore, the two surface integrals are equal, so we can write

$$\nabla \times \mathbf{F} = - \frac{\partial \mathbf{B}}{\partial t} \quad (1-34)$$

This is an expression of Faraday's law and is also the second equation of Maxwell's equations in differential form.

It is obvious that Figure 1-3(a) is the basic arrangement of a transformer, which can transform a low voltage to a high voltage simply by adjusting the turn ratio $N_2/N_1 > 1$, or vice versa.

The time-varying magnetic flux can be achieved by many means. For example, with stationary north and south magnetic poles arranged alternatively in a chain with a small gap between the north and south poles, a copper coil moving along this chain will be linked by the magnetic flux, which changes with time from the north to the south pole and then to the north pole again, and so on. A voltage will be generated in the coil due to $d\phi/dt$. Alternatively, the same effect would result if the coil were kept stationary and the magnetic pole chain were moving through the coil.

A coil or a wire carrying a current placed in a magnetic field will experience a force acting on it. This is the magnetic force given by

$$\vec{F} = q\vec{v} \times \vec{B} \quad (1-35)$$

where $\vec{v}$ is the velocity of the electrons moving in the coil or wire. The direction of this force is perpendicular to both $\vec{B}$ and $\vec{v}$. This is why the electric motor works. The direction of this force follows Fleming's so-called *left-hand rule*, as shown in Figure 1-4(a). This rule states that if the forefinger points in the direction of the magnetic field and the middle finger points in the direction of the current flowing in the conductor (coil or wire), then the thumb will point in the direction of the force which tends to make the coil or wire move. In fact, this phenomenon can be visualized by the crowding of the magnetic flux, which tends to push the conductor from the region with dense flux to the region of less flux, as shown in Figure 1-4(b).

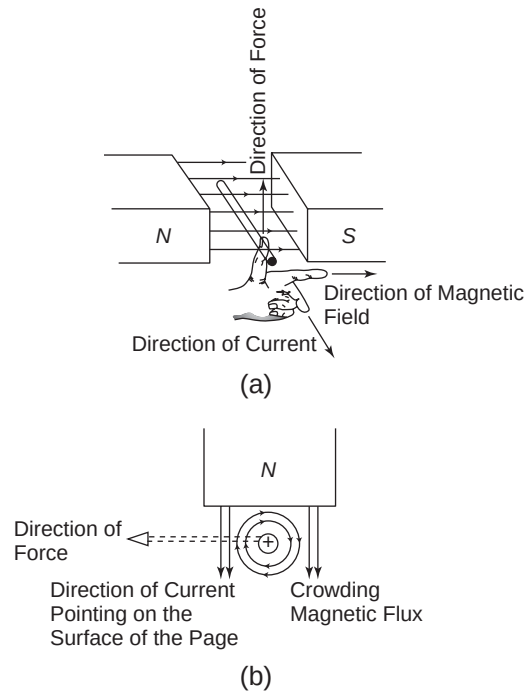


Figure 1-4 (a) Fleming's left-hand rule, and (b) crowding magnetic flux pushing the conductor from a high-flux density region to a weak-flux density region.

The principles of all DC and AC machines—stationary (e.g., electromagnets and transformers) or nonstationary (e.g., generators and motors)—are simply based on these three laws—Ampère's, Faraday's, and Lenz's—coupled with magnetic forces.

1.1.3 Inseparable Magnetic Poles

The third of Maxwell's equations, $\nabla \cdot \mathbf{B} = 0$, merely states the fact that magnetic flux lines are continuous. In other words, the number of magnetic flux lines entering any given volume of a region must equal the number of flux lines leaving the same volume. This means that the magnetic flux lines always close themselves, because the north and south magnetic poles are inseparable, no matter how small the magnets are. Electrons, protons, and neutrons produce their own magnets, and each always comes with both a north pole and a south pole.

Breaking a magnet in two just gives two smaller magnets, not separated north and south poles. No magnetic monopoles exist. Thus, the magnetic effects in a material must originate from the magnetism of the constituent particles as an immutable fact of nature.

1.1.4 Gauss's Law

Divergence of a flux means the excess of the outward flux over the inward flux through any closed surface per unit volume. For electric flux density D , the $\nabla \cdot D$ means

$$\begin{aligned}\int_S D \cdot \vec{n} ds &= \int_V \nabla \cdot D dV = \int \rho dV \\ &= 0 \text{ if } Q \text{ (the net charge) is located} \\ &\quad \text{outside the surface } S \\ &= Q \text{ if } Q \text{ is inside the volume } V \\ &\quad \text{enclosed by the surface } S.\end{aligned}\quad (1-36)$$

If the region enclosed by the surface S has a net charge Q that is equal to $\int \rho dV$, then we can write

$$\nabla \cdot D = \rho \quad (1-37)$$

This is Gauss's law; it is also the fourth of Maxwell's equations.

If the permittivity of the region ϵ is independent of the field and the material is isotropic, then Equation 1-37 can be written as

$$\nabla \cdot F = \frac{\rho}{\epsilon} \quad (1-38)$$

This is Poisson's equation. Since

$$F = -\nabla V \quad (1-39)$$

where V is the scalar potential field in the region. In terms of scalar parameters, Poisson's equation can be written as

$$\nabla \cdot (\nabla V) = \nabla^2 V = -\frac{\rho}{\epsilon} \quad (1-40)$$

If $Q = 0$, that is, if the region is free of charges, then Equation 1-40 reduces to

$$\nabla^2 V = 0 \quad (1-41)$$

Equation 1-41 is known as Laplace's equation. Both Poisson's equation and Laplace's equation will be very useful in later chapters when we

are dealing with space charges and related subjects.

The parameters ϵ , μ , and σ in Equations 1-5 through 1-7 depend on the structure of materials. Therefore, information about the dependence of these parameters on the field strength, frequency, temperature, and mechanical stress may reveal the structure of the materials, as well as the way of developing new materials. Equation 1-7 is Ohm's law, discovered by German scientist Ohm in 1826, referring mainly to metallic conductors at that time. In fact, all of these parameters are dependent on field strength and frequency, even at a constant temperature and pressure condition. It can be imagined that to solve Maxwell's equations with the field-dependent and frequency-dependent ϵ , μ , and σ would be quite involved. For nonmagnetic materials, we can assume that $\mu = \mu_0$ and is constant, but ϵ and σ are always field- and frequency-dependent. In subsequent chapters, we shall deal with these two parameters and discuss how they are related to all dielectric phenomena.

1.2 Magnetization

In this book, we shall deal only with nonmagnetic materials. What kind of materials can we consider nonmagnetic? All substances do show some magnetic effects. In fact, many features of the magnetic properties of matter are similar or analogous to the dielectric properties. Induced magnetization is analogous to induced polarization. For some materials, atoms or molecules possess permanent magnetic dipoles, just as other materials do with permanent electric dipoles. Some materials possess a spontaneous magnetization, just as other materials possess a spontaneous polarization. However, there is a basic difference between magnetic and dielectric behaviors. Individual electric charges (monopoles) of one sign, either positive or negative charges, do exist, but the corresponding magnetic monopoles do not occur. This was explained briefly in Section 1.1.3.

We have mentioned that magnetism is the manifestation of electric charges in motion based on Ampère's and Faraday's laws, so we

can expect that the electrons circulating around the nucleus and the electrons spinning themselves in the atom, as well as the spinning of protons and neutrons inside the nucleus, will produce magnetic effects.

Before discussing the physical origins of the magnetism, we must define some parameters that are generally used to describe the properties of matter. Polarization P describes dielectric behavior, and magnetization; M describes magnetic behavior. P is defined as the total electric dipole moments per unit volume. The same goes for M , which is defined as the total magnetic dipole moments per unit volume. Let us consider a cube-shaped piece of material with a unit volume cut out from the iron core, as shown in Figure 1-3. We can see in this cube that there is a magnetization M , as shown in Figure 1-5. A cube of magnetized material contains of the order of 10^{22} magnetic dipoles, which tend to line up just like a compass needle when they are magnetized in a magnetic field.

For electric polarization, the electric flux inside a polarized material has two components: One component is for setting up an electric field and the other component is due to the polarization. (See Chapter 2, Electric Polarization and Relaxation in Static Electric Fields.) In a similar manner, the magnetic flux inside a magnetized material also has two components: One component is for setting up a magnetic field H , and the other is due to the magnetiza-

tion M . Thus, under an applied magnetic field we have

$$B = \mu_0 H + \mu_0 M \quad (1-42)$$

The magnetization M has the same dimension as H , which is ampere per unit length. M can be expressed as

$$\begin{aligned} M &= \text{Total magnetic dipole moments/volume} \\ &= N(\mu_m) = N(ph) = Nh(p) \end{aligned} \quad (1-43)$$

where N is the number of atoms or molecules per unit volume. The elementary magnetic dipoles are in fact the atoms or molecules, which are the constituent particles of the material. They can become magnetic dipoles pointing in one direction under the influence of a magnetic field. Thus, $\mu_m = ph$ is the magnetic dipole moment, where h is the distance between the north and south magnetic poles. Since the dimension of M is ampere per unit length (or ampere-length⁻¹), the dimension for the pole strength p is ampere-length and that for the magnetic dipole moment ph becomes ampere-length². It is easy to understand the charge q in the electric dipoles. But what is meant by p in the magnetic dipoles? To understand p , we must look into the mechanisms responsible for the magnetic effect of the constituent particles, which are atoms or molecules. Before doing so, return to Equation 1-42. From this equation we have

$$\begin{aligned} M &= \frac{B}{\mu_0} - H = \left(\frac{\mu}{\mu_0} - 1 \right) H = (\mu_r - 1)H \\ &= \chi_m H = N\langle \mu_m \rangle = N\langle ph \rangle \end{aligned} \quad (1-44)$$

where $\langle \rangle$ denotes the average value of μ_m over the whole ensemble, and χ_m is the magnetic susceptibility in analogy to the dielectric susceptibility.

$$\chi_m = \frac{M}{H} = \mu_r - 1 \quad (1-45)$$

Obviously, χ_m reflects the degree of magnetization. In free space, or in gases at normal temperatures and pressures, we can consider $\chi_m = 0$, implying that there is no magnetization: $M = 0$. Depending on the values of χ_m , all materials can be divided into three major groups: diamagnetic materials with μ_r very slightly less than unity, (i.e., $\chi_m < 0$); paramagnetic materials with μ_r very slightly greater than unity (i.e., $\chi_m \gg 0$); and ferromagnetic materials with μ_r enormously

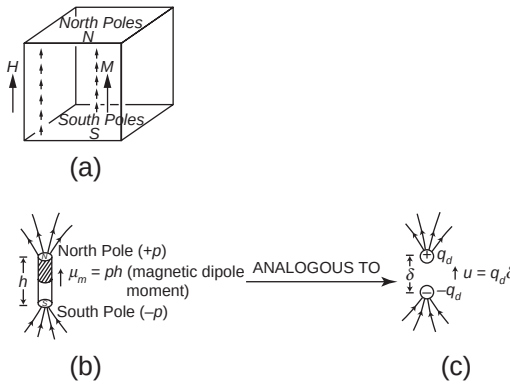


Figure 1-5 Schematic diagrams illustrating (a) magnetization, (b) magnetic dipole of the magnetic dipole moment $u_m = ph$, and (c) electric dipole of the electric dipole moment $u = q_d \delta$.

greater than unity (i.e., $\chi_m \gg 0$). The physical origins of these three kinds of materials are briefly discussed in the following sections.

The Orbital Motion of Electrons around the Nucleus of an Atom

The orbital motion of each electron constitutes a current i circulating around the nucleus, following the path of the orbit of radius r and taking time T_{or} to complete one revolution. Thus, the orbital current i can be expressed as

$$i = -\frac{q}{T_{or}} = -\frac{qv}{2\pi r} = -\frac{q\omega}{2\pi} \quad (1-46)$$

where v and ω are, respectively, the circumferential velocity and the angular velocity (radians/second), and q denotes the electronic charge, which is always positive. Thus, an electron's charge is $-q$, while a proton's charge is $+q$. It is important to remember the sign to avoid confusion. This current i will produce an orbital magnetic moment (called an orbital dipole moment), which is given by

$$\begin{aligned} \mu_{or} &= ph = \frac{B_{or}a}{\mu_0} h \\ &= H_{or}ah = \left(\frac{i}{h}\right)ah = ia \end{aligned} \quad (1-47)$$

where H_{or} and B_{or} are, respectively, the magnetic field and the magnetic flux density produced by i , and a is the area of the orbit.

The revolving electron under the influence of centrifugal force also produces an orbital angular momentum, which is given by

$$L_{or} = mvr = m\omega r^2 \quad (1-48)$$

where m is the electron mass. Substituting Equations 1-46 and 1-48 into Equation 1-47, we have

$$\mu_{or} = \left(-\frac{q\omega}{2\pi}\right)(\pi r^2) = \left(-\frac{q\omega r^2}{2}\right) = \left(-\frac{q}{2m}\right)L_{or} \quad (1-49)$$

The orbital magnetic moment μ_{or} is proportional to the orbital angular momentum L_{or} . Both are normal to the plane of the current loop, but they are in opposite directions, as shown in Figure 1-6(a). The coefficient $(-q/2m)$, i.e., the ratio of μ_{or}/L_{or} , is called the gyromagnetic ratio.⁷

It is very important to understand that electrons of an atom are entirely quantum-mechanical in nature. The classical approach can help us to visualize qualitatively the mechanism, but for quantitative analysis, we must use quantum mechanics. Equation 1-48 gives only the classic angular momentum. In quantum mechanics, the angular momentum is given by

$$L_{or} = m_\ell \left(\frac{h}{2\pi}\right) = m_\ell \hbar \quad (1-50)$$

where h is Planck's constant and $\hbar = h/2\pi$ is generally called the h-bar; m_ℓ is a quantum number. To describe electron behavior in an atom, we need four quantum numbers, defined as follows:

- n : The shell quantum number, defining the energy level of the shell. It takes integers, 1, 2, 3, 4, ...
- ℓ : The orbital or the azimuthal quantum number, defining the orbital type in each shell. It takes integers, 0, 1, 2, ... ($n-1$).
- m_ℓ : The magnetic orbital quantum number, defining the possible ways of electron motion in the orbit. It takes integers with the values limited by $-\ell \leq m_\ell \leq +\ell$.
- s : The spin quantum number, defining the possible ways of electron rotation about its own axis. It takes either $+1/2$ or $-1/2$.

The relation between ℓ and m_ℓ is as follows:

The value of ℓ :	0 1 2 3 4 ($n-1$)
The corresponding orbital type:	s p d f g
The number of possible orbitals in one type:	1 3 5 7 9 ($2\ell+1$)
The corresponding value of m_ℓ :	$\uparrow$ 0 $\uparrow$ (-1, 0, +1) $\uparrow$ (-2, -1, 0, +1, +2) and so on following the relation $-\ell \leq m_\ell \leq +\ell$ $\swarrow$ $\swarrow$ $\swarrow$ $\swarrow$ $\swarrow$

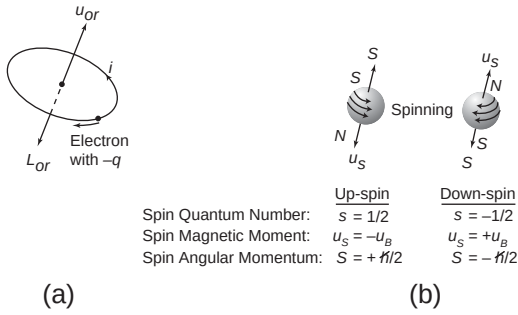


Figure 1-6 Schematic diagrams showing (a) orbital magnetic dipole moment u_{or} created by an electric current i due to the electron orbiting around the nucleus, and orbital angular momentum L_{or} of the electron and (b) spin magnetic dipole moment u_s and the associated spin angular momentum s for the up-spin and the down-spin.

A piece of material contains a large number of atoms or molecules. Each orbital electron has a magnetic moment μ_{or} , but the directions of all the magnetic moments are oriented in a random manner, and their magnetic effects tend to cancel each other out. Thus, the material does not exhibit any magnetic effect because there is no net magnetization in any particular direction without the aid of an external magnetic field. It should also be noted that, if there is only one electron in the s orbital (or s state), the orbital angular momentum $L_{or} = 0$ and hence the orbital magnetic moment $\mu_{or} = 0$, such as hydrogen ($1s^1$). In this case, the magnetic moment is only that of the electron spin. If there are only two electrons in the s orbital, as in helium ($2s^2$), both the orbital and the spin magnetic moments are zero because in this case, the angular momentum $L_{or} = 0$, and the magnetic moments due to the up-spin and the down-spin tend to cancel each other out.

The Rotation of Electrons about Their Own Axis in an Atom

The electron's rotation about its axis, generally referred to as the electron spin, also produces a spin magnetic moment u_s and the accompanying spin magnetic momentum S . The spin motion is entirely quantum mechanical. Similar to orbital electrons, the spin magnetic moment

is also proportional to its angular momentum and the directions of spin magnetic momentum and orbital magnetic momentum are opposite to each other, as shown in Figure 1-6(b), following the relation

$$\mu_s = \left(-\frac{q}{m}\right)S \quad (1-51)$$

where S is the spin angular momentum.

The spin gyromagnetic ratio $(-q/m)$ is twice that for orbital electrons. According to quantum mechanics, the spin angular momentum is given by

$$S = s\left(\frac{h}{2\pi}\right) = s\hbar \quad (1-52)$$

As has been mentioned, the spin quantum number s can take only $+1/2$ (up-spin) or $-1/2$ (down-spin). Thus u_s can be written as

$$\begin{aligned} \mu_s &= \left(-\frac{q}{m}\right)\left(+\frac{\hbar}{2}\right) = -\left(\frac{q\hbar}{2m}\right) \\ &= -\mu_B \text{ for up-spin} \end{aligned}$$

or (1-53)

$$\begin{aligned} \mu_s &= \left(-\frac{q}{m}\right)\left(-\frac{\hbar}{2}\right) = +\left(\frac{q\hbar}{2m}\right) \\ &= +\mu_B \text{ for down-spin} \end{aligned}$$

The quantity $(q\hbar/2m)$ is called the Bohr magneton u_B , which is equal to 9.3×10^{-24} ampere- m^2 . For $m_\ell = \pm 1$, $L = \pm\hbar$ the orbital magnetic moment u_{or} is equal to the spin magnetic moment $u_{or} = \mu_s = u_B$.

The Rotation of Protons and Neutrons inside the Nucleus

The rotation of protons and neutrons also contributes to magnetic moments, but their magnetic effect is much weaker than electrons. Magnetic moments of protons and neutrons are of the order of 10^{-3} times smaller than the magnetic moments of electrons.⁸ In the following discussion, we shall ignore the magnetic effect due to protons and neutrons for simplicity.

In any atoms, the electrons in inner closed shells do not have a net magnetic moment because in these shells all quantum states are filled and there are just as many electron orbital

and electron spin magnetic moments in one direction as there are in the opposite direction, so they tend to cancel each other. Only the electrons in partially filled shells (mainly in outermost shells) may contribute to net magnetic moments in a particular direction with the aid of an external magnetic field. In the following section, we shall discuss briefly the mechanisms responsible for the three kinds of magnetization.

1.2.1 Diamagnetism

In most diamagnetic materials, atoms have an even number of electrons in the partially filled shells, so that the major magnetic moments are due to electron motion in orbits, the spin magnetic moments (one in up-spin and one in down-spin) tend to cancel each other. Under an external magnetic field, the field will exert a torque, acting on the electron. This torque is given by

$$\tau = \frac{d}{dt} L_{or} = \vec{\mu}_{or} \times \vec{B} \quad (1-54)$$

The direction of the torque vector is perpendicular to both u_{or} and B . This torque tends to turn the magnetic dipole to align with the external magnetic field in order to reduce its potential energy. By rewriting Equation 1-54 in the form

$$\frac{dL}{dt} = \left(-\frac{q}{2m} \right) L \times \vec{B} \quad (1-55)$$

or

$$dL = \left(-\frac{q}{m} \right) \vec{L} \times \vec{B} dt$$

it is clear that dL is perpendicular to L and B . Since B is constant, the only possibility that the momentum can change with time is for it to rotate or precess about the B vector, as shown in Figure 1-7.

Thus, Equation 1-54 can be written as

$$\frac{dL}{dt} = -\vec{L} \times \frac{q\vec{B}}{2m} = -L \times \omega_p \quad (1-56)$$

where ω_p is

$$\omega_p = \frac{qB}{2m}$$

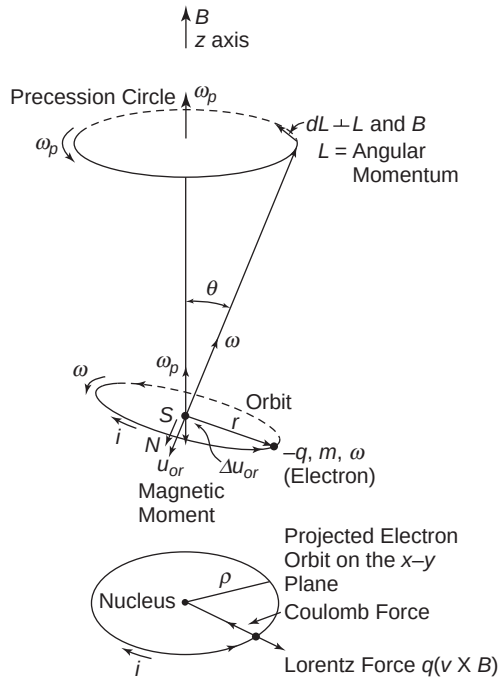


Figure 1-7 Larmor precession of an electron orbit about a magnetic field $H = u_e B$.

which is known as the Larmor frequency. Theoretically, the magnetic dipole does not tend to align itself with the magnetic field, but rather to precess around B without even getting close to the direction of the magnetic field. In a pure Larmor precession, no alignment would take place. However, the precession always encounters many collisions, and the dipole would then gradually lose energy and approach alignment with B because under the condition of alignment, the potential energy $-u_{or}B$ becomes a minimum. Because of the torque, the orientational potential energy of the magnetic dipole can be written as

$$\begin{aligned} U_{or} &= -\mu_{or}B = -\left(-\frac{q}{2m} \right) LB \\ &= \left(\frac{q}{2m} \right) (m_e \hbar) B = m_\ell \mu_B B \end{aligned} \quad (1-57)$$

For example, if $\ell = 1$, m_ℓ takes the value -1 , 0 , and $+1$. This implies that the atomic energy level under the influence of an applied magnetic field B is split into three levels. This splitting caused by a magnetic field was first discovered

by Zeeman; therefore, it is generally called the Zeeman effect or Zeeman splitting. The difference between two adjacent split levels in this case is

$$\Delta U_{or} = \mu_B B \quad (1-58)$$

The lowest level, $m_\ell = -1$, in this case, refers to the minimum potential energy corresponding to the orientation of u_{or} toward the alignment with the magnetic field B ; the highest energy level, $m_\ell = +1$, refers to the maximum potential energy corresponding to the orientation of u_{or} toward the direction opposite to B .

The Zeeman effect also occurs in spin magnetic dipoles, but in this case the potential energy level is split into only two levels: one associated with $s = -1/2$ and the other with $s = +1/2$. The difference between these two levels is

$$\Delta U_s = 2\mu_B B \quad (1-59)$$

At this point, it is desirable to go back to the classical approach because it is easier to visualize the mechanism of precession. In a magnetic field, the motion of the electrons around the nucleus is the same as that in the absence of the field, except that the angular frequency is changed by ω_p due to precession. Thus, the new angular frequency ω is

$$\omega = \omega_o - \omega_p \quad (1-60)$$

where ω_o is the angular frequency in the absence of the field. The reason for the slight decrease in angular frequency is that in the presence of the field B , the orbital electron will experience a Lorentz force $-q(\vec{v} \times \vec{B}) = -q(\omega \vec{r} B)$, as shown in Figure 1-7. This is a radially outward force, which tends to reduce the original centrifugal force by $q\omega r B$, and the new centrifugal force is given by

$$m\omega^2 r = m\omega_o^2 r - q\omega r B$$

which leads to

$$\omega_p = \frac{qB}{2m} \quad (1-61)$$

It is desirable to have some feeling for the order of magnitude of the angular frequencies. ω_o is of the order of 10^{16} radian/sec, $\omega_p \approx 1.05 \times 10^5$ H,

which would be of the order of 10^7 radians/sec for $H = 100$ amperes/m. In this case $\omega_o/\omega_p = 10^{-9}$. So, ω_p is extremely small compared to ω_o . However, the motion of the electron has been slowed down, resulting from this reduction in angular frequency. This implies that the orbital magnetic moment is also decreased by the following amount:

$$\begin{aligned} \Delta\mu_{or} &= -\frac{q\omega_o r^2}{2} - \left(-\frac{q\omega r^2}{2}\right) \\ &= -\frac{q\omega_p r^2}{2} = -\frac{q}{2} \left(\frac{qB}{2m}\right) r^2 \\ &= -\frac{q^2 B r^2}{4m} \end{aligned} \quad (1-62)$$

For the electron circulating counterclockwise, the induced magnetic moment is parallel to B , but its direction is opposite to B , as shown in Figure 1-7. It should be noted that the electron is not orbiting in one circle, but orbiting around a spherical surface. If we choose B in z direction, the orbital magnetic moment μ_{or} of a current loop is iA . The projected area of the loop on the x - y plane is $\pi\rho^2$ with a new radius, ρ . Thus, the mean square of the radius can be written as $\langle\rho^2\rangle = \langle x^2\rangle + \langle y^2\rangle$, which is the mean square of the distance of the orbiting electron normal to the magnetic field (z -axis) through the nucleus. But the orbiting electron is circulating the spherical surface; the mean square of the distance of the electron from the nucleus is $\langle r^2\rangle = \langle x^2\rangle + \langle y^2\rangle + \langle z^2\rangle$. For a spherically symmetrical distribution of the electron charge, we have $\langle x^2\rangle = \langle y^2\rangle = \langle z^2\rangle$. Thus, $\langle r^2\rangle = \frac{3}{2}\langle\rho^2\rangle$.

Suppose a material has N atoms per unit volume and Z electrons per atom, which are undergoing Larmor precession. From Equation 1-62, we can write the magnetization M by replacing r with ρ , since we refer to the electron circulating on the x - y plane, as

$$M = NZ\langle\Delta\mu_{or}\rangle = -\left(\frac{q^2 B \rho^2}{4m}\right) NZ \quad (1-63)$$

and hence, to the susceptibility as

$$\begin{aligned} \chi_m &= \frac{M}{H} = \frac{\mu_o NZ\langle\Delta\mu_{or}\rangle}{B} = -\frac{\mu_o q^2 \rho^2 NZ}{4m} \\ &= -\frac{\mu_o q^2 r^2 NZ}{6m} \end{aligned} \quad (1-64)$$

Diamagnetic susceptibility is negative, resulting in $\mu_r < 1$. For example, for $N = 10^{23} \text{ cm}^{-3}$, $Z = 2$, and $r = 1 \text{ \AA} = 10^{-8} \text{ cm}$, χ_m is of the order of -10^{-6} .

In general, a diamagnetic material does not have permanent magnetic dipoles; the induced magnetization tends to reduce the total magnetic field. This is why χ_m is negative. Materials with complete shells, such as ionic and covalent bonded crystals, are diamagnetic. Their diamagnetic behavior is due mainly to the distortion of the electron orbital motion by the external magnetic field. Dielectric solids, such as insulating polymers involving ionic and covalent bonds, are mainly diamagnetic. Diamagnetic materials generally have an even number of electrons, so the magnetic effects due to the up- and down-spins tend to cancel each other out. In a case such as hydrogen, where there is only one electron in the s orbital (or s state), the orbital motion's contribution to the magnetic effect is zero, and the diamagnetic moment is mainly that of the spin.

In some materials, such as semiconductors and metals, free electrons also contribute to diamagnetic behavior because in the presence of a magnetic field the electrons will experience a magnetic force ($qv \times B$), and their quantum

states and motion will be modified. When this happens, they will produce a local magnetic moment that tends to oppose the external magnetic field, according to Lenz's law. This contributes to a negative magnetization and a negative value of χ_m . Some values of χ_m are χ_m (copper) $= -0.74 \times 10^{-6}$; χ_m (gold) $= -3.7 \times 10^{-5}$; χ_m (silicon) $= -0.4 \times 10^{-5}$; χ_m (silicon dioxide) $= -1.5 \times 10^{-5}$. As χ_m is extremely small, we can consider all diamagnetic materials as nonmagnetic materials because $u_r = 1$ and $u = u_o$.

1.2.2 Paramagnetism

As has been mentioned, substances having an even number of electrons in the shells are inherently diamagnetic. In some substances, however, atoms have nonpaired electrons or an odd number of electrons. In such cases, the magnetic effect due to the total electron spins cannot be zero. An atom of this kind will have a permanent magnetic moment, which arises from the combination of the orbital and the spin motions of its electrons, as shown in Figure 1-8(a). The resultant magnetic moment is given by

$$u_m = g(-q/2m)J \quad (1-65)$$

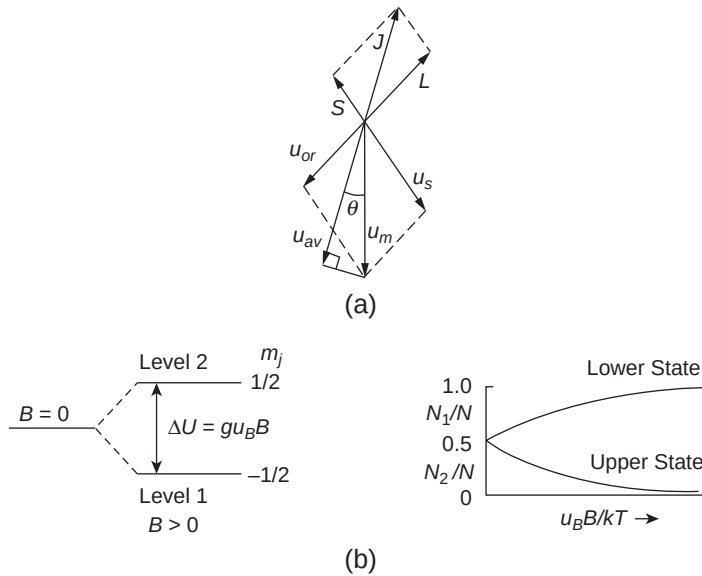


Figure 1-8 (a) The spin-orbit interaction forming a resulting magnetic dipole moment and (b) Zeeman splitting into two potential energy levels and the population of these levels as a function of temperature.

where J is the resultant angular momentum, and g is a constant known as the Lande factor. For a pure orbital magnetic moment, $g = 1$; for a pure spin magnetic moment, $g = 2$. The value of g depends on the relative orientation of both the orbital and the spin angular momenta and must be determined by quantum mechanics. Further discussion about its value is beyond the scope of this book. However, following the quantum-mechanical approach, as in Equations 1-50 and 1-52, J can be written as

$$J = m_j \hbar \quad (1-66)$$

where m_j is the resultant quantum number defining the possible ways of the combined electron orbiting and spinning motions, and j is an equivalent azimuthal quantum number. The permanent magnetic moments of atoms or molecules are randomly distributed, with just as many pointing in one direction as in another in the absence of a magnetic field. With an external magnetic field, all the momenta precess about the magnetic field B with precessional angular velocities, which results in a Zeeman splitting. Following Equation 1-57, the potential energy of the dipole can be written as

$$U = -u_m B = -gm_j u_B B \quad (1-67)$$

For a simple case with only a single spin magnetic moment, then $m_j = m_s = +1/2$ or $m_j = m_s = -1/2$, and $g = 2$, and the Zeeman splitting will produce two levels, as shown in Figure 1-8(b). The difference in potential energy between these two levels is

$$\Delta U = g(1/2)u_B B - g(-1/2)u_B B = gu_B B \quad (1-68)$$

The lower level, corresponding to $m_j = -1/2$, is associated with the magnetic dipole moment in parallel with the magnetic field B because the potential energy of the dipole is a minimum, while the upper level, corresponding to $m_j = +1/2$, is associated with the magnetic dipole moment in the direction opposite to the magnetic field. Thus, we can write the magnetization M as

$$M = \langle u_m \rangle (N_1 - N_2) \quad (1-69)$$

where $\langle u_m \rangle$ is the average dipole moment, and N_1 and N_2 are, respectively, the concentrations

of the magnetic dipoles with the direction n parallel to and that opposite to the magnetic field B in z -direction.

By denoting N as the total concentration of atoms or molecules and following the Boltzmann statistics, we can write

$$N = N_1 + N_2 \quad (1-70)$$

$$N_2/N_1 = \exp(-U/kT) = \exp(-gu_B B/kT) \quad (1-71)$$

From these two equations, we can rewrite Equation 1-69 as

$$M = N \langle u_m \rangle \tanh(gu_B B/kT) \quad (1-72)$$

The ratios of N_1/N and N_2/N are also shown in Figure 1-8(b). Obviously, the condition for $M = 0$ is either $B = 0$ or T being very high, so that $N_1 = N_2$. It should be noted that the axis of the total angular momentum J is generally not the same as that of the total magnetic moment u_m , as shown in Figure 1-8(a). However, the J axis represents the rotation axis of the physical system formed by an electron which is orbiting as well as spinning. For low magnetic fields, $gu_B B/kT \ll 1$, $\tanh(gu_B B/kT) = gu_B B/kT$, Equation 1-72 can be simplified to

$$M = N \langle u_m \rangle (gu_B B/kT) \quad (1-73)$$

Thus, the magnetic susceptibility can be expressed as

$$\chi_m = M/H = Mu_o/B = Nu_o \langle u_m \rangle gu_B/kT \quad (1-74)$$

For general cases involving both the orbital and the spin magnetic moments, the Lande factor g should be between 1 and 2, following the relation⁸

$$g = 1 + \frac{j(j+1) + s(s+1) - \ell(\ell+1)}{2j(j+1)} \quad (1-75)$$

A classical approach to evaluating χ_m is Langevin's method.⁹ Atoms or molecules with a permanent magnetic dipole moment experience a torque in the presence of a magnetic field B , tending to orient themselves toward the direction of the field. However, the ensemble of atoms or molecules will gradually attain a statistically quasi equilibrium. Langevin, in 1909,

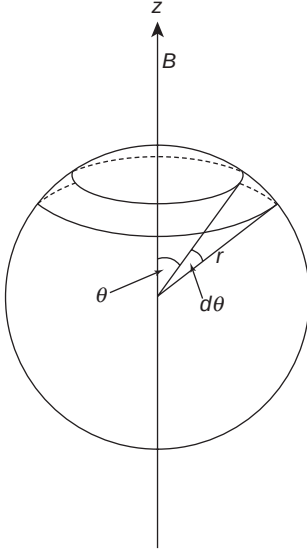


Figure 1-9 The element of the spherical shell between solid angle Ω and $\Omega + d\Omega$.
 $\Omega = \text{solid angle} = 2\pi(1 - \cos \theta)$ steradians.
 $d\Omega = 2\pi \sin \theta d\theta$.

was the first to develop a method of calculating the orientational magnetic susceptibility for paramagnetic materials. Later, Debye used the same method for its electrical analog in dipolar dielectric materials.¹⁰ The mean permanent dipole moment in the direction of the magnetic field is

$$\langle u_m \rangle = u_m \langle \cos \theta \rangle \quad (1-76)$$

where θ is the angle between the dipole moment and the magnetic field B . Thus, the potential energy of the dipole at angle θ is $-u_m B \cos \theta$. So, the probability of finding a dipole in the direction θ from the z -axis (direction of the magnetic field B) is governed by the Boltzmann distribution function

$$f(\theta) = \exp(-U/kT) = \exp(u_m B \cos \theta / kT) \quad (1-77)$$

Thus, the mean dipole moment can be determined by the following equation

$$\langle u_m \rangle = \frac{\int u_m \cos \theta f(\theta) d\Omega}{\int f(\theta) d\Omega} \quad (1-78)$$

As shown in Figure 1-9, the solid angle Ω formed by the circular cone is equal to

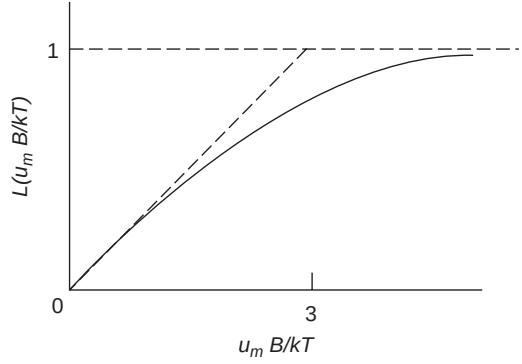


Figure 1-10 The Langevin function as a function of $u_m B / kT$.

$2\pi(1 - \cos \theta)$ steradians, so $d\Omega = 2\pi \sin \theta d\theta$. Substituting Equation 1-77 into Equation 1-78 and changing $d\Omega$ in terms of $d\theta$, we obtain

$$\langle u_m \rangle = u_m L(u_m B / kT) = u_m \langle \cos \theta \rangle \quad (1-79)$$

where $L(u_m B / kT)$ is known as the Langevin function, which is given by

$$\begin{aligned} L(u_m B / kT) &= \langle \cos \theta \rangle \\ &= \frac{\int_0^\pi \exp(u_m B \cos \theta / kT) \cos \theta \sin \theta d\theta}{\int_0^\pi \exp(u_m B \cos \theta / kT) \sin \theta d\theta} \\ &= \text{Coth}(u_m B / kT) - (u_m B / kT)^{-1} \end{aligned} \quad (1-80)$$

$L(u_m B / kT)$ as a function of $(u_m B / kT)$ is shown in Figure 1-10. For low values of $u_m B / kT$, Equation 1-80 can be simplified to

$$L(u_m B / kT) = \langle \cos \theta \rangle = u_m B / 3kT \quad (1-81)$$

From Equations 1-79 and 1-81, the magnetization M can be written as

$$M = N \langle u_m \rangle = N (u_m)^2 B / 3kT \quad (1-82)$$

and the magnetic susceptibility χ_m as

$$\chi_m = N u_o (u_m)^2 / 3kT = N u_o (g m_j u_B)^2 / 3kT \quad (1-83)$$

in which $u_m = g m_j u_B$, and $g m_j$ can be considered as the effective number of Bohr magnetons per magnetic dipole moment.

Notice that the derivation of Equation 1-74 is based on the simplest type of Zeeman

splitting, involving only two levels. If j is larger than $1/2$, then the number of levels would be $2j + 1$, and Equation 1-74 may be written as

$$\chi_m = \frac{Nu_o(gm_j u_B)(gu_B)}{3kT} \quad (1-84)$$

It can be seen that χ_m , derived on the basis of the quantum mechanical approach, is very close to that derived on the basis of the classical approach.

If $N = 10^{23}$ atoms per cm^3 and each atom has only one magnetic moment and $T = 300\text{ K}$, then $\chi_m = 0.87 \times 10^{-4}$. For any actual atoms or molecules in a real material, the value given by Equation 1-83 must be adjusted in accordance with the number of orbital and spin-formed permanent magnetic dipoles involved to give the magnetic moment u_m per atom as a whole. We can replace N with NZ , with Z denoting the number of permanent magnets per atom. For example, χ_m (aluminum) $= 0.21 \times 10^{-4}$, χ_m (platinum) $= 2.9 \times 10^{-4}$, χ_m (oxygen gas) $= 1.79 \times 10^{-6}$. Note that the magnetic effect given for paramagnetism is stronger than the inherent diamagnetic effect, which is always there. However, the resultant magnetic moment is always in the same direction as the external magnetic field, resulting in u_r as slightly larger than unity and χ_m as positive, even though its value is very small, of the order of 10^{-4} . Because of the extremely small values of χ_m , we can also consider the paramagnetic materials as nonmagnetic materials, and we can assume $u_r \approx 1$ for these materials.

1.2.3 Ferromagnetism

We have discussed, to some extent, diamagnetism and paramagnetism. The purpose of including these topics in the present book is twofold: to show what kind of materials can be considered nonmagnetic materials, and to show that the mechanisms of magnetization are similar to those of electric polarization. With a basic knowledge of magnetism, the reader may better appreciate the discussion in the later chapters about dielectric phenomena. Ferromagnetic materials are magnetic materials; the mechanisms responsible for magnetization and related

ferromagnetic behavior are complicated and must be treated quantum mechanically. To do so would be far beyond the scope of the present book. However, for completeness we shall discuss briefly, and only qualitatively, the origins of ferromagnetic behavior, which may have some bearing on later chapters.

General ferromagnetic properties are summarized as follows:

- The magnetization is spontaneous.
- The relative permeability attains very high values, as high as 10^6 in some solids.
- The magnetization may reach a saturation value by a weak magnetizing field.
- The material may have zero magnetization at zero (or very small) magnetic fields, and the magnetization remains after the removal of the magnetic field. Thus, when the applied magnetic field is changed from a positive polarity to a negative polarity, and then back to the positive polarity, a hysteresis loop will be formed in the relation of B and H , resulting in energy loss due to the rotation of the dipoles. The area enclosed by the hysteresis loop represents the energy loss involved.
- The ferromagnetic behavior (i.e., the spontaneous magnetization) completely disappears at temperatures higher than a critical temperature, called the Curie temperature T_c . At $T > T_c$, spontaneous magnetization is destroyed, and the material becomes paramagnetic, not ferromagnetic.

The commonly used ferromagnetic materials are iron (Fe), cobalt (Co), nickel (Ni), and their alloys, which are of importance to technical applications. Two other rare-earth metals, gadolinium (Gd) and dysprosium (Dy), are of less practical interest because of their low Curie temperatures. The Curie temperatures for Fe, Co, Ni, Gd, and Dy are, respectively, 1043 K, 1400 K, 630 K, 280 K, and 105 K. However, the elemental metals and their alloys are metallic in nature, and their electric conductivity is generally very high. For some applications, insulating or semi-insulating ferromagnetic materials are preferable. Some oxides, such as

CrO_2 , MnOFe_2O_3 , and FeOFe_2O_3 , have low electric conductivity and also possess spontaneous magnetization under suitable circumstances. They have a reasonable range of Curie temperatures much higher than normal room temperature.

In all of the elemental metals, spontaneous magnetization arises from an incomplete 3d shell for Fe, Co, and Ni, and an incomplete 4f shell for Dg and Dy. In this section, we use Fe as an example to discuss the mechanisms of ferromagnetic behavior. The same mechanisms apply to other elements with incomplete 3d or 4f shells. An iron atom has six electrons in the 3d shell; a fully filled 3d shell should contain ten electrons. Of these six electrons in the 3d shell, five spin in the same direction and only one spins in the opposite direction. This makes the spin magnetic dipole moment of one atom equal to $4u_B$.

Why are these six electrons in the 3d shell divided in such a way that five electrons have the same spin and only one electron has the opposite spin, producing a net spin magnetic moment of $4u_B$? In other words, why must these six electrons form a magnetized state (with five electrons in one spin and one in opposite spin), instead of having a nonmagnetized state (three electrons in one spin and the other three in opposite spin)? This can only be explained through quantum mechanics. The exchange energy principle, coupled with the Pauli's exclusion principle, shows that the potential energy of the electrons in the magnetized state is lower than when they are in the nonmagnetized state.¹¹⁻¹³ This conclusion can also be realized on the basis of band theory, by considering that the electrons with up-spin form one band and those with down-spin form another. These two bands would be completely filled if there were ten electrons in the 3d shell per atom, i.e., five electrons with up-spin and the other five with down-spin. However, iron atoms have only six electrons in the 3d shell. For the unmagnetized state, we would expect three electrons in the up-spin band and the other three in the down-spin band. In this case, both bands are partially filled. For the magnetized state, we would expect five electrons with same

spin in one band, which is completely filled, and one electron with opposite spin in the other band, which is partially filled. By comparing these two states, it can be imagined that the magnetized state is more stable than the unmagnetized one. This also implies that the potential energy is lower in the magnetized state than in the unmagnetized state, because for the former, one band is completely filled, while for the latter, each band is only partially filled. However, the difference in potential energy between these two states, referred to as the exchange energy, is small; it is of the order of 0.1 eV. Thermal agitation tends to destroy such a magnetized state (the state with a minimum potential energy) and hence, the spontaneous magnetization. Thus, we would expect that there is a critical temperature, at or above which the magnetized state would become an unmagnetized state, and spontaneous magnetization would be completely destroyed. This critical temperature is called the Curie temperature T_c . Its value is different for different magnetic materials, as mentioned previously.

Ferromagnetic materials, such as Fe, Co, and Ni, have definite types of crystalline structure, such as body-centered cubic for Fe, hexagonal for Co, and face-centered cubic for Ni. In such crystals, there are certain principal directions, which are (100), (110), and (111), assigned according to the Miller system of indices. In general, for body-centered cubic Fe, magnetization is easy in the (100) direction, medium in (110), and hard in (111). Most magnetic materials are polycrystalline, i.e., the material is composed of a large number of very tiny single crystals, called grains. Between grains, there exist grain boundaries. In each grain, there are many spin magnetic dipoles. Even their magnetic moments tend to be parallel to the easy direction, but there are six possible easy directions in a cubic unit cell: $\pm x$, $\pm y$, and $\pm z$. In order to reduce the dipole interaction energy (i.e., magnetostatic or magnetic field energy) to a minimal or lowest level, the grain must be divided into many small regions called domains. In each domain, the net spin magnetic moments are pointing in one easy direction, which is termed the parallel direction"; or in the

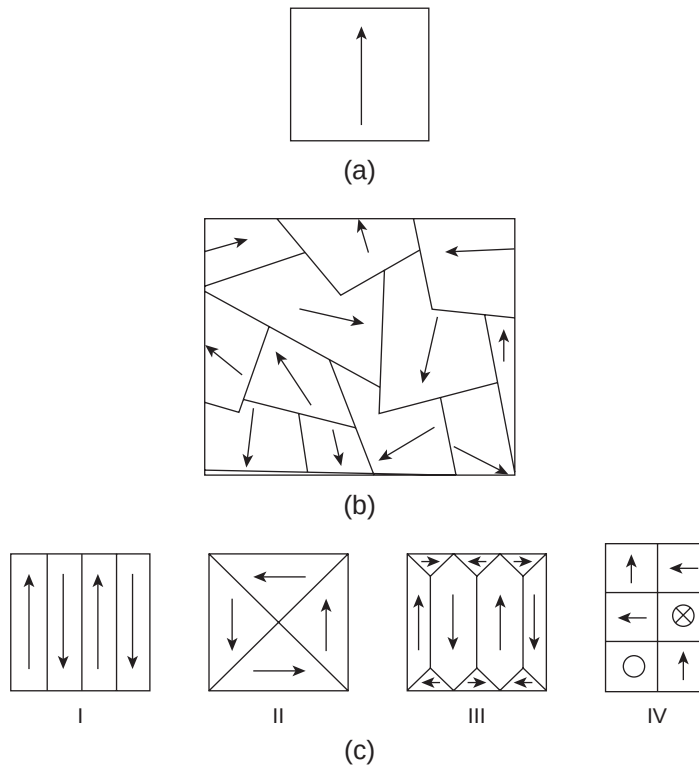


Figure 1-11 Schematic diagrams showing the directions of spin magnetic moments in the domains inside a piece of ferromagnetic material (a) net magnetic moment pointing in only one direction with a large potential energy, (b) each domain having its magnetic moment pointing in different directions with a total minimal potential energy, and (c) some possible configurations, $\otimes$ the magnetic moment pointing into the plane of the page and $\circ$ the magnetic moment pointing out of it. The potential energy is lower with the magnetic moments of neighboring domains pointing in opposite directions.

reverse easy direction, which is termed the antiparallel direction. At $T = 0$, all spins in one domain will be pointing in the parallel direction, and this domain is said to be spontaneously magnetized to the maximum degree without the aid of any external magnetic fields.

The formation of domains inside a crystal is a natural process to minimize the potential energy of the system, which contains a large number of spin magnetic dipole moments. It can be imagined that, if the crystal had only one domain with a net magnetic moment pointing in one direction, this magnetic moment would be large and produce a large external magnetic field; therefore, a large amount of energy is involved, as shown in Figure 1-11(a). However, if the crystal is divided into four or more domains with the magnetic moments in the

domains tending to cancel one another out, as shown in Figure 1-11(b), then a lower energy for the crystal can be achieved. The mutual reaction force between the dipoles pointing in one direction acts as a driving force to cause the formation of more domains, with the magnetic moments pointing different directions so that they can cancel one another, thus reducing the potential energy of the system. The partial distribution of domains tends to assume the configuration so that adjacent domains have opposite magnetic moments. Some possible configurations are shown in Figure 1-11(c).

The potential magnetic energy in Figure 1-11(c) part I is smaller than that in Figure 1-11(a), and the potential energy in Figure 1-11(c) parts III and IV is much smaller than that in Figure 1-11(c) part II. It seems that the larger

the number of domains formed, the lower the energy associated with the system. But when domains are formed, block walls (or domain walls) are required to separate the domains. Energy is required to create a wall between domains. This, in turn, results in an increase of the energy in the system. It is likely that the increase in energy for the creation of domain walls balances the decrease in energy due to the formation of domains. There must be an optimal point at which the number of domains is limited by the minimization of the total energy of the system. In other words, the number and the configuration of domains in a magnetic material are determined by the minimization of the magnetic field energy and the domain wall energy. We have already mentioned the reduction of magnetic field energy by the formation of many domains. The domain wall energy arises from both the exchange energy and the isotropy energy; the former is related to the energy difference between the unmagnetized state and the magnetized state, while the latter is related to the energy involved for magnetization in one direction relative to the crystal axes and in another direction. In a material as a whole, there are many tiny crystals (grains), and each grain has many domains. The magnetic moments are pointing randomly in all directions, resulting in a zero net magnetization, as shown in Figure 1-11(c).

It is desirable to have a general picture about the sizes of grains and domains. Suppose a piece of iron of one cubic centimeter (cm^3) has about 10^{23} atoms, which may contain about 10^4 tiny single crystals (grains), and each grain may be divided into about 10^5 domains in average. This means that in average, there are about 10^{19} atoms per grain and about 10^{14} atoms per domain. Assuming that both the grain and the domain are cubic in shape, the linear dimensions would be about $5 \times 10^{-2} \text{cm}$ for a grain and about 10^{-3}cm for a domain. Note that the sizes of the grains and the domains may be different from grain to grain and from domain to domain, and their shape is not necessarily cubic. The example given here is only to establish a feeling about the average sizes of a grain and a domain.

Let us now consider the influence of an applied magnetic field on the orientation of the magnetic moments in the domains toward the direction of the magnetic field H to produce an overall magnetization of the whole material. Suppose that we apply a magnetic field H to a nonmagnetized ferromagnetic solid, starting with zero field and gradually increasing the field in steps. We can see the gradual increase of the magnetization (or the magnetic flux density B), as shown in Figure 1-12(a). The B - H curve has four regions, each is related to different magnetization processes:

- Region I: Magnetization by translation increasing the size of the magnetized domains parallel to H and reducing the size of those with magnetization antiparallel to H , as shown in Figure 1-12(b).
- Region II: B increases in a series of jerky steps with increasing H , corresponding to jerky rotation of domains from the original “easy” crystal direction to other “easy” direction more nearly parallel to H . This phenomenon is known as the Barkhausen effect.
- Region III: The magnetic moments in all domains rotate gradually from the “easy” direction toward the direction of H .
- Region IV: Further increase in H will make the magnetic moments of all domains parallel to H , implying that the magnetization has reached a saturation value.

Now we start to reduce the magnetic field H , also in steps. As can be seen in Figure 1-12(a), B at point 5 is almost the same as at point 4, but with further decrease of H to zero, B does not go back to zero, but decreases following a different curve and reaches a finite value B_R at $H = 0$. B_R is referred to as the remanent flux density. To bring B to zero, we need a reverse magnetic field $-H_C$, which is generally called the coercive force or demagnetizing field. This is the magnetic field opposite to the magnetization field, required to bring the magnetization to the vanishing point. The formation of a hysteresis loop implies that the magnetization of a ferromagnetic material is irreversible. The

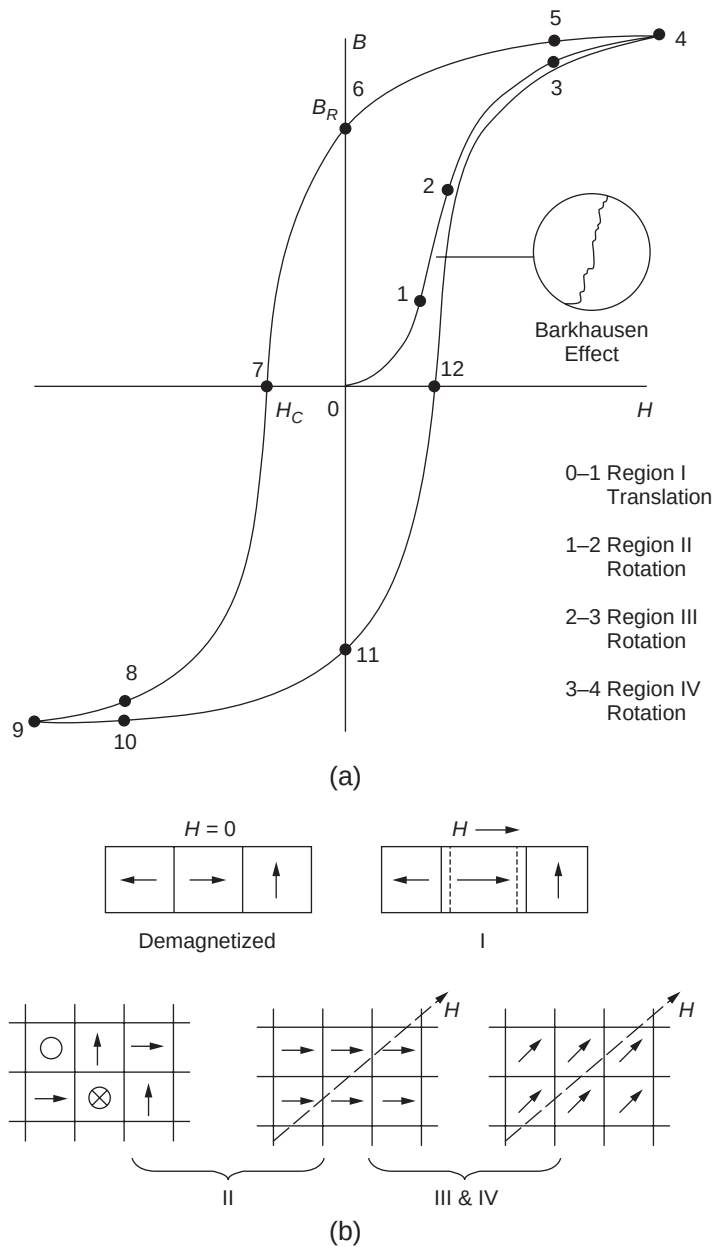


Figure 1-12 (a) The B - H hysteresis loop of a typical ferromagnetic material, and (b) the magnetization processes in Regions I, II, III, and IV.

paths 3–4 and 4–5 are almost identical, indicating that the domains' rotation from an "easy" to a "hard" direction is reversible. However, the paths 0–1–2–3 and 5–6–7 are quite different, indicating that the domain motion is irreversible because of the unavoid-

able presence of all kinds of defects: structural defects and chemical defects or impurities in a real material. The domain walls move in a series of jerky steps or in a jerky rotation due to the collision with various defects, which consume energy.

When a magnetization process consumes energy, this process will be irreversible, and this is why a hysteresis loop is formed. For the hysteresis loop around the loop 5–6–7–8–11–12–5, the hysteresis loss is equal to the loop area, which is given by

$$W_m = \oint H dB \quad \text{joules/m}^3 - \text{cycle} \quad (1-85)$$

when H is in ampere/m and B in Webers/m². The power loss is proportional to the frequency of the magnetization field. Thus, the total power loss, i.e., the total hysteresis energy loss per second, is

$$P = \left[\oint H dB \right] f \times (\text{volume of the material}) \quad (1-86)$$

where f is the frequency of the magnetic field.

Note that the magnetic moment in one direction in one domain does not change abruptly to the reverse direction in a neighboring domain because of the exchange energy, which tends to keep the angle between adjacent magnetic moments small. Thus, the domain wall is a transitional region, enabling a gradual change in the spin direction, as shown in Figure 1-13. The domain wall width W_B that minimizes the

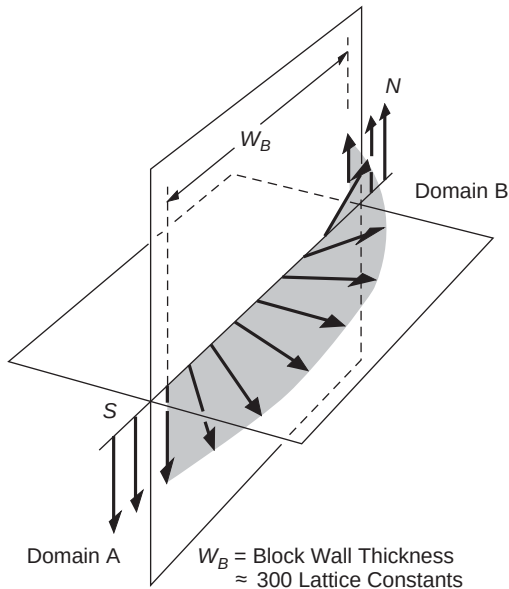


Figure 1-13 Gradual change of spin direction inside the domain wall.

total energy—exchange energy and anisotropy energy—is about 300 lattice constants (1000 Å). The wall energy involved is about 10^{-7} joule cm^{-2} for a wall separating domains magnetized in opposite directions (a 180-degree wall).

It can be imagined that the transition due to the motion of the domain walls can only be made with the expenditure of work. This constitutes the hysteresis loss for each cycle of an alternating magnetic field. In fact, hysteresis loss is a major component of power losses in electrical machines. To reduce this loss, we must use very soft magnetic materials with a very narrow B – H loop and also laminate the material in order to reduce the eddy current loss.

1.2.4 Magnetostriction

It may be said that all mechanical forces developed inside a material are originated by the interaction of electric charges. This is also true for magnetostriction, which is the magnetic counterpart of electrostriction. The spontaneous magnetization of a domain in a ferromagnetic material is always accompanied by an elongation or contraction in the direction of the magnetizing field. This physical deformation phenomenon is called magnetostriction. It is associated with the inherent anisotropy of the crystal structure and the preferred direction of magnetization in certain crystallographic directions. Magnetostriction is fundamentally a result of crystal anisotropy. Any physical deformation due to mechanical forces, such as expansion or contraction, will result in an increase in strain energy. Magnetostriction transforms electrical energy into mechanical energy in magnetostrictive transducers,⁸ but it also causes damage by introducing serious vibrations in the magnetic cores of the electrical machines and humming noise in transformers.

1.2.5 Magnetic Materials

In terms of the arrangement of spin magnetic moments in the fully magnetized state, there are four general arrangements, as shown in

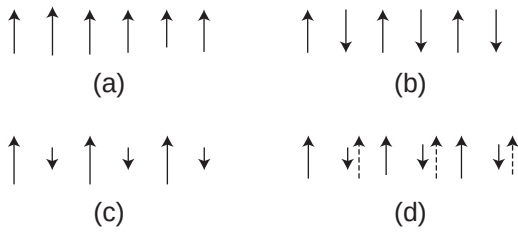


Figure 1-14 The arrangement of spin magnetic moments in four different types of magnetic materials: (a) ferromagnetic, (b) antiferromagnetic, (c) ferrimagnetic, and (d) garnet ferromagnetic (a mixture of spin and orbital magnetic moments: — spin magnetic moments, ---- orbital magnetic moments).

Figure 1-14. These general arrangements, briefly discussed, are as follows:

- In ferromagnetic materials such as Fe, Co, and Ni, the spins of neighboring atoms are parallel to each other, which would result in a great increase in potential energy of the system. To reduce this energy, the formation of many tiny domains, with their magnetic moments tending to cancel each other is necessary in nature, as explained in the previous section.
- In antiferromagnetic materials including chromium, manganese, MnO, and FeO, the spin magnetic moments alternate atom by atom as a result of the exchange interaction, which is essentially based on Pauli's exclusion principle. These materials are still magnetic from a behavioral point of view, but they do not produce an external magnetic effect, because the magnetic moments tend to cancel each other. Therefore, they are of no technical interest.
- Ferrimagnetism occurs only in compound materials. In such materials, there are two sets of spin magnetic moments that are unequal in magnitude and antiparallel to each other in direction, as shown in Figure 1-14(c). For most practical ferrimagnetic materials, the resultant spin magnetic moments are quite large. The materials behave like ferromagnetic materials but generally have a lower saturation value of magnetization. Ferrimagnetic materials are

generally referred to as ferrites. They have a very important feature, i.e., they are insulators, so ferrites are sometimes called the ferromagnetic insulators. They have very small eddy current losses, so they are suitable for high frequency applications. Ferrites are mainly transition-metal oxides. One of the commonly used ferrites is MFe_2O_3 , where M is a metal, typically Fe, Ni, Al, Zn, or Mg. If M is Fe, the material becomes FeOFe_2O_3 or Fe_3O_4 , which is the mineral magnetite, also known as lodestone, which has been exploited in navigation for centuries. It is the best known ferrite.

- The typical example of the garnet ferromagnetic materials is yttrium-iron garnet ($\text{Y}_3\text{Fe}_5\text{O}_{12}$), also known as YIG. In this material, the spin magnetic moments of the yttrium atoms are opposite to the spin magnetic moments of iron atoms, but yttrium atoms also have orbital magnetic moments, which are larger than the spin magnetic moments. Thus, this compound is ferromagnetic. The garnet structure is very tight. All the large cages between oxygen ions are occupied by positive ions. As a result, these materials are very stable compared to spinel ferrites. Furthermore, very thin garnet films can be fabricated by the epitaxial technique on a suitable substrate for memory or storage on magnetic tapes or discs.

Based on the magnitude of the coercivity, magnetic materials can also be divided into two major classes: soft magnetic materials and hard magnetic materials, as shown in Figure 1-15. Soft magnetic materials refer to the materials that can provide a large magnetic flux density at a small magnetizing field with a high differential relative permeability $(u_r)_{\text{max}}$, which is defined as

$$(u_r)_{\text{max}} = \frac{1}{\mu_0} \frac{dB}{dH} \quad (1-87)$$

as shown in Figure 1-15. Soft materials have a low coercivity (or demagnetizing field $-H_c$) and hence, a small area enclosed by the B - H loop, thus giving a low hysteresis loss per cycle of an applied magnetic field. These materials are

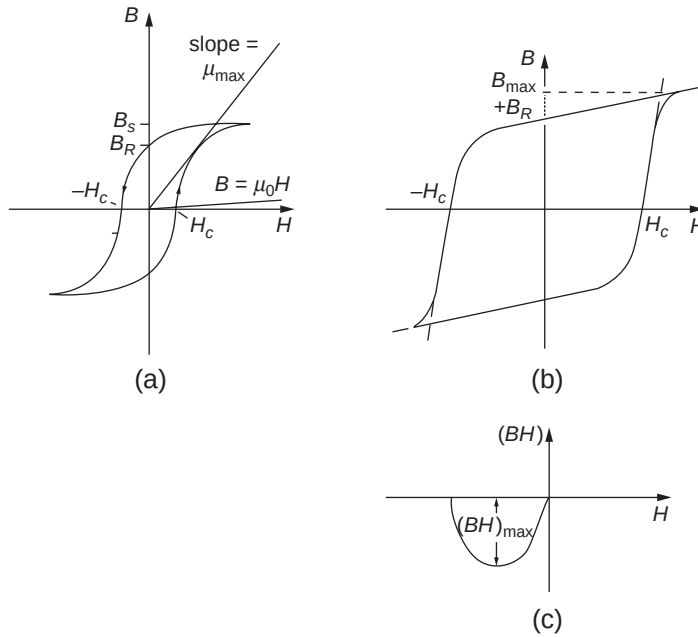


Figure 1-15 (a) Typical B - H curve for soft magnetic materials, (b) typical B - H curve for hard magnetic materials, and (c) typical BH product as a function of H for the portion of the B - H curve where B is positive and H is negative (demagnetizing).

generally used for transformer cores, electrical machines, electromagnets, inductors, etc., which require easy magnetization and low energy losses. The most commonly used soft material is not pure iron, because practical iron contains various impurities and its properties depend on its fabrication processes. In practice, to reduce the coercivity, some suitable impurities are introduced into iron to make the motion of domain walls easier. For example, for iron, the easy magnetizing direction is along the (100) crystal direction, while for nickel, the easy magnetizing direction is along the (111) crystal direction. It can be imagined that an alloy made of Fe and Ni would be magnetized in any direction between (100) and (111) crystal directions, making domain motion much easier.

Furthermore, the deformation sign of the magnetostriction of iron is opposite to that of nickel. An alloy of Fe and Ni with 70–80% of Ni is called permalloy, with a negligible net magnetostriction effect. It is a very soft material, easy to magnetize and demagnetize. Permalloys with 22% of Fe and 78% of Ni have

$\mu_r = 10^5$, and those with 16% of Fe, 79% of Ni, and 5% of Mo have μ_r as high as 10^6 . These materials are used for high-quality transformers and other electromagnetic apparatus. However, these materials have a very high electrical conductivity, and therefore, they have a high eddy-current loss, particularly at high frequencies. We need a material with high μ_r but a low electrical conductivity. Silicon-iron alloy, with about 2% of silicon to reduce the electrical conductivity, has been used for most electrical machines. Some ferrites with a very low electrical conductivity have also been used for electromagnetic apparatus. Some commonly used soft magnetic materials are listed in Table 1-1.

Hard magnetic materials are used for permanent magnets, storage media in tapes and discs, etc. In order to have a large coercivity, we need a flat B - H curve. The figure of merit generally used to characterize the hard magnetic materials is the maximum value of the B - H product, i.e., $(BH)_{\max}$ in the second quadrant of the B - H curve, as shown in Figure 1-15. The larger the

Table 1-1 Some soft magnetic materials and their properties.

Soft Magnetic Material	Curie Temperature T_C (K)	Saturated B_S (Wb m ⁻²)	Coercive H_C (Am ⁻¹)	Relative Permeability μ_r	Resistivity ρ (Ω - m)
Fe	1043	2.15	4	5×10^3	1.0×10^{-7}
Fe (with 3% Si)	1030	1.97	12	4×10^4	6.0×10^{-7}
Permalloy (22% Fe and 78% Ni)	800	1.08	4	1×10^5	1.6×10^{-7}
Supermalloy (16% Fe, 79% Ni and 5% Mo)		0.80	0.16	1×10^6	6.0×10^{-7}
Manganese-Zinc Ferrite [MnZn(Fe ₂ O ₃) ₂]	570	0.25	0.80	2×10^3	0.2

Table 1-2 Some hard magnetic materials and their properties.

Hard Magnetic Material	T_C (K)	Remanent B_R (Wb m ⁻²)	H_C (Am ⁻¹)	$(BH)_{max}$ (AWb m ⁻³)
Magnetite [FeO Fe ₂ O ₃] (Iron Ferrite—Lodestone)	850	0.27	25×10^3	
Carbon Steel (With 0.9% C and 1% Mn)		0.90	4×10^3	8×10^2
Alnico 5 (51% Fe, 8% Al, 14% Ni, 24% Co and 3% Cu)	1160	1.35	64×10^3	2×10^4
Barrium Ferrite [BaO(Fe ₂ O ₃) ₆] (Ferroxdur)	720	0.35	160×10^3	12×10^4

value of $(BH)_{max}$, the harder the magnetic material. To have a larger coercivity, the material must have a large anisotropy and contain impurities, which make domain motion difficult, so that after the removal of the magnetizing field H , the magnetized material will still provide a magnetic flux density B_R and not relax back to its original thermal equilibrium state with the magnetic moments randomly arranged. Carbon-incorporated steel and a series of alnico alloys are good hard magnetic materials. Alnico alloys are composed mainly of Fe, Al, Ni, Co, and Cu, such as alnico 5, which consists of 51% Fe, 8% Al, 14% Ni, 24% Co, and 3% Cu and yields a very high coercivity. Ferrites are also good hard magnetic materials. Some com-

monly used hard magnetic materials are listed in Table 1-2.

1.2.6 Magnetic Resonance

So far, we have dealt with only the magnetic effects under static DC or low frequency AC magnetic fields. There is much more information that can be obtained using high frequency AC magnetic fields. The term resonance generally refers to the condition in which a system involving oscillation with its natural frequency will absorb a maximum energy from a driving force source when it is driven at the frequency equal to the natural frequency. In other words, the response of a physical system to a periodic

excitation is greatly enhanced when the excitation frequency approaches the natural frequency of the system. In fact, resonance is commonly evidenced by a large oscillation amplitude, which results when a small amplitude of a periodic driving force has a frequency approaching one of the natural frequencies of the driven system. We have mentioned that the rotation of electrons and spins in the atoms, and the rotation of protons and neutrons inside the nuclei, will produce magnetic moments. Application of a magnetic field $H(= B/\mu_o)$ will cause a Larmor precession of the magnetic moments around the field axis at a Larmor frequency ω_p (Equation 1-56). It can be imagined that the Larmor precession under a static field will become gradually attenuated, but the superposition of a small alternating field will make it start a precession again around the static field axis. Magnetic resonance occurs when the frequency of the exciting field is equal to the Larmor frequency of the system (i.e., the natural frequency of the system). There are several types of magnetic resonance, depending on whether the magnetic effects are associated mainly with the spin angular momentum of nuclei or of electrons. These are nuclear magnetic resonance (NMR), electron paramagnetic resonance (EPR), electron spin resonance (ESR), nuclear quadrupole resonance (NQR), ferromagnetic resonance (FMR), antiferromagnetic resonance (AFMR), etc. Most commonly used are NMR, EPR, and ESR. In this section, we shall discuss only briefly the basic principles of NMR, EPR, and ESR, which are widely used for many diagnostic and analytical applications.

Nuclear Magnetic Resonance (NMR)

As mentioned, for ferromagnetic materials, the magnetic effect due to atomic nuclei can be neglected because it is extremely small compared to that due to electrons, which is a thousand times stronger and plays the key role in ferromagnetism. However, for nonferromagnetic materials, the magnetic dipole moment of the nucleus resulting from the intrinsic

moments of the protons and neutrons, plus the moment of the current loop formed by the orbiting protons, exhibits some features different from those due to electrons. Magnetically, all electrons are identical, but all nuclei are not. Different nuclei have different magnetic moments and spins. Following Equation 1-65, the resultant magnetic moment of a nucleus may be written as

$$u_N = g_N(q/2m_p)I \quad (1-88)$$

where g_N is the nucleus g factor analogous to the Lande g factor, m_p is the mass of the proton, and I is the resultant angular momentum of the nucleus.

Following the quantum mechanical approach, as used in Equation 1-66, the nuclear angular momentum I must be quantized in the same way as any other angular momentum. Thus, we can write

$$I = m_i \hbar \quad (1-89)$$

where m_i is the quantum number for the nuclear spin and assumes the value

$$m_i = i, (i-1), \dots, -(i-1), -i \quad (1-90)$$

where i is the spin quantum number for the nucleus. The magnetic moments and spins depend on the structure of the nucleus.^{12,14-16} The magnetic moments and the spin quantum numbers for some nuclei are listed in Table 1-3. The isotope number is the total number of protons and neutrons in the nucleus.

Similar to electrons, the number of allowed quantum energy levels is $2i + 1$. For $i = 3/2$, m_i

Table 1-3 Nuclear magnetic moments and spins for some nuclei.

Nucleus (Isotope Number)	Magnetic Moment (in Nuclear Magnetons)	Spin (i)
Hydrogen (1)	+2.79	1/2
Carbon (13)	+0.70	1/2
Oxygen (17)	-1.89	5/2
Sodium (23)	+2.22	3/2
Silicon (29)	-0.55	1/2
Phosphorus (31)	+1.13	1/2
Copper (63)	+2.22	3/2

can be $3/2$, $1/2$, $-1/2$, or $-3/2$. The potential energy of the dipole is given by

$$\begin{aligned} U_N &= u_N B = g_N (q/2m_p) I u_o H \\ &= g_N (q/2m_p) (m_i \hbar) u_o H \\ &= g_N u_{NM} m_i u_o H \end{aligned} \quad (1-91)$$

where $u_{NM} = q\hbar/2m_p$ is called the nuclear magneton, in analogy to the Bohr magneton u_B . For $i = 3/2$, the energy difference between two levels, as shown in Figure 1-16, is

$$\Delta U = g_N u_{NM} (3/2 - 1/2) u_o H = g_N u_{NM} u_o H \quad (1-92)$$

If the nucleus is excited by an alternating electromagnetic field to make a transition from one energy level to the other, the frequency of the electromagnetic field must, based on $hf = g_N u_{NM} u_o H$, be

$$f = g_N u_{NM} u_o H / h \quad (1-93)$$

Since m_p (proton mass) is much larger than m (electron mass), the frequency required for nuclear magnetic resonance is usually in the megahertz range for a suitable choice of the magnetic field $B (= u_o H)$. The basic experimental arrangement for NMR measurements is shown in Figure 1-17.

NMR has been a powerful tool for studying the physical properties of solids. It is also a very useful technique in nuclear physics, chemistry, and biology. A recent and fast-growing application of NMR is in the area of medical

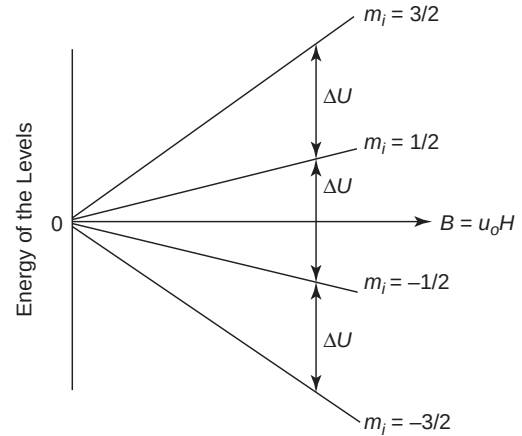


Figure 1-16 Splitting of $2i + 1$ energy levels in a magnetic field B for a nucleus with $i = 3/2$.

imaging. For example, NMR is used in measuring the concentration of protons in tissues, the decay time of absorption using short rf pulses for the diagnosis of cancer tumors, etc.

Electron Paramagnetic Resonance (EPR)

In principle, energy absorption from the high-frequency electromagnetic field in EPR is strikingly similar to that in NMR. However, in paramagnetic systems, the interaction between dipoles due to electrons is weak. So, when a material containing atoms with quantum number $j \neq 0$ is subjected to a magnetic field

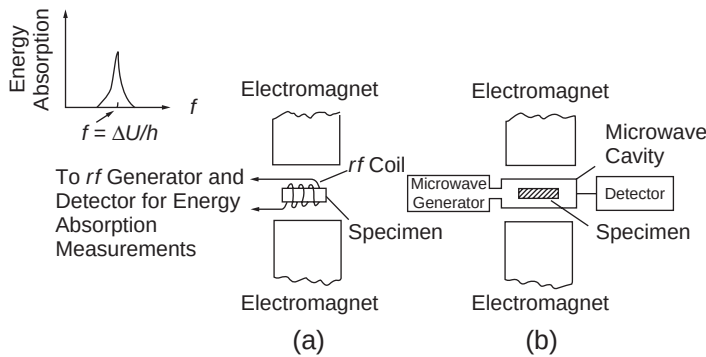


Figure 1-17 Schematic diagrams showing the basic experimental arrangements for the measurements of (a) the nuclear magnetic resonance (NMR) in the radio frequency (rf) megahertz region, and (b) the electron paramagnetic resonance (EPR) in the microwave frequency gigahertz region.

$B (= u_o H)$, all atoms will have some magnetic moments parallel to the direction of the magnetic field. The possible magnetic moment of the dipole is given by (see Equations 1-65 and 1-66).

$$u_m = g m_j u_B \quad (1-94)$$

where $m_j = j, (j-1), \dots, -(j-1), -j$, and $j = 0, 1, 2, 3, \dots$. The possible potential energy of an atom in a magnetic field is governed by Equation 1-67, depending mainly on the value of m_j . There are $2j + 1$ energy levels. For example, if $j = 1$, there are three levels corresponding to $m_j = 1$, $m_j = 0$, and $m_j = -1$. Thus, the separation between levels for $j = 1$ is

$$\Delta U = g u_B B = g u_B u_o H \quad (1-95)$$

The splitting of the $2j + 1$ energy levels for $j = 1$ is shown in Figure 1-18.

If such a paramagnetic material is subjected to a static magnetic field superposed with a small microwave field with frequency f , when the frequency reaches a value such that $hf = \Delta U$ or $f = \Delta U/h$ (which is equal to the Larmor frequency or natural frequency of the system), the incident energy will excite the dipole from a lower energy level to a higher energy level. At the same time, the energy will be absorbed by the system from the incident microwave source. There is a dip in the transmission spectrum, indicating the paramagnetic resonant absorption, as shown in Figure 1-19.

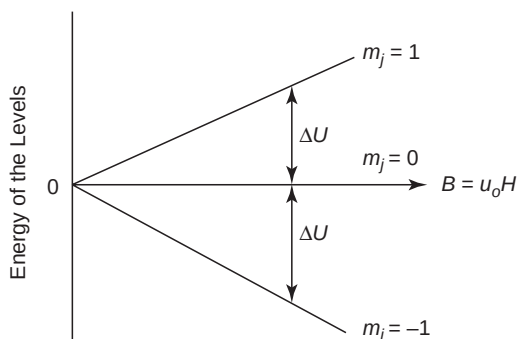


Figure 1-18 Splitting of $2j + 1$ energy levels in a magnetic field B for an electron in an atom resulting from spin-orbit interaction with $j = 1$.

The basic experimental arrangement for EPR measurements is similar to that shown in Figure 1-17, except that the range of the incident excitation frequencies for EPR is in the microwave gigahertz region simply because m for electrons is much smaller than m_p for protons.

The EPR phenomenon was first observed by Zavoisky in 1945 on the paramagnetic salt ($\text{CuCl}_2 \cdot 2\text{H}_2\text{O}$).^{17,18} Studies of paramagnetic resonance in crystalline solids provide a great deal of information about the crystalline structure.¹⁸ The technique of EPR has been widely used in physics, chemistry, biology, and other fields. It is one of the important tools employed in the analysis of matter,¹⁸⁻²⁰ such as the measurements of free radicals, trapped electrons, and excited molecules in dielectric materials. For example, there is no electron spin resonance in pure and perfect ionic crystals, since all electron shells are completely filled. Imperfections, including impurities in such crystals, however, usually create uncompensated spins, thus producing an absorption line or lines. An F color center (F refers to *Farbe*, the German word for color) in an alkali halide crystal is considered an electron trapped at a negative ion vacancy, the electron being shared by the six surrounding positive ions. Such electrons will exhibit electron spin resonance.²¹

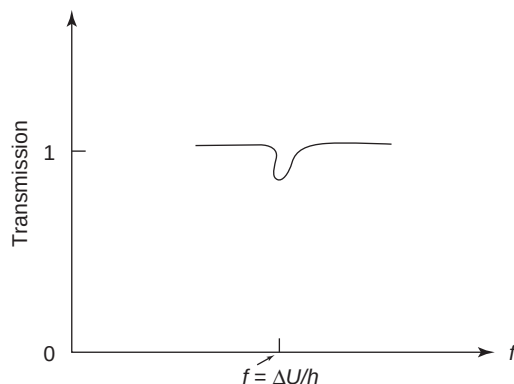


Figure 1-19 Transmission of electromagnetic waves as a function of frequency. The dip indicates the resonant absorption at $f = \Delta U/h$.

Electron Spin Resonance (ESR)

ESR can be considered a special case of EPR only if the spin of the electron plays the key role in magnetic resonance. The ESR technique is widely used as a tool to analyze chemical reactions. When chemical bonds break up, electrons may be left unpaired. Such fragments may possess spins and are usually called free radicals. The magnetic resonant absorption from the electromagnetic excitation source indicates the presence of free radicals, and the magnitude of the absorption peaks can serve as a measure of their concentration.

1.2.7 Permanent Magnets

Lodestone was possibly the first natural permanent magnet, discovered on the earth long before Christ. Chinese scientists invented the compass and called it the “south-pointing needle”—a magnetized needle, possibly made of lodestone, having the composition Fe_3O_4 (magnetite), mounted on a pivot to permit the needle to swing freely in the horizontal plane. To fabricate a permanent magnet, we first must granulate a hard magnetic material into very small particles, each no larger than the domain (about 20–100 nm). These particles are then suspended in a substance solution, which itself need not be ferromagnetic. When an external magnetic field is applied, these suspended particles will orient to make themselves parallel to the direction of the magnetic field. As soon as this magnetization process is completed, the substance solution is gradually hardened, and the magnetized particles become frozen inside the substance to form a permanent magnet. There is another fabrication process called sintering. In this process, the granular magnetic particles are first subjected to an external magnetic field, so that they are aligned in one direction. Then they are heated to a high temperature to allow them to stick together. After that, the temperature is gradually lowered to consolidate the stuck particles, forming a solid permanent magnet.

The most commonly used material for permanent magnets is alnico 5 (see Table 1-2). A permanent magnet with tiny, needle-like parti-

cles of Fe in a nickel and aluminum matrix can be formed by controlled heat treatment and by annealing in a magnetic field. Single domain particles with a high uniaxial anisotropy should have a large coercivity. Barium ferrites called Ferroxdur [$\text{BaO}(\text{Fe}_2\text{O}_3)_6$] belong to this kind of large coercivity material. These materials have been used to form ceramic permanent magnets by sintering (or bonding in a matrix) the finely powdered barium ferrites under a magnetic field. Under this guideline, many new materials have recently been developed for sophisticated permanent magnets, including samarium-cobalt (SmCo_5), $\text{Nd}_2\text{Fe}_{14}\text{B}$, etc. The permanent magnets formed by single domain particles are often referred to as ESD magnets because the particles are elongated single domain particles.^{22–25}

1.2.8 Magnetic Memories

The discussion of various applications of magnetic materials is far beyond the scope of this book. However, it is worth mentioning a few, such as the use of magnetic materials for storing information. It is also good for the reader to think whether the dielectric materials can have a function similar to or better than magnetic materials. In fact, the sales of audio and video tapes, computer discs, and other magnetic information products in the United States exceeded those for all other high-technology products. Storage capacity has been greatly improved in the past decade, an upper-bound density has reached about 10^{14} bits m^{-2} . The basic principle for storage of information as patterns of magnetization in a magnetic material has been well established for more than 100 years. The main difference between the old and the new magnetic memories lies in technology—in materials and recording methods.

In general, to write is simply to use an electromagnet moving relative to the recording tape or disc. After recording (writing), to read is to move the tape or disc that has already been stored with information. This moving tape or disc will give rise to varying magnetic flux, which will induce voltages in a coil and convert the signal to the original information. For tapes

and discs, magnetic particles of sizes substantially less than one micron, with carefully tailored magnetic properties, are deployed in a passive binder on a flexible polymeric sheet, such as polyethylene terephthalate (PET or trade name Mylar) or on a rigid aluminum disc. To promote high coercivity, it is necessary to make a certain magnetization direction preferable to the others (easy axis). Particles commonly used for tapes and discs are γ iron oxides including γ Fe_2O_3 , Co γ Fe_2O_3 , CrO_2 , $\text{BaO}(\text{F}_2\text{O}_3)_6$, etc. Synthetic γ iron oxide is the oxide containing γ iron that has a face-centered cubic structure. γ iron oxides can be used to produce needle-shaped particles. For more information about this field, see references 23–25.

Semiconductor memories are presently preferred for short-term information storage, mainly because of low cost, easy handling capacity, fast access time, and small size. Magnetic tapes and discs, on the other hand, are chosen for long-term, large-scale information storage, and particularly, for the fact that no electrical energy is needed to retain information storage. However, conventional tapes and discs do have some shortcomings: recording and playback heads are easily worn down, and coercivity decreases as packing density increases. Thin film technologies were developed to reduce the bit size so as to increase the bit density and make the reader easier to read by putting the thin film medium closer to the head. About four decades ago, small, bulk ferromagnetic ferrite cores were dominantly employed as computer memory elements; now, magnetic films deposited in the presence of a magnetic field exhibit a square hysteresis loop capable of switching from one state to another as a bistable element. Thin magnetic films consisting of Co, Ni, and Pt; Co, Cr, and Ta; or Co (75%), Cr (13%), and Pt (12%) are frequently used for hard disc memories. They may be deposited on an aluminum substrate by means of vapor deposition, sputtering, or electroplating. Such films are subsequently covered with a thin carbon layer about 40 nm thick for lubrication and protection against corrosion. These films can have coercivities ranging from 60 to 120 kA m⁻¹ and are capable of providing a density of 1.8 Mbits mm⁻².

Thin films can also be used for optical recording. In fact, magneto-optical recording relies on thermomagnetic effects. Suppose that a small region of a magnetic film is heated to a temperature higher than its Curie temperature and then allowed to cool in the presence of an external magnetic field. This small region would become magnetized to the direction of the magnetic field. If this heating and cooling process were carried out by means of a focused laser beam, the whole operating process could be very fast, within a microsecond. This could be considered as the “write” step in a recording system. The information stored in the magnetized regions of a thin film can be read by using the Kerr or Faraday magneto-optic effects by means of polarized light. By analyzing the direction of rotation of the polarization plane, it is easy to find the magnetized regions that hold the stored information. Several amorphous alloys, including TbFe, GdCo, GdFe, GdTbFe, GdTbCo, etc., have been used for magneto-optic memories because they have a larger Kerr rotation angle and hence a better signal-to-noise ratio.^{25–27}

There is another magnetic storage device, which is based on magnetic bubbles. This device was first demonstrated in 1967. A great deal of research was carried out for a while, but this technology is presently not much in use. Single crystal mixed garnets, such as yttrium-iron-garnet ($\text{Y}_3\text{Fe}_5\text{O}_{12}$), are used for magnetic bubble media. Basically, a thin film of a few μm in thickness, consisting of such a magnetic material, can be epitaxially grown on a suitable substrate. Such a film has a highly uniaxial and anisotropic feature, implying that there is an easy axis for the generation of bubbles, perpendicular to the plane of the film. In this film are tiny cylindrical domains, each about 1 μm in diameter. All the domains can be aligned in a weak magnetic field normal to the film. However, by applying a strong, localized magnetic field in the opposite direction, it is possible to produce a cylindrical domain called a magnetic bubble with its magnetic axis inverted. Such a bubble will be stable after the strong field is removed because of its large domain-wall coercivity. If the north pole of the

bubble is on the top surface of the film, a small, south-pole magnetic field placed near the bubble will attract it, causing it to move. This small magnetic field can be produced simply by employing soft magnetic permalloy overlays on the top surface of the film. The actual devices are fairly complicated. Our intention here is to describe briefly the basic principle of how magnetic bubbles can be generated, persuaded to move from one place to another, and erased by magnetic fields. The crystals are transparent to red light. Thus, by viewing the transparent film through crossed polarizers, it is easy to find the domain from the direction of rotation of the plane of polarization (Faraday effect in transmission or Kerr effect in reflection). Each such domain constitutes one bit of stored information. For more details about magnetic bubbles, see references 26–30.

1.3 Electromagnetic Waves and Fields

There are two kinds of waves: longitudinal and transverse. In longitudinal waves, the oscillatory movement is in the direction of wave propagation, as in sound waves; in transverse waves, the oscillatory movement is perpendicular to the direction of wave propagation, as in waves on the surface of water. An electromagnetic wave is a transverse wave with its electric field F and magnetic field H perpendicular to each other and also perpendicular to the direction of wave propagation, as shown in Figure 1-20.

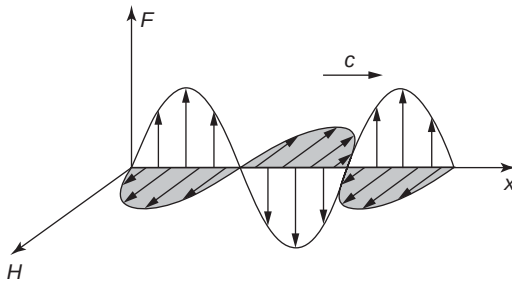


Figure 1-20 Schematic diagram of a traveling plane electromagnetic wave.

From Maxwell's equations, the following can be seen:

- Wherever there exists a time-varying electric field, there must also exist a time-varying magnetic field, and vice versa.
- An electrostatic field does not induce a magnetic field, and a magnetostatic field does not induce an electric field.
- Electric flux lines terminate on charges.
- Magnetic flux lines do not end, and their associated magnetic fields are solenoidal.

In general, space containing air only is essentially equivalent to free space, as far as electromagnetic wave propagation is concerned. We consider the medium to be air in which $\epsilon = \epsilon_0$, $u = u_0$, $\sigma = 0$ and $\rho = 0$. The variables F and H produce each other, so, for simple presentations, we can keep one variable and eliminate the other by the following process:

Taking the curl of Equation 1-2, we have

$$\begin{aligned}\nabla \times (\nabla \times F) &= \nabla(\nabla \cdot F) - \nabla^2 F \\ &= -\mu_0 \frac{\partial}{\partial t} (\nabla \times H)\end{aligned}\quad (1-96)$$

Since $\rho = 0$, $\nabla \cdot F = 0$. Substitution of Equation 1-1 into Equation 1-96 yields

$$\nabla^2 F = \epsilon_0 \mu_0 \frac{\partial^2 F}{\partial t^2} = \frac{1}{c^2} \frac{\partial^2 F}{\partial t^2} \quad (1-97)$$

This is the wave equation for free space (or in air medium), and c is the velocity of light, which is $(\epsilon_0 \mu_0)^{-1/2}$. Equation 1-97 provides information about the electric field F at any position and at any time. The amplitude of its electric field intensity vector will vary with respect to x , y , z and t if Cartesian coordinates are employed. For simplicity, if we consider only the y component F_y and assume that there is no variation in H with respect to y and in F with respect to z directions, then Equation 1-97 reduces to

$$\frac{\partial^2 F_y}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 F_y}{\partial t^2} \quad (1-98)$$

The general solution of Equation 1-98 is

$$F_y = f_1(x + ct) + f_2(x - ct) \quad (1-99)$$

If we are interested principally in sinusoidal function, then the solution of Equation 1-98 can be expressed simply as

$$F_y = A \cos(\omega t - kx) \quad (1-100)$$

where k is the wave factor, which is given by

$$k = \frac{\omega}{v} = \frac{\partial \pi f}{\lambda f} = \frac{2\pi}{\lambda} \quad (1-101)$$

which depends on the wavelength or the velocity of the electromagnetic wave. In other words, it depends on ϵ and μ , which are inherent in the properties of matter. Equation 1-100 represents a component of F in y direction, which varies with time. Because the variation is perpendicular to the direction of wave propagation, which is in x direction, the wave is a transverse wave, as shown in Figure 1-20. It is also a plane wave, because a surface passing through all points of equal phase would be a plane, which is perpendicular to the x direction for the present case.

From Equations 1-1 and 1-100, it can easily be shown that the associated magnetic field in z direction is given by

$$H_z = A(\epsilon_0/u_0)^{1/2} \cos(\omega t - kx) \quad (1-102)$$

The ratio of F_y to H_z is the intrinsic impedance of the medium

$$Z = \frac{F_y}{H_z} = \left(\frac{\mu_0}{\epsilon_0} \right)^{1/2} \quad (1-103)$$

For free space, $Z = 377$ ohms. Both the electric field and the magnetic field represent the stored energy of the electromagnetic wave. By considering an enclosed surface with a volume V , the energy stored in the electric field is

$$U_F = \frac{\epsilon_0}{2} \int_V F^2 dV \quad (1-104)$$

and the energy stored in the magnetic field is

$$U_H = \frac{u_0}{2} \int_V H^2 dV \quad (1-105)$$

By denoting the power conveyed by the electromagnetic wave or power flow in the direction of wave propagation per unit area and per unit time as P , the total power flow out of the enclosed surface per second can be expressed

as $\int_a P da$. The rate of decrease of the stored energy would be $\frac{d}{dt} \left[\frac{1}{2} \int_V (\epsilon F^2 + \mu H^2) dV \right]$, which must be equal to the outflow of the power. Thus,

$$\int_a P da = -\frac{d}{dt} \left[\frac{1}{2} \int_V (\epsilon F^2 + \mu H^2) dV \right] \quad (1-106)$$

From Equations 1-1 and 1-2 it can easily be shown that

$$\int_a P da = \int_V \nabla \cdot (F \times H) dV \quad (1-107)$$

Based on Gauss' law

$$\int_V \nabla \cdot (F \times H) dV = \int_a (F \times H) da \quad (1-108)$$

Thus, the power flow per unit area can be written as

$$P = F \times H \quad (1-109)$$

This is called the Poynting vector, showing that the direction of the power flow is the same as the direction of wave propagation.

Wave theory can explain many optical phenomena, such as reflection, diffraction, and interference, but it cannot explain phenomena related to the exchange of energy, including the emission and absorption of light, photoelectric effects, etc. According to Einstein and Planck, electromagnetic wave or light energy is emitted or absorbed in multiples of a certain minimum energy particle, which is called a quantum or a photon.^{31,32} Depending on the frequency of the electromagnetic wave, the energy of a quantum or a photon is given by

$$E = h\nu = h \frac{c}{\lambda} \quad (1-110)$$

where ν is the frequency. (Note that the frequency is sometimes denoted by f .) It is generally accepted that electromagnetic waves or light apparently has a dual nature. Depending on the situation, the particle nature dominates when dealing with the exchange of energy, while the wave nature dominates when dealing with reflection, diffraction, etc. Figure 1-21 shows the electromagnetic spectrum. Equation 1-110 can be written for the photon energy in free space in a simple form as

$$E = \frac{1.24}{\lambda} \quad (1-111)$$

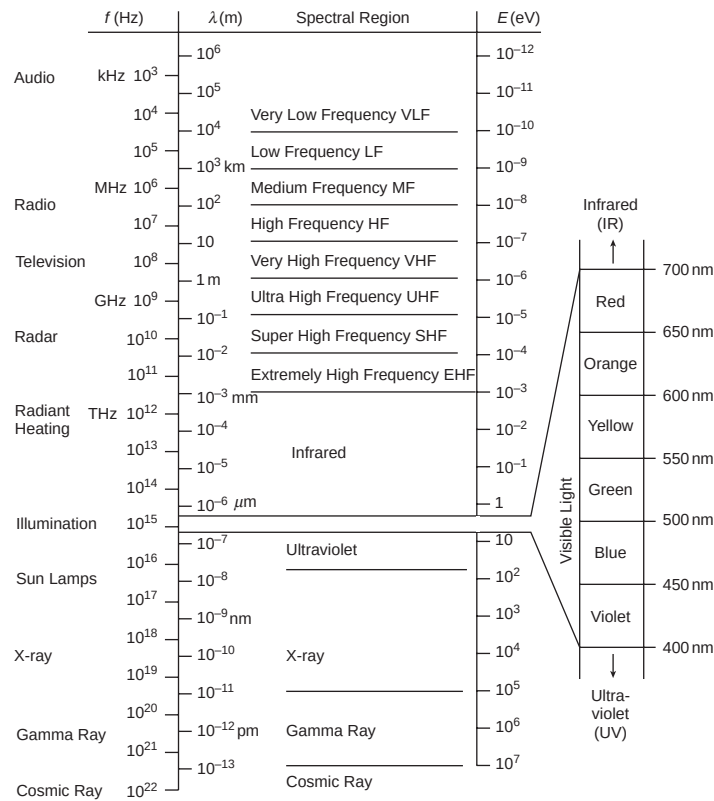


Figure 1-21 Electromagnetic spectrum. The wavelength λ depends on the medium in which the electromagnetic wave propagates. In this figure, the medium is air, so the wavelength is for $\epsilon = \epsilon_0$ and $u = u_0$.

in which E is in eV and the wavelength λ in μ m.

1.4 Dimensions and Units

Length, mass, time, and electric charge are generally used as the fundamental dimensions for describing physical quantities. On the basis of the International System of Units (*Système Internationale*, SI), the units for the four fundamental dimensions are

L (length): m (meter)
 M (mass): kg (kilogram)
 T (time): s (second)
 Q (electric charge): C (coulomb)

1.4.1 Length

The unit for length is m (meter), but sometimes it is more convenient to use smaller units, particularly for microscopic dimensions. The commonly used smaller units are

1 cm (centimeter) = 10^{-2} m
 1 mm (millimeter) = 10^{-3} m
 1 μ m (micron) = 10^{-6} m = 10^{-4} cm
 1 nm (nanometer) = 10^{-9} m = 10^{-7} cm
 1 Å (Angstrom) = 10^{-10} m = 10^{-8} cm
 1 pm (picometer) = 10^{-12} m = 10^{-10} cm

Some old publications sometimes use British units: inch, foot, or yard. The conversion of British units to SI units is

$$\begin{aligned}
 1\text{ inch} &= 0.0254\text{ m} = 2.54\text{ cm} \\
 1\text{ foot} &= 12\text{ inches} = 0.3048\text{ m} = 30.48\text{ cm} \\
 1\text{ yard} &= 3\text{ feet} = 0.9144\text{ m} = 91.44\text{ cm} \\
 1\text{ mile} &= 5280\text{ feet} = 1609.344\text{ m}
 \end{aligned}$$

1.4.2 Mass

The unit for mass is kg (kilograms). Again, it is sometimes more convenient to use smaller units. Those most commonly used are

$$\begin{aligned}
 1\text{ g (gram)} &= 10^{-3}\text{ kg} \\
 1\text{ mg (milligram)} &= 10^{-6}\text{ kg} = 10^{-3}\text{ g} \\
 1\mu\text{g (microgram)} &= 10^{-9}\text{ kg} = 10^{-6}\text{ g} \\
 1\text{ ng (nanogram)} &= 10^{-12}\text{ kg} = 10^{-9}\text{ g}
 \end{aligned}$$

British units for mass are ounce and pound. The conversion is as follows:

$$\begin{aligned}
 1\text{ ounce (avoirdupois)} &= 28.3495\text{ g} \\
 &= 28.3495 \times 10^{-3}\text{ kg} \\
 1\text{ pound (avoirdupois)} &= 16\text{ ounces} = 453.59\text{ g} \\
 &= 0.45359\text{ kg} \\
 &= 1\text{ lb (from Latin } \textit{Libra}) \\
 1\text{ ton (short ton, S.T.)} &= 2000\text{ pounds} \\
 &= 907.18\text{ kg} \\
 1\text{ ton (long ton, L.T.)} &= 2240\text{ pounds} \\
 &= 1016.04\text{ kg}
 \end{aligned}$$

1.4.3 Time

The unit for time is s (second). The commonly used smaller units are

$$\begin{aligned}
 1\text{ ms (millisecond)} &= 10^{-3}\text{ s} \\
 1\mu\text{s (microsecond)} &= 10^{-6}\text{ s} \\
 1\text{ ns (nanosecond)} &= 10^{-9}\text{ s} \\
 1\text{ ps (picosecond)} &= 10^{-12}\text{ s}
 \end{aligned}$$

1.4.4 Electric Charge

The fourth unit is C (coulomb). Smaller units are also frequently used, namely μC (micro-coulomb), nC (nanocoulomb), and pC (pico-coulomb). The smallest existing charge, called the elementary charge, is the charge of an electron, which is $-1.602 \times 10^{-19}\text{ C}$. We use q to denote this charge

$$q = 1.602 \times 10^{-19}\text{ C}$$

So, an electron has a negative charge $-q$ and a proton has a positive charge $+q$.

1.4.5 Derived Units

Many derived units are also generally used to shorten the dimensional expressions for many physical parameters, including the ampere, volt, ohm, etc. The derived units can be easily expressed in terms of the fundamental units. However, it is important to understand the meaning of the derived units when analyzing any dielectric phenomena.

1. Current I: The unit is ampere.

$$1\text{ A (Ampere)} = 1\text{ Cs}^{-1} \quad (\text{T}^{-1}\text{Q})$$

2. Force F: The unit is newton.

$$1\text{ N (newton)} = 10^5\text{ dynes} = 1\text{ kg ms}^{-2} \quad (\text{LMT}^{-2})$$

3. Pressure X: The unit is Pa (pascal).

$$1\text{ Pa (pascal)} = 1\text{ newton m}^{-2} = 1\text{ kg m}^{-1}\text{ s}^{-2} \quad (\text{L}^{-1}\text{MT}^{-2})$$

One conventional unit frequently used in the literature is torr, which is

$$\begin{aligned}
 1\text{ torr} &= 1\text{ mm Hg} = 1.01325/760\text{ bar} \\
 1\text{ bar} &= 10^6\text{ dynes cm}^{-2} \\
 1\text{ torr} &= 1.01325 \times 10^6 / 760 = 1330\text{ dynes cm}^{-2} \\
 &= 133\text{ Pa}
 \end{aligned}$$

Standard temperature and pressure (STP) refers to the condition at one atmosphere (760 mm Hg) and 270 K (0°C).

4. Energy E: The unit is J (joule).

$$\begin{aligned}
 1\text{ J (joule)} &= 1\text{ Nm} = 1\text{ kg m}^2\text{ s}^{-2} \quad (\text{L}^2\text{MT}^{-2}) \\
 1\text{ J (joule)} &= 10^7\text{ erg}
 \end{aligned}$$

One most frequently used unit is eV (electron-volt).

$$\begin{aligned}
 1\text{ eV} &= 1.602 \times 10^{-19}\text{ coulomb} \times 1\text{ V} \\
 &= 1.602 \times 10^{-19}\text{ joule} \\
 1\text{ keV} &= 10^3\text{ eV} \\
 1\text{ MeV} &= 10^6\text{ eV}
 \end{aligned}$$

There is another unit frequently used by chemists, calorie, called the thermo-chemical unit.

$$1 \text{ cal (calorie)} = 4.184 \text{ J (joule)}$$

$$1 \text{ J (joule)} = 0.2389 \text{ cal (calorie)}$$

Thermal energy is also expressed in eV, such as kT (at 300 K) = 0.0259 eV. British thermal unit is called the btu, 1 btu = 252 calorie.

5. Power P: The unit is W (watt).

$$1 \text{ W (watt)} = 1 \text{ J s}^{-1} = 1 \text{ kg m}^2 \text{ s}^{-3} \quad (\text{L}^2 \text{MT}^{-3})$$

6. Electric Potential (voltage) V: The unit is V (volt).

$$1 \text{ V (volt)} = 1 \text{ W A}^{-1} = 1 \text{ C}^{-1} \text{ kg m}^2 \text{ s}^{-2} \quad (\text{L}^2 \text{MT}^{-2} \text{Q}^{-1})$$

7. Resistance R: The unit is Ω (ohm).

$$1 \text{ (ohm)} = 1 \text{ V A}^{-1} = 1 \text{ C}^{-2} \text{ kg m}^2 \text{ s}^{-1} \quad (\text{L}^2 \text{MT}^{-1} \text{Q}^{-2})$$

8. Conductance G: The unit is S (siemens).

$$1 \text{ S (siemens)} = 1 \Omega^{-1} = 1 \text{ C}^2 \text{ kg}^{-1} \text{ m}^{-2} \text{ s} \quad (\text{L}^{-2} \text{M}^{-1} \text{TQ}^2)$$

9. Electric flux ψ : The unit is C (coulomb) or electric flux lines.

$$\text{Electric flux density D: The unit is } \text{C m}^{-2} \text{ or electric flux lines per m}^2. \quad (\text{L}^{-2} \text{Q})$$

10. Capacitance C: The unit is F (farad).

$$1 \text{ F (farad)} = \text{QV}^{-1} = 1 \text{ C}^2 \text{ kg}^{-1} \text{ m}^{-2} \text{ s}^2 \quad (\text{L}^{-2} \text{M}^{-1} \text{T}^2 \text{Q}^2)$$

11. Magnetic flux ϕ : The unit is Wb (weber).

$$1 \text{ Wb (weber)} = \text{Vs} = 1 \text{ C}^{-1} \text{ kg m}^2 \text{ s}^{-1} \quad (\text{L}^2 \text{MT}^{-1} \text{Q}^{-1})$$

Magnetic flux density (or called magnetic induction) B: The unit is T (tesla).

$$1 \text{ T (tesla)} = 1 \text{ Wb m}^{-2} = 1 \text{ C}^{-1} \text{ kg s}^{-1} \quad (\text{MT}^{-1} \text{Q}^{-1})$$

B is frequently expressed in cgs unit, that is gauss.

$$1 \text{ gauss} = 10^{-4} \text{ Wb m}^{-2} = 10^{-8} \text{ Wb cm}^{-2} \quad (\text{MT}^{-1} \text{Q}^{-1})$$

12. Inductance L: The unit is H (henry).

$$1 \text{ H (henry)} = 1 \text{ Wb A}^{-1} = 1 \text{ C}^{-2} \text{ kg m}^2 \quad (\text{L}^2 \text{MQ}^{-2})$$

A list of SI units for some physical parameters frequently encountered is given in Table 1-4.

1.4.6 Cgs System of Units and Cgs/SI Conversion

The basic units of the cgs system are cm (centimeter), g (gram), and s (second). But, the fourth unit depends on whether it is based on the cgs electrostatic system of units (esu) or based on the cgs electromagnetic system of units (emu). For the esu system, the permittivity in free space is assigned to be 1 (unra-

Table 1-4 SI units for some physical parameters frequently used.

Physical Parameter	Symbol	SI Unit	Dimensions
Temperature	T	K (kelvin)	K
Frequency	f or ν	Hz	s^{-1} (T^{-1})
Force	$\mathcal{F}$	N (newton)	kg/m/s^{-2} (LMT^{-2})
Pressure	X	Pa (pascal)	N/m^{-2} ($\text{L}^{-1} \text{MT}^{-2}$)
Energy	E	J (joule)	N/m ($\text{L}^2 \text{MT}^{-2}$)
Power	P	W (watt)	J/s^{-1} ($\text{L}^2 \text{MT}^{-3}$)
Current	I	A (ampere)	C/s^{-1} ($\text{T}^{-1} \text{Q}$)
Electric Potential	V	V (volt)	W/A^{-1} ($\text{L}^2 \text{MT}^{-2} \text{Q}^{-1}$)
Resistance	R	Ω (ohm)	V/A^{-1} ($\text{L}^2 \text{MT}^{-1} \text{Q}^{-2}$)
Conductance	G	S (siemens)	A/V^{-1} ($\text{L}^{-2} \text{M}^{-1} \text{TQ}^2$)
Electric Flux	ψ	C (coulomb)	C (Q)
Electric Flux Density	D	C/m^{-2} (coulomb/m ²)	C/m^{-2} ($\text{L}^{-2} \text{Q}$)
Capacitance	C	F (farad)	Q/V^{-1} ($\text{L}^{-2} \text{M}^{-1} \text{T}^2 \text{Q}^2$)
Magnetic Flux	ϕ	Wb (weber)	V/S ($\text{L}^2 \text{MT}^{-1} \text{Q}^{-1}$)
Magnetic Flux Density	B	T (tesla)	Wb/m^{-2} ($\text{MT}^{-1} \text{Q}^{-1}$)
Inductance	L	H (henry)	Wb/A^{-1} ($\text{L}^2 \text{MQ}^{-2}$)

tionalized dielectric constant), the fourth basic unit is statcoulomb, and each unit is designated by prefixing the syllable *stat* to the name of the corresponding unit. For the emu system, the permeability μ_0 in free space is assigned to be 1 (unrationalized magnetic constant), the fourth basic unit is abampere, and most units are designated by prefixing the syllable *ab* to the name of the corresponding units, exceptions are the maxwell, gauss, oersted, and gilbert.

In the stat-cgs (esu) system, $\epsilon_0 = 1$ (unrationalized), the electric flux is defined in such a manner that 4π lines of electric flux emanate from each statcoulomb of charge. That is

$$\psi = 4\pi Q_{\text{stat}}$$

Thus, the rationalized stat-cgs (esu) system is characterized by the fact that the electric flux density is expressed as statcoulomb per cm^2 . Based on this relation, we can write the electric flux density D in terms of mks-(SI) units and that in terms of stat cgs-(esu) units as follows:

$$\frac{D(\text{in coulomb m}^{-2})}{D(\text{in cgs lines cm}^{-2})} = \frac{\epsilon_0 \epsilon_r (\text{in volts m}^{-1})}{1 \epsilon_r F (\text{in statvolts cm}^{-1})}$$

or

$$\begin{aligned} \epsilon_0 (\text{in mks-SI units}) &= \frac{\text{coulombs}}{4\pi \text{ statcoulombs}} \frac{\text{statvolts}}{\text{volts}} \frac{\text{centimeters}}{\text{meters}} \\ &= \frac{1}{36\pi \times 10^9} \text{ farad meter}^{-1} \\ &= 8.854 \times 10^{-12} \text{ Fm}^{-1} \\ &= 8.854 \times 10^{-2} \text{ C}^2 \text{ kg}^{-1} \text{ m}^{-3} \text{ s}^2 \quad (\text{L}^{-3} \text{ M}^{-1} \text{ T}^2 \text{ Q}^2) \end{aligned}$$

The following is the transformation of the quantities between mks (SI) and cgs (esu) units:

$$\text{Charge: } \frac{\text{statcoulomb}}{\text{coulomb}} = 3 \times 10^9$$

$$\text{Resistance: } \frac{\text{statohm}}{\text{ohm}} = \frac{1}{9 \times 10^{11}}$$

$$\text{Voltage: } \frac{\text{statvolt}}{\text{volt}} = \frac{1}{300}$$

$$\text{Capacitance: } \frac{\text{statfarad}}{\text{farad}} = 9 \times 10^{11}$$

$$\text{Current: } \frac{\text{statampere}}{\text{ampere}} = 3 \times 10^9$$

In the unrationalized ab-cgs (emu) systems, $\mu_0 = 1$. Following the same procedure, we can write the magnetic flux density B in terms of mks-(SI) units and that in terms of ab cgs-(emu) units as follows:

$$\begin{aligned} \frac{B(\text{in webers m}^{-2})}{B(\text{in maxwells cm}^{-2})} &= \frac{\mu_0 \mu_r H(\text{in amperes m}^{-1})}{1 \mu_r H(\text{in gilberts cm}^{-1})} \\ &= \frac{\text{webers}}{\text{maxwells}} \frac{\text{gilberts}}{\text{amperes}} \frac{\text{centimeters}}{\text{meters}} \end{aligned}$$

or μ_0 [in mks-(SI) units]

$$\begin{aligned} &= 4\pi \times 10^{-7} \text{ henry per meter} \\ &= 1.254 \times 10^{-6} \text{ Hm}^{-1} = 1.254 \times 10^{-6} \text{ C}^{-2} \text{ kg m} \\ &\quad (\text{LMQ}^{-2}) \end{aligned}$$

The following is the transformation of the quantities between the mks (SI) and the cgs (emu) units:

$$\text{Charge: } \frac{\text{abcoulomb}}{\text{coulomb}} = 10^{-1}$$

$$\text{Capacitance: } \frac{\text{abfarad}}{\text{farad}} = 10^{-9}$$

$$\text{Voltage: } \frac{\text{abvolt}}{\text{volt}} = 10^8$$

$$\text{Magnetic flux: } \frac{\text{maxwell}}{\text{weber}} = 10^8$$

$$\text{Current: } \frac{\text{abampere}}{\text{ampere}} = 10^{-1}$$

$$\begin{aligned} \text{Magnetic flux density:} \\ \frac{\text{gauss}(\text{maxwell cm}^{-2})}{\text{weber m}^{-2}} &= 10^4 \end{aligned}$$

$$\text{Resistance: } \frac{\text{abohm}}{\text{ohm}} = 10^9$$

$$\begin{aligned} \text{Magnetomotive force (mmf):} \\ \frac{\text{abampere-turn (gilbert)}}{\text{ampere-turn}} &= \frac{4\pi}{10} \end{aligned}$$

$$\text{Inductance: } \frac{\text{abhenry}}{\text{henry}} = 10^9$$

$$\begin{aligned} \text{Magnetic field strength:} \\ \frac{\text{abampere cm}^{-1} (\text{oersted})}{\text{ampere m}^{-1}} &= \frac{4\pi}{10^3} \end{aligned}$$

Based on the above relations, we obtain

$$\frac{\epsilon_o(esu)}{\epsilon_o(emu)} = c^2$$

$$\frac{u_o(esu)}{u_o(emu)} = \frac{1}{c^2}$$

1.4.7 The Unit of Debye

The unit of debye is frequently used in the literature to give a numerical value of dipole moments. One debye is one unit of dipole moment formed by positive and negative charge of the charge equal to q (electron charge), which is 4.803×10^{-10} statcoulomb, and the positive and negative's charge separation of $0.2\text{\AA} = 2 \times 10^{-9}\text{ cm}$. Thus,

$$1 \text{ debye} = 4.803 \times 10^{-10} \text{ statcoulomb} \\ \times 2 \times 10^{-9} \text{ cm} \approx 10^{-18} \text{ esu}$$

In terms of SI units, we have

$$1 \text{ debye} = 1.602 \times 10^{-19} \text{ coulomb} \times 2 \times 10^{-11} \text{ m} \\ = 3.2 \times 10^{-30} \text{ cm} \approx 1/3 \times 10^{-29} \text{ cm}$$

1.4.8 The Chemical Unit of Mole

The definition of Avogadro's number of $6.022 \times 10^{23}/\text{mole}$ is the number of atoms or molecules per one gram atomic weight. For one gram atomic weight of hydrogen with atomic weight of one gram, one mole of hydrogen contains 6.022×10^{23} hydrogen atoms. For one gram atomic weight of oxygen with atomic weight of 16 grams, one mole of oxygen also contains 6.022×10^{23} oxygen atoms. Similarly, for one gram atomic weight of silicon with atomic weight of 28 grams, one mole of silicon still contains 6.022×10^{23} silicon atoms. Thus, one mole of silicon oxide (SiO_2) with a molecular weight of $28 + 2 \times 16 = 60$ grams contains 6.022×10^{23} SiO_2 molecules.

To transform this number into practical units in terms of the number of atoms or the number of molecules per cm^3 , we need to know the density of the material ρ (grams per cm^3). For example, the number of silicon atoms per cm^3 is

$$N_{\text{Si}} = (6.022 \times 10^{23} \text{ mole}^{-1} / 28 \text{ grams mole}^{-1}) \\ \times 2.33 \text{ g cm}^{-3} = 5 \times 10^{22} \text{ atoms/cm}^3$$

Similarly, the number of SiO_2 molecules per cm^3 is

$$N_{\text{SiO}_2} = (6.022 \times 10^{23} \text{ mole}^{-1} / 60 \text{ grams mole}^{-1}) \\ \times 2.22 \text{ g cm}^{-3} \\ = 2.23 \times 10^{22} \text{ molecules/cm}^3$$

where 2.33 g cm^{-3} and 2.22 g cm^{-3} are, respectively, the densities of Si and SiO_2 .

References

1. H.H. Skilling, *Exploring Electricity* (Ronald Press, New York, 1948).
2. M. Faraday, Phil. Trans. 128, 1, 79, 265 (1837/38).
3. J.C. Maxwell, *Treatise on Electricity and Magnetism* (Dover, New York, 1954).
4. R.P. Feynman, R.B. Leighton, and M. Sands, *The Feynman Lectures in Physics* (Addison-Wesley, New York, 1989).
5. W.R. Smythe, *Static and Dynamic Electricity* (McGraw-Hill, New York, 1950).
6. J.A. Stratton, *Electromagnetic Theory* (McGraw-Hill, New York, 1941).
7. A.R. von Hippel, *Molecular Science and Molecular Engineering* (John Wiley and Sons, New York, 1959).
8. C. Kittel, *Introduction to Solid State Physics*, 5th Edition (John Wiley and Sons, New York, 1976).
9. P. Langevin, *J. Phys.* 4, 678 (1905).
10. P. Debye, *Polar Molecules* (Dover, New York, 1945).
11. R.M. Bozorth, *Rev. Mod. Phys.*, 19, 29 (1947) and also *Ferromagnetism* (Van Nostrand, New York, 1951).
12. A.H. Morrish, *Physical Principles of Magnetism* (John Wiley and Sons, New York, 1965).
13. C. Kittel and J.K. Galt, "Ferromagnetic Domains," *Solid State Physics*, 3, 437 (1956).
14. A. Abragam, *Nuclear Magnetism* (Oxford University Press, Oxford, 1961).
15. G.E. Pake, "Nuclear Magnetic Resonances" in *Solid State Physics*, Ed. by F. Seitz and D. Turnbull, Vol. 2 (Academic Press, New York, 1956).
16. R.T. Schumacher, *Introduction to Nuclear Magnetic Resonance* (W.A. Benjamin Publishers, New York, 1970).
17. E.K. Zavoisky, *J. Phys. U.S.S.R.*, 9, 211, 245, and 447 (1945), and also 10, 170, and 197 (1946).

18. B. Bleany and K.W.H. Stevens, Rep. Progr. Phys., 16, 108 (1955).
19. G.E. Pake, *Paramagnetic Resonance* (W.A. Benjamin Publishers, New York, 1962).
20. W. Low, in "Paramagnetic Resonance in Solids," *Solid State Physics Supplement 2* (Academic Press, New York, 1962).
21. A.F. Kip, C. Kittel, R.A. Levy, and A.M. Portis, Phys. Rev., 91, 1066 (1953).
22. H.H. Stadelmaier and E. Th. Honig, Permanent Magnetic Materials, J. of Metals, 43, 32 (1991).
23. R.E. Hummel, *Electronic Properties of Materials*, 2nd Edition (Springer-Verlag, New York, 1993).
24. J.C. Anderson, *Magnetism and Magnetic Materials* (Chapman and Hall, London, 1968).
25. N. Braithwaite and G. Weaver, *Electronic Materials* (Butterworth-Heinemann, London, 1990).
26. J.C. Mallinson, *The Foundation of Magnetic Recording*, (Academic Press, New York, 1987).
27. M. Ohring, *The Materials Science of Thin Films* (Academic Press, New York, 1992).
28. A.H. Bobeck and E. Della Torre, *Magnetic Bubbles*, (North-Holland, Amsterdam, 1975).
29. G.S. Almasi, Proc. IEEE., 61, 438 (1972).
30. E.A. Giess, *Science*, 208, 938 (1980).
31. A.R. von Hippel, *Dielectrics and Waves* (John Wiley and Sons, New York, 1954).
32. E. Goldin, *Waves and Photons: An Introduction to Quantum Optics*, (John Wiley and Sons, New York, 1982).

2 Electric Polarization and Relaxation

The task of science is both to extend the range of our experience and to reduce it to order. Only by experience itself do we come to recognize those laws which grant us a comprehensive view of the diversity of phenomena.

The classical theories have contributed so fundamentally to our knowledge of atomic structure.

Neils Bohr

2.1 Fundamental Concepts

We can say that all dielectric phenomena arise from an electric force which is primarily due to the attraction and the repulsion of electric charges. The action of the force tends to reduce the potential energy of the whole electrically stressed system to a minimum. This is a universal law of nature. Photocopies stick together, dust collects on television screens, dirty particles adhere to outdoor power transmission line insulators, etc.: Such phenomena are caused by this force. In the following sections, we shall discuss the origin of this force.

2.1.1 Electric Charge Carriers and Their Motion

The most important electric charges are electrons and protons, which are the major elementary particles of any atom of any material. Both the electron and the proton are the elementary charges; they are equal in magnitude, but opposite in charge polarity. An electron has a negative charge $-q$, and a proton has a positive charge $+q$, where $q = 1.602 \times 10^{-19}$ C.

An electric field can be considered a region in which a particle with a charge q would experience an electric force acting on it. This force, denoted by $\mathcal{F}$, is given by

$$\mathcal{F} = qF \quad (2-1)$$

where F is the electric field strength at the point the particle is placed. The electric field strength

F is determined by the magnitudes and the locations of all other charges, as well as the surrounding medium. For a single-point charge with total net charge $+Q$ in free space, F can be written as

$$F = \frac{Q}{4\pi\epsilon_0 r^2} \quad (2-2)$$

where r is the distance between the location of $+Q$ and a point P , and ϵ_0 is the permittivity of the medium. From Equations 2-1 and 2-2, it can be seen that the force acting on the charge q can either be attractive or repulsive, depending solely on the relative polarity of Q and q , as shown in Figure 2-1.

In a neutral atom or molecule, or any neutral particle, the number of electrons is equal to the number of protons. When such a neutral entity loses one electron, it becomes a positive ion (i.e., the entity has one proton which is not neutralized by the electron); when the entity receives an electron, it becomes a negative ion. Free ions are just like free electrons and will move in the presence of an electric field. A charged particle moving due to an electric force will have an acceleration a , based on Newton's law, given by

$$a = \frac{\mathcal{F}}{M} = \frac{qF}{M} \quad (2-3)$$

Since the charged particle is accelerated in the direction of the field, it exhibits a kinetic energy

$$\frac{1}{2} Mv^2 = \int \mathcal{F} dr = \int qF dr \quad (2-4)$$

where M is the mass, v is the velocity of the particle, (i.e., charge carrier), and r is the length of the path along which the particle has traveled. In a nonuniform field, such as that produced by a point charge $+Q$, the velocity of the charge carrier varies from point to point. Supposing that the charge carrier is an electron and it is placed at point P at $t = 0$, $v = 0$, and $r = r_1$, then its velocity at point P' where $r = r_2$ can be expressed as

$$v = \left[\frac{2q}{m} \int F dr \right]^{1/2} = \left[\frac{2qQ}{4\pi\epsilon_0 m} \int_{r_1}^{r_2} \frac{dr}{r^2} \right]^{1/2} \quad (2-5)$$

$$= \left[\frac{qQ}{2\pi\epsilon_0 m} \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \right]^{1/2}$$

It can be seen from Equations 2-4 and 2-5 that the velocity and the kinetic energy of an electron or a negative ion increase as it travels toward a convergent field, and decrease when it travels in a divergent field.

Suppose now we place a neutral particle with a permittivity ϵ_2 in a nonuniform field such as

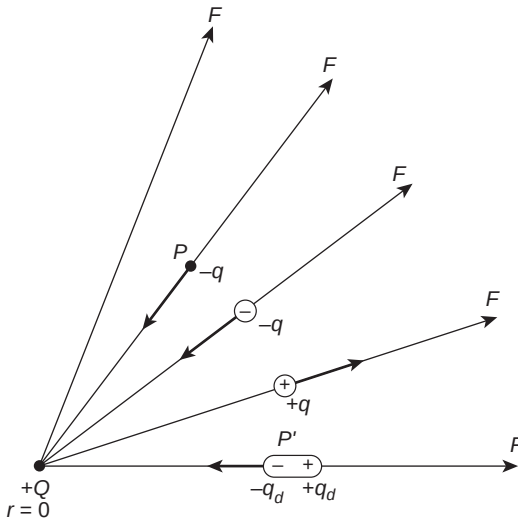


Figure 2-1 Nonuniform electric field F produced by a single point charge $+Q$, and the force between $+Q$ and various charged particles.

- Electron with a charge $-q$.
- ⊖ Negative ion with a charge $-q$.
- ⊕ Positive ion with a charge $+q$.
- ⊕⊖ Dipole formed by a polarized neutral particle with a net negative charge $-q_d$ on one end and a net positive charge of $+q_d$ on the other at a separation of δ .
- ⇒ The direction of the force.

that produced by a point charge $+Q$ shown in Figure 2-1. In this case, the neutral particle will be polarized, shifting the normally symmetrical electron clouds of atoms or orienting existing permanent dipoles, if any, toward the direction of the field. Electric polarization processes will be discussed in Section 2.3. This neutral particle, polarized by the field, will become an overall large dipole consisting of two charges equal in magnitude but opposite in polarity, $+q_d$ and $-q_d$, separated by a small distance δ . At point P' , where this dipole is located, the charge $-q_d$ will experience an attractive force given by

$$\mathcal{F}_1 = qF = -q_d \frac{Q}{4\pi\epsilon_0 r^2} \quad (2-6)$$

and the charge $+q_d$ will also experience a force, which is repulsive, given by

$$\mathcal{F}_2 = qF = +q_d \frac{Q}{4\pi\epsilon_0 (r+\delta)^2} \quad (2-7)$$

In this case, $\mathcal{F}_1$ is larger than $\mathcal{F}_2$ by an amount equal to

$$\Delta\mathcal{F} = |\mathcal{F}_1 - \mathcal{F}_2| = \left| \frac{q_d Q}{4\pi\epsilon_0} \left(\frac{1}{r^2} - \frac{1}{(r+\delta)^2} \right) \right| \quad (2-8)$$

$$= \left| \frac{q_d Q}{2\pi\epsilon_0} \right| \frac{\delta}{r^3}$$

since r is much larger than δ .

This force always tends to move the particle from the weak field region to a strong field region. It should be noted that, apart from this force, there is another force arising from the release of some stored potential energy of the system: in the present case, the permittivity of the particle $\epsilon_2 > \epsilon_0$. This force also helps to move the particle to a strong field region because this action would reduce the potential energy of the system. This is why dust is attracted to television screens and dirty particles are collected on the surface of power transmission line insulators, which may lead to a very harmful effect. This force depends on the relative magnitude of the permittivity of the particles and that of the medium. We shall discuss this force further in a later section.

Now let us consider the case with a constant uniform electric field F . Such a uniform field can be produced simply by connecting a steady DC voltage supply across two parallel metallic

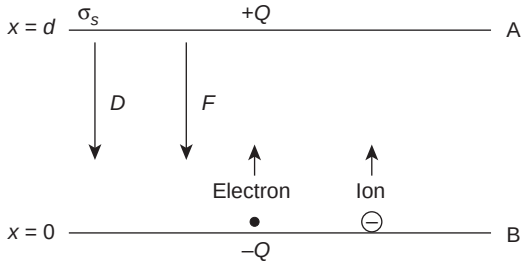


Figure 2-2 Uniform electric field F produced inside a capacitor by the uniformly distributed charge $+Q$ on plate A and $-Q$ on plate B, the separation between two parallel plates being d and the area of each plate being A .

plates and charging this system as a capacitor. After charging, the supply is disconnected and removed from the system, leaving positive charge $+Q$ on plate A and a negative charge $-Q$ on plate B, as shown in Figure 2-2. The electric field created by the charges on the plates can be considered to be uniform in the region between the plates if the edge effect is ignored. This field F can be written as

$$F = \frac{Q}{\epsilon A} = \frac{\sigma_s}{\epsilon} = \frac{D}{\epsilon} \quad (2-9)$$

where ϵ is the permittivity of the medium, A is the area of the plates, σ_s is the surface charge density on the plates, and D is the electric flux density in the medium, which is equal to σ_s if the medium is free space. Now, if we place an electron or a negative ion at a point very close to $x = 0$, as shown in Figure 2-2, this charge carrier will move freely from $x = 0$ to $x = d$ if the medium is free space with $\epsilon = \epsilon_0$. This carrier will be accelerated along the field direction without encountering any collision. Thus, we can write

$$\begin{aligned} \mathcal{F} &= qF = Ma \\ a &= \frac{qF}{M} \end{aligned} \quad (2-10)$$

where M is the mass of the carrier, and $M = m$ if it is an electron. The acceleration a is equal to dv/dt , thus the velocity of the carrier can be expressed as

$$v = \int_0^t a dt = at \quad (2-11)$$

where a is constant since F is constant, but v increases linearly with traveling time t . Since v

increases with t , the time required to travel a unit distance decreases as x increases from 0. Suppose it takes t_d time for the carrier to travel to $x = d$ from $x = 0$, then the velocity of the carrier at $x = d$ can be expressed as

$$v_d = at_d \quad (2-12)$$

Thus, the average velocity can be written as

$$\bar{v} = \frac{at_d}{2} \quad (2-13)$$

Since $t_d = d/\bar{v}$, the average velocity can be expressed in the form

$$\bar{v} = \left(\frac{qFd}{2M} \right)^{1/2} \quad (2-14)$$

The mobility of the carrier is defined as the velocity per unit electric field strength, so the average mobility of the carrier $\bar{\mu}$ is the average velocity per unit field strength. Thus,

$$\bar{\mu} = \frac{\bar{v}}{F} = \left(\frac{qd}{2MF} \right)^{1/2} \quad (2-15)$$

The average velocity of the carrier in this case is proportional to $(F)^{1/2}$, while the average mobility of the carrier is inversely proportional to $(F)^{1/2}$.

Now we will consider the case for solids, such as semiconductors or dielectric materials. In this case, there are about 10^{22} to 10^{23} atoms or molecules per cm^3 . It can be imagined that in solids, a charge carrier will not move freely as in free space. It will suffer many, many collisions with phonons and impurities during its movement from $x = 0$ to $x = d$ in the same electrode configuration shown in Figure 2-2. Based on Equation 2-9, for the same charges on plates A and B, F will be smaller by a factor of

$\frac{\epsilon}{\epsilon_0} = \epsilon_r$. An electron (or a hole) will react with the lattice vibrations (i.e., phonons) and other imperfections in the material, and its quantum state is governed by Pauli's exclusion principle. To take into account the effects of the inertia due to all the interactions and to make the motion of electrons or holes to follow the classical Newton's law, we use the effective mass m^* instead of the conventional electron mass m_0 in free space, which is practically equal to the rest mass of the electron because, for most

cases, the electron velocity is much smaller than the light velocity, so the relativistic effect can be ignored.

The electric field exerts on the electron an electric force qF , but the electron encounters a friction force (called an inertia force or a viscous force) due to the collisions with phonons and imperfections (structural defects and chemical defects or impurities). Assuming that this friction force, which opposes the electric force, has the form m^*v/τ , where v is the drift velocity of the electron and τ is the collision time, then based on Newton's law, we can write

$$m^* \frac{dv}{dt} = qF - m^* \frac{v}{\tau} \quad (2-16)$$

The friction force tends to retard the motion of the electron. In the steady state condition, $dv/dt = 0$, and Equation 2-16 reduces to

$$v = \frac{q\tau}{m^*} F = \mu F \quad (2-17)$$

where μ is called the electron mobility, which is defined as the drift velocity per unit field strength

$$\mu = dv/dF = \frac{q\tau}{m^*} \quad (2-18)$$

Electrons in a solid also have their thermal velocity v_{th} , due to thermal agitation. The direction of thermal velocity is random, and hence the random motion of electrons does not contribute to electrical conductivity. Drift velocity leads the electrons to move in the direction of the electric field and hence contributes to the electrical conductivity, which is defined as the electrical current density J per unit electric field strength. Thus, the electrical conductivity σ is given by

$$\sigma = \frac{J}{F} = \frac{qn v}{F} = q\mu n \quad (2-19)$$

where n is the concentration of electrons.

Both thermal velocity and drift velocity are temperature dependent. Thermal velocity is given by

$$v_{th} = \left(\frac{2kT}{3m^*} \right)^{1/2} \quad (2-20)$$

which is proportional to $(T)^{1/2}$. It is usually much larger than the drift velocity. The temperature dependence of drift velocity or mobility is mainly due to the temperature dependence of τ . The collision time τ is sometimes called the mean free time or relaxation time. It is really the time between collisions. Based on the quantum mechanics, an electron has a wave character. Thus, a wave passing through a perfect periodic lattice should travel freely without scattering (or collision).¹ However, the lattice is far from perfect periodicity because thermal agitation causes lattice vibrations (or atomic vibrations), which generate phonons, and also various defects (structural and chemical) in the material form scatterers to collide with electrons. In general, the probability of electron collision with phonons increases with increasing temperature, while the probability of electron collision with defects decreases with increasing temperature. That the mobility decreases with increasing temperature is mainly due to the decrease of τ with temperature.

Ions are much larger in size than electrons. It can be imagined that it would be much more difficult for ions to move in solids. Small ions may be able to move by jumping from one interstitial site to the next, but large ions must move from one lattice site to the next, which must be vacated for their occupation. Thus, the probability of ion movement depends on the probability of the creation of a vacancy next to the moving ion. For large ions, the activation energy for their movement involves several eV. Donors or acceptors in semiconductors or insulators, when ionized, become positive or negative ions, respectively. These ions, for most practical cases, cannot move at normal temperature. Charge carrier transport processes will be discussed in Ionic Conduction and Electronic Conduction in Chapter 7.

2.1.2 Electromechanical Effects

In an electrically stressed system, there is always a force tending to reduce the stored potential energy of the system. For a simple capacitor, shown in Figure 2-2, the force tends to increase the capacitance in order to reduce

the potential between the two charged plates so as to reduce the potential energy, since the decrease in V would result in a decrease in potential energy, which is equal to $QV/2$ or $CV^2/2$. If the medium inside the capacitor is air, such a force would tend to attract any substance with a permittivity larger than ϵ_0 into the capacitor to replace the air medium. In the following discussion, it is assumed for mathematical simplicity that the dielectric system is isotropic and linear, and heat losses and other effects, such as thermal agitation and gravitation, are ignored. If the initial charges in the system are fixed, then any mechanical work done for the reduction of the potential energy of the system must be at the expense of the stored energy in the system. The change of the stored potential energy ΔU in the system, due to the introduction of a dielectric body of volume V_2 and permittivity ϵ_2 into a dielectric system of volume V_1 and permittivity ϵ_1 , can be evaluated by

$$\Delta U = \frac{1}{2} \int (\epsilon_2 - \epsilon_1) F_1 F_2 dV \quad (2-21)$$

where F_1 is the electric field in the region of the dielectric medium in the system of permittivity ϵ_1 where a dielectric body of permittivity ϵ_2 is subsequently introduced, and F_2 is the electric field in the dielectric body itself.² The stored potential energy of the system after the introduction of the body into the system is decreased when ΔU is positive, and increased when ΔU is negative. Using Equation 2-21, we have derived the electromechanical forces resulting from the change of the stored energy of the system for the following cases.³

Before presenting our analysis of the cases, it is desirable to review briefly some formulas about the field distribution inside and outside a dielectric body of permittivity ϵ_2 placed in a medium of permittivity ϵ_1 , which is stressed in a uniform static electric field F_1 . The field inside the dielectric body F_2 depends on the geometric shape of the body. The formulas for some common configurations shown in Figure 2-3 are summarized as follows.⁴⁻⁶

1. Medium 1 of permittivity ϵ_1 and medium 2 of permittivity ϵ_2 separated by a plane boundary.

Case i (a): The field in the media parallel (or tangential) to the boundary, as shown in Figure 2-3(a). For this case, the tangential fields in the two media are equal.

$$F_{2t} = F_{1t} \quad (2-22)$$

Case i (b): The field in the media normal (or perpendicular) to the boundary, as shown in Figure 2-3(b). For this case, the relation of the fields in both sides is given by

$$F_{2n} = \frac{\epsilon_1}{\epsilon_2} F_{1n} \quad (2-23)$$

Case i (c): The field inclined with an incident angle θ_1 incident on the boundary, as shown in Figure 2-3(c). For this case, we have

$$\begin{aligned} F_{1t} &= F_1 \sin \theta_1, & F_{1n} &= F_1 \cos \theta_1 \\ F_{2t} &= F_2 \sin \theta_2, & F_{2n} &= F_2 \cos \theta_2 \end{aligned}$$

Based on Equations 2-22 and 2-23, we have

$$\frac{\tan \theta_1}{\tan \theta_2} = \frac{\epsilon_1}{\epsilon_2}$$

This is the law of refraction (see Wave Theory in Chapter 3).

2. A dielectric sphere with a radius a . For this case, the field inside the sphere is given by

$$F_2 = \frac{3\epsilon_1}{2\epsilon_1 + \epsilon_2} F_1 \quad (2-24)$$

$F_2 < F_1$ for $\epsilon_2 > \epsilon_1$, as shown in Figure 2-3(d) and $F_2 > F_1$ for $\epsilon_2 < \epsilon_1$, as shown in Figure 2-3(e).

3. A dielectric ellipsoid with its major axis $2c$ and 2 equal minor axes $2a$ and $2b$, i.e., a prolate spheroid with $c > a = b$.

Case iii (a): The field in the medium parallel to the major axis. For

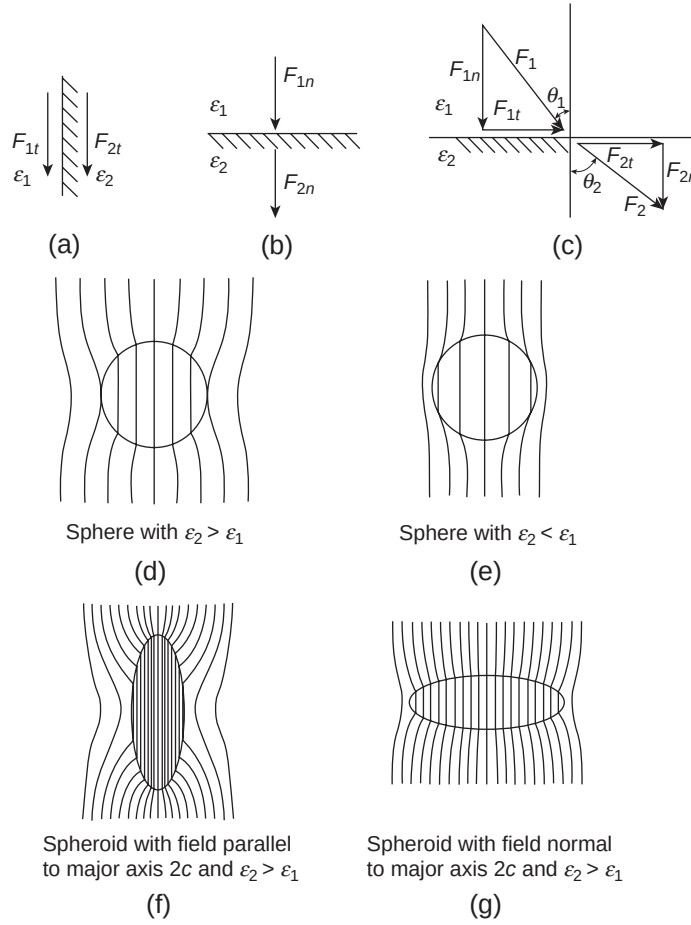


Figure 2-3 Electric field distribution in two media separated by a boundary. (a) Tangential field parallel to a plane boundary, (b) normal field perpendicular to a plane boundary, (c) inclined field with an incident angle θ_1 incident on a plane boundary, (d) a sphere of radius a having permittivity ϵ_2 in a medium having permittivity ϵ_1 with $\epsilon_2 > \epsilon_1$, (e) same as (d) but with $\epsilon_2 < \epsilon_1$, (f) a spheroid of major axis $2c$ having permittivity ϵ_2 in a medium of permittivity ϵ_1 with $\epsilon_2 > \epsilon_1$ and the field parallel to the major axis $2c$, and (g) same as (f) but with the field normal to the major axis $2c$.

this case, the field inside the spheroid is given by

$$F_2 = \frac{\epsilon_1}{\epsilon_1 - (\epsilon_1 - \epsilon_2)G_{//}} F_1 \quad (2-25)$$

where $G_{//}$ is given by

$$G_{//} = \frac{a^2 c}{2} \int_0^\infty \frac{ds}{(s+c^2)^{3/2}(s+a^2)} \quad (2-26)$$

In this case, $G_{//} < \frac{1}{3}$, $F_2 < F_1$ for $\epsilon_2 > \epsilon_1$, as shown in Figure 2-3(f).

Case iii (b): The field in the medium normal to the major axis. For this case, the field inside the spheroid is given by

$$F_2 = \frac{\epsilon_1}{\epsilon_1 - (\epsilon_1 - \epsilon_2)G_{\perp}} F_1 \quad (2-27)$$

where $G_{\perp}$ is given by

$$G_{\perp} = \frac{ac^2}{2} \int_0^\infty \frac{ds}{(s+a^2)^{3/2}(s+c^2)} \quad (2-28)$$

In this case, $G > 1/3$, $F_2 < F_1$ for $\epsilon_2 > \epsilon_1$, as shown in Figure 2-3(g).

With these formulas, we can now analyze the following cases.

The Force Acting on the Boundary between Two Different Dielectric Materials

At the boundary between two different materials 1 and 2, the electric flux is usually refracted, and the tangential components of the electric field F_{1t} and F_{2t} , as well as the normal components, F_{1n} and F_{2n} , follow the relationship based on Equations 2-22 and 2-23, which are

$$F_{1t} = F_{2t} \quad (2-29)$$

$$F_{1n} = \epsilon_2 F_{2n} / \epsilon_1 \quad (2-30)$$

where ϵ_1 and ϵ_2 are, respectively, the permittivities of materials 1 and 2. According to Equation 2-21, there is a force acting on the boundary with the tendency of expanding material 2 into material 1 if $\epsilon_2 > \epsilon_1$. The decrease in stored energy due to the expansion of a small volume dV of material 2 into material 1, due to the tangential component of the field, is then

$$dU_t = \frac{1}{2}(\epsilon_2 - \epsilon_1)F_{1t}^2 dV \quad (2-31)$$

Thus, the force acting normally on the boundary is

$$d\mathcal{F}_t = \nabla(dU_t) = \frac{1}{2}(\epsilon_2 - \epsilon_1)\nabla F_{1t}^2 dV \quad (2-32)$$

since $d\mathcal{F}_t/dV = \nabla p_t$, where p_t is the mechanical stress, which is the mechanical force per unit area. Therefore, the mechanical stress due to the tangential components of the field can be written as

$$p_t = \frac{1}{2}(\epsilon_2 - \epsilon_1)F_{1t}^2 \quad (2-33)$$

Similarly, it can easily be shown that the mechanical stress due to the normal components of the field is

$$p_n = \frac{1}{2} \frac{\epsilon_1(\epsilon_2 - \epsilon_1)}{\epsilon_2} F_{1n}^2 \quad (2-34)$$

Both p_t and p_n are normal to the boundary, tending to cause the dielectric material having a higher permittivity to move into the space occupied by the other, having a lower permit-

tivity. Because of the presence of this force, we would expect the occurrence of the following electromechanical effects:

- If material 1 is air with $\epsilon_1 = \epsilon_0$, and material 2 is a dielectric solid or liquid with $\epsilon_2 > \epsilon_1$, there is always a pulling force to pull material 2 into the space of air. Since the force is proportional to F^2 , a big mechanical impact on the dielectric solid may occur when the insulation system is subjected to a high voltage transient or switching surge.
- If material 1 is a dielectric fluid, such as transformer oil or hydrocarbon liquids with $\epsilon_1 > \epsilon_0$, and material 2 is a particle with $\epsilon_2 < \epsilon_1$, then under a nonuniform electric field, such as that shown in Figure 2-1, the particle may be ejected from the strong field region because the force given by Equations 2-33 and 2-34 is more dominant than the force given by Equation 2-8. In this case, these two forces are present but, in opposite directions. For example, in the strong field region near the high voltage windings of a power transformer, in which transformer oil is generally used for both heat transfer and insulation purposes, we would expect that any gas bubbles formed in the transformer oil would be ejected from the strong field region, and any particles with $\epsilon_2 > \epsilon_1$ and conductive particles, particularly moisture-absorbed particles, would be attracted to the strong field region, resulting in many insulation problems in power transformers.

The Force Acting on Conductor Surfaces

The attractive force between two parallel metal plates with opposite electric charges, in contact with a dielectric material of permittivity ϵ , will bring about deformation within the volume of the dielectric material in order to reduce the stored energy of the system. The decrease in stored energy resulting from a reduction of the volume of the material by a small amount of dV due to this compressive force can be written as

$$dU = \frac{1}{2}\epsilon F^2 dV \quad (2-35)$$

where F is the electric field strength produced by the charges on the metal plates. Following

the same derivation procedure, the compressive pressure acting on the dielectric material is

$$P_c = \frac{1}{2} \epsilon F^2 \quad (2-36)$$

This mechanical stress will result in a compression of the dielectric material, and its effect on the breakdown strength of polymers has been reported by several investigators.^{7,8} For example, a dielectric material, such as an insulating polymer, having a relative permittivity of 4 stressed at a field of 2 MV cm^{-1} , may suffer a compressive pressure of about 273 newtons cm^{-2} . This pressure, which increases proportionally with permittivity and the square of applied electric field, may be sufficient, especially under transient overvoltage conditions, to cause reduction of thickness or fracture of the insulation, followed by dielectric breakdown.

The Force Elongating a Bubble or a Globule in a Dielectric Fluid

A small bubble of gaseous or vapor phase or a globule of liquid phase of permittivity ϵ_2 in a dielectric fluid of permittivity ϵ_1 always takes on a spherical shape because of surface tension. In the presence of an electric field F_1 in the dielectric fluid, the field strength inside a spherical bubble or globule is given by Equation 2-24. Thus, by substituting Equation 2-24 into Equation 2-21, we would obtain the change in stored energy in the electrically stressed dielectric fluid due to the introduction of a bubble or globule, which is

$$\Delta U_s = \frac{1}{2} \left[\frac{3\epsilon_1(\epsilon_2 - \epsilon_1)}{2\epsilon_1 + \epsilon_2} \right] F_1^2 \left(\frac{4}{3} \pi r^3 \right) \quad (2-37)$$

where r is the radius of the spherical bubble or globule, and F_1 is the field strength in the region of the fluid which the bubble or globule subsequently occupies. It is assumed that a spheroidal shape is a good approximation to that into which a spherical bubble or globule may be expected to deform, and that the volume of the bubble or globule remains unchanged while its shape deforms from a sphere into a spheroid. As the field strength inside a spheroidal bubble or globule is given by Equation 2-25, substitu-

tion of Equation 2-25 into Equation 2-21 gives the change in stored energy in the electrically stressed dielectric fluid due to the presence of a spheroidal bubble or globule, which is

$$\Delta U_e = \frac{1}{2} \left[\frac{\epsilon_1(\epsilon_2 - \epsilon_1)}{\epsilon_1 - (\epsilon_1 - \epsilon_2)G} \right] F_1^2 \left(\frac{4}{3} \pi r^3 \right) \quad (2-38)$$

Thus, the work done at the expense of the stored energy for the deformation of the bubble or globule is

$$\begin{aligned} \Delta U &= \Delta U_e - \Delta U_s \\ &= \frac{1}{2} \left[\frac{\epsilon_1(\epsilon_2 - \epsilon_1)^2(1 - 3G)}{(2\epsilon_1 + \epsilon_2)[\epsilon_1 - (\epsilon_1 - \epsilon_2)G]} \right] \\ &\quad \bullet F_1^2 \left(\frac{4}{3} \pi r^3 \right) \end{aligned} \quad (2-39)$$

When ΔU is positive, the stored energy is further reduced after the bubble or globule is degenerated into a spheroid. For $\Delta U > 0$, G must be less than $1/3$, according to Equation 2-26. This corresponds to a prolate spheroid, implying that the spherical bubble or globule is elongated along the direction of the applied field, irrespective of whether the permittivity of the globule is larger or smaller than that of the fluid.

The force involved in the elongation can be derived as follows: By assuming that the volume of the bubble or globule is unchanged during elongation (we make this assumption purely for mathematical simplicity; there should be some change during the elongation process), we can write

$$r^3 = ca^2 \quad (2-40)$$

When the sphere is slightly deformed by elongation of the major axis from the original $2r$ to $2(r + dc)$, then we have

$$c = r + dc \quad (2-41)$$

and hence

$$\frac{a^2}{c^2} = \frac{r^3}{(r + dc)^3} \approx 1 - \frac{3dc}{r} \quad (2-42)$$

the terms of higher powers in the binomial expansion being ignored. Substituting Equation 2-42 into Equation 2-26 and then integrating it, we obtain

$$G = \frac{1}{3} - \frac{dc}{r} \quad (2-43)$$

Substitution of Equation 2-43 into Equation 2-39 gives the decrease in stored energy due to a bubble or globule with its major axis elongated from $2r$ to $2(r + dc)$

$$d(\Delta U) = \frac{9}{2} \left[\frac{\epsilon_1(\epsilon_2 - \epsilon_1)^2}{(2\epsilon_1 + \epsilon_2)^2} \right] F_1^2 \left(\frac{4}{3} \pi r^3 \right) \frac{dc}{r} \quad (2-44)$$

Thus, the elongation force is

$$\begin{aligned} \mathcal{F}_e &= \frac{d(\Delta U)}{dc} \\ &= \frac{9}{2} \left[\frac{\epsilon_1(\epsilon_2 - \epsilon_1)^2}{(2\epsilon_1 + \epsilon_2)^2} \right] \frac{F_1^2}{r} \left(\frac{4}{3} \pi r^3 \right) \end{aligned} \quad (2-45)$$

The elongation phenomenon was first predicted based on a simple analysis by Kao,^{3,9} and later verified experimentally by Kao¹⁰ and further analyzed rigorously by Garton and Krasucki.¹¹ Electric breakdown of dielectric liquids has been attributed to the formation and subsequent elongation of vapor bubbles in the liquids.¹²

The Dielectrophoretic Force

Unlike electrophoresis, in which the motion of particles depends on the charge of the particles, dielectrophoresis is defined as the motion of neutral particles in a nonuniform electric field, depending only on the force acting on polarized particles. The dielectrophoretic force tends to draw the particles whose permittivities are larger than that of the dielectric medium from the weak field to the intense field region, and to reject those with lower permittivities from the intense field to the weak field region in the same manner described in Section 2.1.2. On the assumption that the particles are spherical in shape, the change in stored energy due to the presence of spherical particle of radius a and permittivity ϵ_2 in a dielectric medium of permittivity ϵ_1 and stressed at a field F_1 can be calculated from Equation 2-37. Thus, the dielectrophoretic force can be written as

$$\begin{aligned} \mathcal{F}_d &= \nabla(\Delta U_d) \\ &= \frac{1}{2} \left[\frac{3\epsilon_1(\epsilon_2 - \epsilon_1)}{2\epsilon_1 + \epsilon_2} \right] \nabla F_1^2 \left(\frac{4}{3} \pi a^3 \right) \end{aligned} \quad (2-46)$$

Consider a nonuniform field, such as that produced by a single point charge Q given by Equation 2-2; the dielectrophoretic force can be expressed as

$$\mathcal{F}_d = \frac{1}{2} \left[\frac{3\epsilon_1(\epsilon_2 - \epsilon_1)}{2\epsilon_1 + \epsilon_2} \right] \left(\frac{4}{3} \pi a^3 \right) \left[\frac{Q^2}{4\pi^2 \epsilon_1^2 r^5} \right] \quad (2-47)$$

This force depends on the distance from the point charge r and is opposed by a viscous drag of the medium, which, according to Stokes's law, is $6\pi\eta av$, where η is the viscosity of the medium and v is the velocity of the particle. Hence, the velocity of the particle in such a nonuniform field can be approximately evaluated by

$$v = \frac{1}{2} \left[\frac{3\epsilon_1(\epsilon_2 - \epsilon_1)}{2\epsilon_1 + \epsilon_2} \right] \left(\frac{2a^2}{9\eta} \right) \left[\frac{Q^2}{4\pi^2 \epsilon_1^2 r^5} \right] \quad (2-48)$$

The dielectrophoretic effect may be used to produce pumping action of nonconducting liquids, to cause separation of the components in suspension in a dielectric fluid, to cause precipitation, or to produce mixing.¹³

The Electrostriction Force

Electrostriction is defined as the elastic deformation of a dielectric material under the force exerted by an electric field. A dielectric material can be thought of as consisting of dielectric particles (atoms or molecules) uniformly distributed throughout a vacuum space. Thus, the local field acting on each particle is higher than the apparent or measured field by an amount resulting from the polarization field of the surrounding polarized particles. This local field is given by

$$F_{loc} = \frac{\epsilon + 2\epsilon_o}{3\epsilon_o} F \quad (2-49)$$

where ϵ and ϵ_o are the permittivities of the dielectric material and vacuum, respectively. (See Section 2.5.) The polarization involves the displacement of electric charges, shifting the originally symmetrical charge distribution to an nonsymmetrical one. This implies that the polarization tends to elongate the particles in the direction of the field, i.e., to cause the

material to expand in the direction parallel to the field and to contract in the direction perpendicular to the field. This process is accompanied by a slight decrease in volume of the material and hence a slight increase in density of the material. Suppose a material of volume V suffers a slight decrease in volume by dV and hence a slight increase in density by $d\rho$, due to elongation; then from Equation 2-21, the decrease in stored energy of this piece of material due to the decrease in volume by dV is

$$d(\Delta U)_v = \frac{1}{2} \left[\frac{(\epsilon - \epsilon_o)(\epsilon + 2\epsilon_o)}{3\epsilon_o} \right] F^2 dV \quad (2-50)$$

Therefore, the hydrostatic pressure tending to contact the material is

$$P_v = \frac{d(\Delta U)_v}{dV} = \frac{(\epsilon - \epsilon_o)(\epsilon + 2\epsilon_o)}{6\epsilon_o} F^2 \quad (2-51)$$

It should be noted that most solids are incompressible because the particles (atoms or molecules) are rigidly bonded together. For slightly compressible materials—such as some liquids, polymers, or ceramics—electrostriction phenomenon may be detectable. Also, the hydrostatic pressure may become significant for materials with high dielectric constants.

The Torque Orienting a Solid Body

If a body is not symmetrical about its center, it will experience a torque, tending to orient itself in such a manner that the stored energy in the electric field will become minimal. For simplicity, a prolate spheroidal shape is chosen as the approximation to the general shape of a solid body, and its principal axes, which follow the relationship $c > a = b$, form a rectangular coordinate frame, x directing the major axis $2c$, y and z , the minor axes $2a$ and $2b$. If one of the minor axes is originally in the direction of the field, the decrease in stored energy when its major axis orients and causes an angular displacement θ , from Equations 2-21 and 2-25, is

$$\Delta U_T = \frac{2\pi c a^2 \epsilon_1 (\epsilon_2 - \epsilon_1)^2 (G_y - G_x) \sin^2 \theta}{3[\epsilon_1 - (\epsilon_1 - \epsilon_2)G_x][\epsilon_1 - (\epsilon_1 - \epsilon_2)G_y]} F_1^2 \quad (2-52)$$

Hence, the torque exerted on the body is

$$\begin{aligned} T &= \frac{d(\Delta U_T)}{d\theta} \\ &= \frac{2\pi c a^2 \epsilon_1 (\epsilon_2 - \epsilon_1)^2 (G_y - G_x) \sin 2\theta}{3[\epsilon_1 - (\epsilon_1 - \epsilon_2)G_x][\epsilon_1 - (\epsilon_1 - \epsilon_2)G_y]} F_1^2 \end{aligned} \quad (2-53)$$

where

$$\begin{aligned} G_x &= \frac{c a^2}{2} \int_0^\infty \frac{ds}{(s + c^2)^{3/2} (s + a^2)} \\ &= -\frac{1}{2} \frac{(a/c)^2}{[1 - (a/c)^2]^{3/2}} \left\{ 2 \left[1 - (a/c)^2 \right]^{1/2} \right. \\ &\quad \left. + \log \frac{1 - [1 - (a/c)^2]^{1/2}}{1 + [1 - (a/c)^2]^{1/2}} \right\} \end{aligned} \quad (2-54)$$

$$\begin{aligned} G_y &= \frac{c a^2}{2} \int_0^\infty \frac{ds}{(s + c^2)^{1/2} (s + a^2)^2} \\ &= \frac{1}{4} \frac{(a/c)^2}{[1 - (a/c)^2]^{3/2}} \left\{ \frac{2 \left[1 - (a/c)^2 \right]^{1/2}}{(a/c)^2} \right. \\ &\quad \left. + \log \frac{1 - [1 - (a/c)^2]^{1/2}}{1 + [1 - (a/c)^2]^{1/2}} \right\} \end{aligned} \quad (2-55)$$

F_1 is the field strength in the region of the dielectric medium where the solid body is subsequently placed, and ϵ_1 and ϵ_2 are the permittivities of the medium and the body, respectively.

The torque tends to orient the major axis of the body to the direction of the field and is independent of the relative magnitudes of ϵ_1 and ϵ_2 . The magnitude of the torque depends on the size and the position of the body. It is zero when the major axis of the body is perpendicular to the field, and is maximal when it is at the angle of 45° to the field. The torque exerted on the solid particles may accentuate the effect of impurities on the dielectric breakdown of liquids, but on the other hand, it may be used as a means to determine the permittivity of a dielectric sample of known shape.

These six electromechanical forces may be appreciable when the impressed electric fields

are large. The directions of these forces are independent of the directions of the fields, implying that unidirectional and alternating fields will produce the same effects.

2.1.3 Electrostatic Induction

Before discussing electric polarization and relaxation, it is desirable to review briefly the electrostatic induction phenomenon.

Suppose that we have a positively charged insulator A with the charge distribution, as shown in Figure 2-4(a). If we now place an uncharged conductor B near the charged insulator A, the electrostatic field produced by A will shift the free electrons in B toward the surface close to A, resulting in the formation of a negative charge on one side and a positive charge on the opposite side, where there is a deficit of electrons, as shown in Figure 2-4(a). This phenomenon is referred to as electrostatic induction. The electric field inside B is zero because this field is composed of two components: one is the original field from A, which should always exist inside B, and the other is the field produced by the induced negative and positive charges in B, which is opposite to that

produced by A. If we now connect the conductor B to ground by a conducting wire, electrons will flow to B from the ground to make up the deficit of electrons on the positive-charge side, as shown in Figure 2-4(b). This is equivalent to saying that some positive charge flows out from B to ground. Thus, the negative charge on the side close to A is called the bound induced charge, while the positive charge on the opposite side is called the free induced charge. When the free induced charge disappears, both the magnitude and the distribution of the negative bound charge will change in such a way that the electric field inside B remains zero. Now, if the ground connecting wire is cut off and the conductor B is removed from the field produced by A, then the conductor B becomes an isolated negatively charged conductor, and the bound induced charge becomes free to move to redistribute itself on the surface in such a way that the electric field inside B is zero and the electric flux outside B is perpendicular to the surface of B, because any tangential components of the electric field would cause the movement of surface charges. The charge distribution is dependent solely on the geometric shape of the conductor B. It should be noted

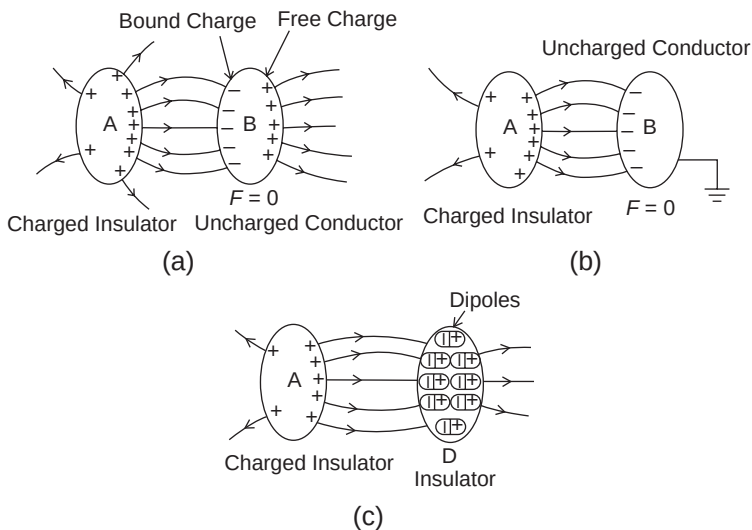


Figure 2-4 Electrostatic induction: (a) positively charged insulator A and uncharged conductor B (isolated), (b) positively charged insulator A and uncharged conductor B (grounded), and (c) positively charged insulator A and isolated insulator D.

that any change in charge distribution in conductor B due to different arrangements does not affect the charge distribution in the charged insulator A, simply because charges in insulator A cannot freely move. However, if A is a charged conductor instead of a charged insulator, then the induced charge in conductor B will also cause a change in charge distribution in charged conductor A, in order to make the electric field inside both A and B be zero.

A conductor, or a conducting material, generally refers to the material consisting of a great many mobile free charge carriers, such as a metal in which the concentration of free electrons (mobile charge carriers) is of the same order as that of atoms, i.e., about 10^{22} to 10^{23} cm^{-3} , and electrolytes in which ions are the charge carriers. In those conducting materials, the electrostatic induction phenomenon prevails. But in other materials, such as semiconductors and dielectric materials, the number of mobile charge carriers is far less than the number of atoms. Although an electric field would cause the movement of mobile charge carriers to produce space charge polarization in a manner similar to electrostatic induction, space charge polarization usually plays an insignificant role because it involves only a small number of mobile charge carriers, as compared to electric polarization, which involves all atoms.

Now, suppose the uncharged conductor B, shown in Figure 2-4(a), is replaced by an isolated uncharged insulator D of the same shape, shown in Figure 2-4(c). In this case, the field produced by the charged insulator A will not induce charge in D because there are practically no mobile free charge carriers in insulator D. Instead, the field will polarize the material by shifting slightly the normally symmetrical distribution of electron clouds of atoms and by orienting dipolar molecules toward the direction of the field to form dipoles. So, each atom or molecule contributes a tiny dipole to form a big dipole, as shown in Figure 2-4(c). If we remove the insulator D after it has been completely polarized, the polarized insulator D will undergo relaxation and gradually become depolarized, due to thermal agitation. Electric

polarization and relaxation processes are discussed in the following sections.

2.2 Electric Polarization and Relaxation in Static Electric Fields

Electric polarization refers to a phenomenon of the relative displacement of the negative and positive charges of atoms or molecules, the orientation of existing dipoles toward the direction of the field, or the separation of mobile charge carriers at the interfaces of impurities or other defect boundaries, caused by an external electric field. A detailed account of various polarization processes will be given in Section 2.3. Electric polarization can also be thought of as charge redistribution in a material caused by an external electric field. The work done for the charge redistribution and the energy loss involved in the redistribution process require an energy supply. Where does this energy come from? We can say that the whole polarization process is performed at the expense of the potential energy released from this process, because the total potential energy of the system in an electric field is smaller after electric polarization than before it.

We shall apply Gauss's law to explain some dielectric phenomena resulting from electric polarization. As discussed in Gauss's Law in Chapter 1, Gauss's law simply states that the electric flux outward from a volume is equal to the total net charge enclosed inside it. We use a simple system consisting of two metal plates parallel to each other with an area A each and a separation d , which is much smaller than the linear dimension of the plates, so that the fringing effect at the edges can be ignored by approximation. Suppose we introduce a positive charge $+Q$ on the upper plate and a negative charge $-Q$ of the same magnitude on the lower plate. This can be easily done by connecting a steady DC voltage supply across the plates and charging the system as a capacitor, and then disconnecting the supply from the system as soon as the charge has been accumulated to the desired value Q . These

charges on the plates will create a potential difference between the plates V that is proportional to Q . Thus, we can write

$$V \propto Q$$

$$V = \frac{Q}{C} \quad \text{or} \quad Q = CV \quad (2-56)$$

where C is the proportionality constant, generally called the capacitance. By denoting the surface charge density on the plates as σ_s , the charge Q can be expressed as

$$Q = C(Fd) = \sigma_s A \quad (2-57)$$

where F is the electric field strength, which is simply equal to V/d . From Equation 2-57 we can express C in the form

$$C = \frac{Q}{V} = \frac{\sigma_s A}{Fd} = \epsilon \frac{A}{d}$$

From this we obtain

$$\epsilon = \frac{\sigma_s}{F} \quad (2-58)$$

Where ϵ is the permittivity of the material filled between the two plates, which is equal to the ratio of $\frac{\sigma_s}{F}$. In the following sections, we shall discuss this ratio for three general cases.

2.2.1 Vacuum Space

By true vacuum space, or free space, we mean that in the space there are no detectable particles, although there may be a hypothetical ether, thought of as the medium for the propagation of electromagnetic waves in free space. However, a true vacuum is not available on the earth. The best vacuum system that can be produced by today's technology can achieve only a vacuum of about 10^{-14} torr, in which there are still about 300 particles per cm^3 . Assuming that the space between the plates is free space, σ_{so} is equal to the electric flux density D_o based on Gauss's law, as shown in Figure 2-5.

So we can write

$$\sigma_{so} = D_o = \epsilon_o F_o$$

or

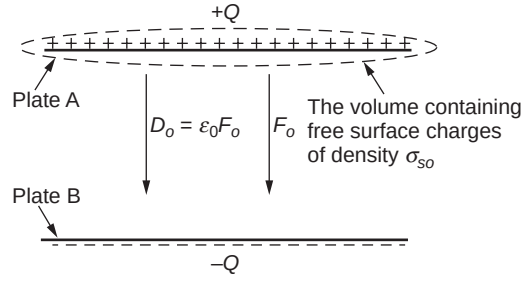


Figure 2-5 The electric field F_o and the associated electric flux density D_o produced by the free surface charges of density σ_{so} on the plates in vacuum space (free space).

$$\epsilon_o = \frac{\sigma_{so}}{F_o} \quad (2-59)$$

where F_o is the electric field strength in the vacuum space, which is equal to Q/V_o and V_o is the potential between two plates created by the charge Q ; and ϵ_o is the permittivity of vacuum or free space, which is practically the same as the permittivity of dielectric gases, such as air, at normal pressure and temperature. The permittivity of gases, depending on the molecular structure, has only negligibly small deviation from the ideal value of ϵ_o at normal pressure and temperature.¹⁴

2.2.2 Conducting Materials

A conducting material, or a conductor, refers to a material containing many mobile free charge carriers. For example, in a piece of copper or sodium, each atom contributes one free electron, making the total electron concentration equal to the concentration of atoms, i.e., 10^{22} to 10^{23} cm^{-3} . These electrons move freely and randomly and distribute themselves with statistical uniformity in the bed of regularly arranged positive ions. If such a piece of metal with the thickness of d -2s were inserted into the space between the two charged plates, with a small gap s to separate the plates and the metal conductor, then the charge on the plates would cause the electrons in the conductor to move toward the surface close to the positively charged plate but could not leave the surface, resulting in a nonuniform distribution of the net

charge. This creates a net negative space charge on one surface and a net positive space charge on the other in a manner similar to the electronic polarization of a neutral atom. Thus, in this case, we can say that charge Q on plate A induces a negative charge equal in magnitude but opposite in polarity on one surface of the metal close to plate A. Similarly, a positive charge is induced on the other surface of the metal by the negative $-Q$ on plate B. So, in the vacuum space between plate A and the metal surface, or between plate B, and the other surface of the metal, we have

$$\sigma_{so} = D_o = \epsilon_o F_o$$

$$\epsilon_o = \frac{\sigma_{so}}{F_o}$$

which is the same as that as given by Equation 2-59. The response in the vacuum space is exactly the same as that described in Section 2.2.1.

However, inside the metal there is no electric flux based on Gauss's law, as shown in Figure 2-6, and hence, $F = 0$.

$$\epsilon_s = \frac{\sigma_s}{F} \rightarrow \infty \quad (2-60)$$

In metal, the permittivity is ∞ under static fields only. Under a time-varying field or an alternating field, the charges induced on the metal surfaces may not follow instantaneously the time-varying field. In this case, the permittivity is not infinite, but a finite value. We shall discuss this phenomenon further in later sections.

2.2.3 Dielectric Materials

One of the important electrical properties of dielectric materials is permittivity (or relative permittivity, which is generally referred to as the dielectric constant). For most materials, the dielectric constant is independent of the electric field strength for fields below a certain critical field, at or above which carrier injection into the material becomes important. The dielectric constant depends strongly on the frequency of the alternating electric field or the rate of the change of the time-varying field. It

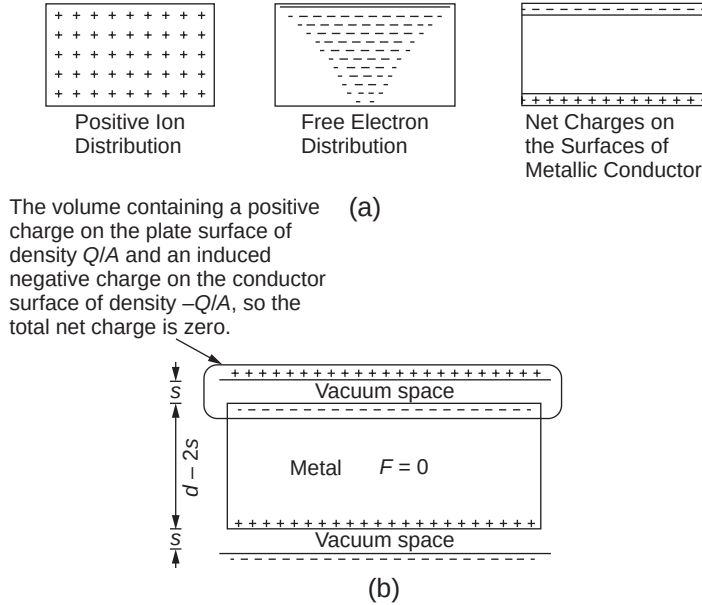


Figure 2-6 (a) Induced net charges on the surfaces of a metallic conductor in the presence of a static electric field, and (b) the charge on the plate with the surface charge density σ_s inducing a charge equal in magnitude but opposite in polarity on the surface of the conductor, resulting in zero electric field inside the conductor. The vacuum space s is exaggerated for clarity.

also depends on the chemical structure and the imperfections (defects) of the material, as well as on other physical parameters including temperature and pressure, etc. This chapter is devoted to the discussion of electric polarization and relaxation under various conditions.

A dielectric material is made up of atoms or molecules that possess one or more of five basic types of electric polarization:

1. Electronic polarization
2. Atomic or ionic polarization
3. Dipolar polarization
4. Spontaneous polarization
5. Interface or space charge polarization.

Each type of polarization requires time to perform; this is why the degree of the overall polarization depends on the time variation of the electric field. In the present section, we consider only the polarization under static fields. In the first place, we consider a perfect dielectric material. The so-called “perfect” material implies that inside the material no mobile charge carriers (electrons or ions) are present. If a piece of such a dielectric material is put near the system shown in Figure 2-5, the system will tend to attract this piece of material into the vacuum space between the two plates in order to reduce the potential energy of the system.

Suppose now that the piece of the material is inside the space between the metal plates with the original surface of density σ_s unaltered, as shown in Figure 2-7. This causes the potential between the plates produced by the original charge Q on the plates to decrease to a smaller value. In fact, with Q on the plates remaining constant, the ratio of the electric field in free space F_o to that filled with the dielectric material F is the so-called dielectric constant or relative static permittivity, $\epsilon_{sr} = \epsilon_s/\epsilon_o$.

With the dielectric material, one portion of σ_s is used to compensate the polarization charges on the surfaces of the material in contact with the metal plates. This portion of σ_b is bound at the locations with its charge opposite in polarity and equal in magnitude to the polarization charges of the material, as shown

The volume containing free surface charge of density $\sigma_s - \sigma_b$ and bound surface charge of density σ_b , which is for compensating the polarization surface charge of the dielectric material.

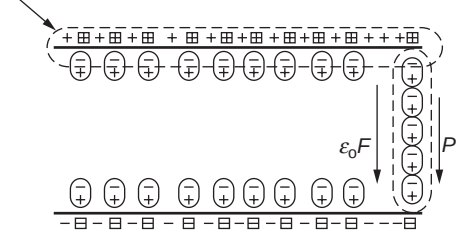


Figure 2-7 The surface charge of density σ_s consisting of two portions: the bound charge σ_b and the free charge $\sigma_s - \sigma_b$. The free charge $\sigma_s - \sigma_b$ produces the electric field F and the electric flux density of $\epsilon_o F$, while the bound charge portion produces polarization P . + and - denote the free positive and negative charges, respectively, and $\boxplus$ and $\boxminus$ denote the bound positive and negative charges, respectively.

in Figure 2-7. This portion of charge is termed the bound charge density σ_b . By denoting the number of atoms (or molecules) per unit volume of the material as N , and the average separation between charges $+q_d$ and $-q_d$ of each dipole (each atom or molecule produces one dipole) as δ , we can write

$$\sigma_b = N\langle q_d \delta \rangle \quad (2-61)$$

The other portion $\sigma_s - \sigma_b$ is the free surface charge density, which acts in exactly the same manner as that in a vacuum, creating an electric flux density $D_o = \epsilon_o F$. Thus, we can write

$$\begin{aligned} \sigma_s &= (\sigma_s - \sigma_b) + \sigma_b \\ &= D = D_o + P \\ &= \epsilon_s F = \epsilon_o F + (\epsilon_s - \epsilon_o) F \end{aligned} \quad (2-62)$$

Therefore, polarization P can be expressed as

$$\begin{aligned} p &= (\epsilon_s - \epsilon_o) F = \sigma_b = \text{polarization} \\ &= \frac{\text{Bound charge}}{\text{Surface area}} = N\langle q_d \delta \rangle \\ &= \frac{\text{Number of Induced Dipole Moments}}{\text{Volume}} \\ &= N\langle \vec{u} \rangle \end{aligned} \quad (2-63)$$

where $\langle \vec{u} \rangle$ is the average value of the dipole moment, which is given by

$$\langle \vec{u} \rangle = \langle q_a \delta \rangle = \alpha \vec{F}_{loc} \quad (2-64)$$

where α is called the polarizability and is a scalar quantity if the constituent particles (atoms or molecules) in a material are spherically symmetrical in shape and $\langle \vec{u} \rangle$ is in the direction of F . However, if the particles (atoms or molecules) are not spherically symmetrical in shape, particularly those with permanent dipoles, then $\langle \vec{u} \rangle$ is not exactly in the direction of F . In this case, α becomes a tensor. Throughout this book, we are interested mainly in physical concepts and, therefore, to maintain clarity and simplicity rather than to involve lengthy mathematics, we consider only homogeneous materials and electric fields that are independent of space coordinates, although they may vary with time. However, it should be noted that $\langle \vec{u} \rangle$ depends on the local field at which the individual particle is polarized. This field F_{loc} is different from the applied external field F ; therefore, by making $\langle \vec{u} \rangle$ proportional to F , the polarizability α must be related to the local field. From Equations 2-63 and 2-64, we can also write

$$\alpha = \frac{\epsilon_s - \epsilon_o}{N} \quad (2-65)$$

We can also say that the polarizability α of a particle in a dielectric material is induced by the local field. We will see this point more clearly in Section 2.3. It is sometimes convenient to use the relative permittivity ϵ_{sr} (i.e., the dielectric constant) to describe the dielectric properties

$$\epsilon_{sr} = \frac{\epsilon_s}{\epsilon_o} = \left(1 + \frac{N\alpha}{\epsilon_o} \right) = 1 + \chi \quad (2-66)$$

where $\chi = \frac{N\alpha}{\epsilon_o}$ is called the electric susceptibility, which can also be written as

$$\begin{aligned} \chi = \epsilon_{sr} - 1 &= \frac{P}{D_o} \\ &= \frac{\text{Bound surface charge density on the plates}}{\text{Free surface charge density on the plates}} \end{aligned} \quad (2-67)$$

χ is the electric counterpart of the magnetic susceptibility described in Magnetization in Chapter 1.

It should be noted that this discussion is limited to the phenomena under a static field. This means that the charges $+Q$ and $-Q$ already put on the plates are constant and that the dielectric material has been inserted into the vacuum space for a time period long enough for all types of polarization to settle down to their steady state values. Now, supposing that it takes a negligibly small amount of time for the insertion of the dielectric material into the vacuum space, we can assume that at time $t = 0$ the material is already inside the vacuum space and starts to be polarized by the field created by the charges on the metal plates. The polarization due to the elastic displacement of electron clouds of the particles (atoms and molecules) requires a very little time, while the polarization involving the movement of the particles, such as the orientation of permanent dipoles or the migration of charge carriers (electrons or ions), requires a much longer time to perform. All types of polarization encounter some inertia counteracting the change and, therefore, involve some dielectric loss, as shown in Figure 2-8.

In the system being considered, σ_s on the plates is constant, but the permittivity starts to rise from $\epsilon = \epsilon_o$ and the corresponding field to decrease from $F = F_o$ at $t = 0$. It takes a very little time for ϵ to increase to ϵ_∞ and for F to decrease to F_∞ due to the polarization contributed by the elastic displacement of electron clouds of atoms or molecules (electron and atomic polarization). The time involved in this polarization is of the order of 10^{-14} to 10^{-13} second. We shall discuss ϵ_∞ further in Section 2.6. However, for the permittivity to increase from ϵ_∞ to its steady state value ϵ_s and for the corresponding field to decrease from F_∞ to its final value F_s , a considerably larger amount of time is required, because this change is caused by the polarization associated with the inelastic movement of particles such as the sluggish collective orientation of dipoles or the migration of charge carriers to form space charges near the electrodes or grain boundaries. Ignoring the small loss due to elastic displacement of electron clouds, the transition from $P_\infty = (\epsilon_{o/r} - 1)\epsilon_o F_\infty$ to $P_s = (\epsilon_{sr} - 1)\epsilon_o F_s$ involves some

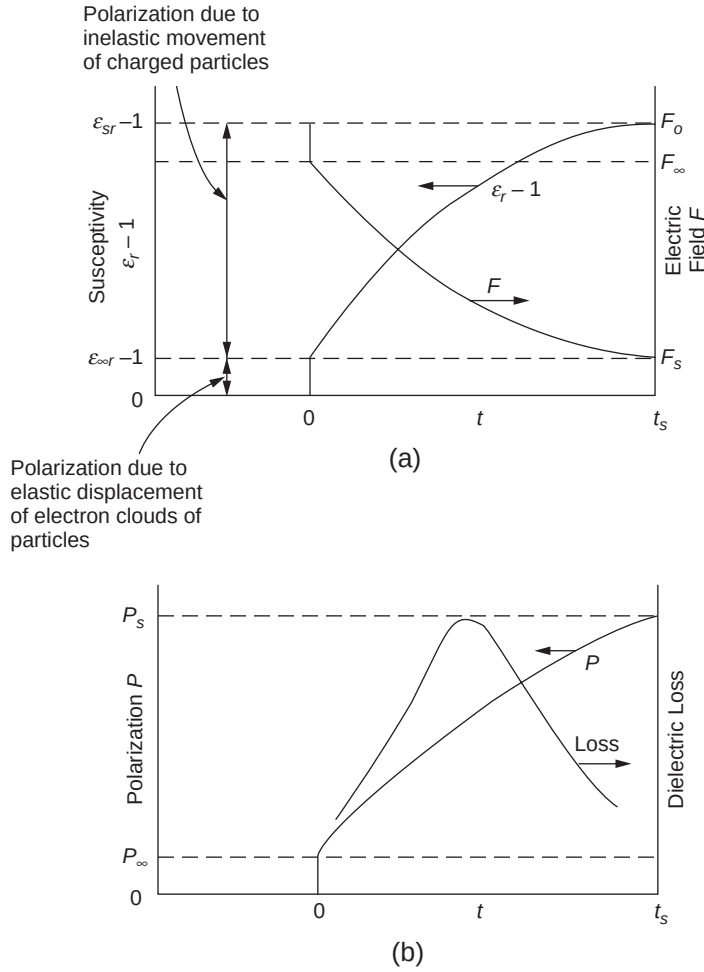


Figure 2-8 (a) The variation of the susceptibility $\epsilon_r - 1$ and the electric field F with time, and (b) the polarization $P = (\epsilon_r - 1) \epsilon_0 F$ and the dielectric loss with time, where $P_\infty = (\epsilon_\infty - 1) \epsilon_0 F_\infty$, and $P_s = (\epsilon_{sr} - 1) \epsilon_0 F_s$.

energy loss, which must be consumed to overcome the inertia resistance. This energy loss is called the dielectric loss, which always accompanies time-varying electric fields. At static fields, it appears only during the transient period in which the electric field across the dielectric material is time-varying, although the charges on the metal plates are constant. This loss per unit time will become zero when the polarization has reached its final steady value $\epsilon = \epsilon_s$. The loss is proportional to the PF product; thus the loss per unit time becomes maximal when the PF product is maximal, as shown in Figure 2-8(b).

After the overall polarization process is completed, we short-circuit the two metal plates. Under this short-circuiting condition, the charges on plate A and plate B will be immediately neutralized, leaving only the polarized particles in the dielectric material to be gradually depolarized. It would take some time, depending on the environmental temperature, for all polarized particles to become completely depolarized, particularly for those whose polarization involves the orientation of permanent dipoles or the migration of charge carriers. After the short-circuit current, due to the neutralization of the charges on the metal plates, a

reverse depolarization current will follow, due to the depolarization of the polarized particles. The thermally activated depolarization process will be discussed in Charge Storage Involving Dipolar Charges in Chapter 5, which deals with electrets.

In reality, there is no perfect dielectric material. All real dielectric materials contain, more or less, some mobile charge carriers. If such a real material has the total charge carriers n per unit volume, each carrying a charge q and having an average mobility u_n , then the conductivity of this material σ is $qu_n n$. For a dielectric material with low conductivity, we can use an equivalent circuit, shown in Figure 2-9, to describe the response of this material under an electric field produced by two charged metal plates similar to that shown in Figure 2-7. As mentioned, the parameters ε and R_d are derived directly from the polarization process. Thus, at $t = 0$, just after the insertion of the material into the vacuum space, $\varepsilon \rightarrow \varepsilon_\infty$ and $R_d \rightarrow \infty$, and at $t = t_s$, $\varepsilon \rightarrow \varepsilon_s$ and $R_d \rightarrow \infty$. The value of the resistance R_d is finite only during the period $0 < t < t_s$. But the leakage current through the leakage resistance R_c will continue to drain the charge from plate A to plate B, making σ_s decay with time. By expressing the voltage between the two plates as

$$Fd = \frac{d\sigma_s}{dt} R_c$$

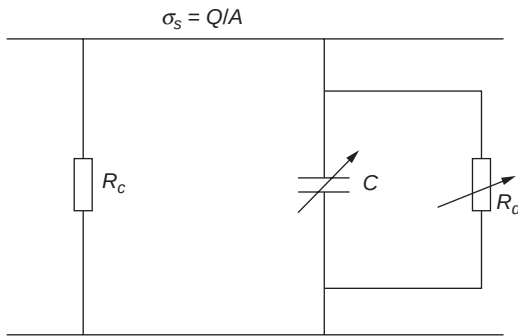


Figure 2-9 The equivalent circuit of a real dielectric material. $R_c = \rho d$ is the leakage resistance of the material where ρ is the resistivity; $C = \varepsilon/d$ is the capacitance per unit area, where ε varies with time; R_d is the resistance per unit area derived directly from the polarization process and varies with time.

or

$$\frac{d\sigma_s}{dt}(\rho d) = \left[\frac{\sigma_s(t)}{\varepsilon_s} \right] d \quad (2-68)$$

using the boundary condition $\sigma_s(t=0) = \sigma_s(0)$, the solution of Equation 2-68 gives

$$\sigma_s(t) = \sigma_s(0) e^{-t/\tau_d} \quad (2-69)$$

where $\tau_d = \rho \varepsilon_s$ is the dielectric relaxation time and ρ is the resistivity of the material. τ_d is the time required for the originally induced charge to decay to 36.7% of its original value. τ_d can be very large for materials with a very high resistivity (i.e., low conductivity). So, it is quite dangerous to touch an electrically disconnected capacitor, because some charges may still remain on the electrodes even though the capacitor may have been de-energized for some time. For example, if $\varepsilon_s = 3\varepsilon_0$ and $\rho = 10^{18} \Omega \cdot \text{cm}$, then $\tau_d = 2.65 \times 10^5$ seconds = 3 days. This means that it may take many days for the charges to be completely neutralized.

2.3 The Mechanisms of Electric Polarization

The polarizability defined by Equation 2-64 implies that if $\langle \vec{u} \rangle$ is proportional to the local field F_{local} at which the particles are polarized, then $\langle u \rangle$ is expressed as $\langle \vec{u} \rangle = \alpha F_{local}$ and the polarizability α depends only on the mechanism of polarization and can be defined simply as the average dipole moment per unit field strength of the local field F_{local} , which is different from the externally applied field F . But if $\langle \vec{u} \rangle$ is expressed as $\langle \vec{u} \rangle = \alpha_{eff} F$, i.e., $\langle \vec{u} \rangle$ is proportional to F , in this case, α_{eff} should also be defined similarly to the effective dipole moment per unit field strength of the applied field F . However, α_{eff} depends not only on the mechanism of polarization, but also on the factor of F_{loc}/F . We shall consider first the mechanisms responsible for the polarizability, α , and the local field later.

There are three major mechanisms of electric polarization. At moderate electric fields (i.e., at fields much lower than the inner atomic or molecular fields), and for materials with a

very low conductivity (i.e., the concentration of charge carriers inside the materials is so low that its effect can be neglected), these mechanisms prevail.

- **Electronic polarization** (also called optical polarization): The electric field causes deformation or translation of the originally symmetrical distribution of the electron clouds of atoms or molecules. This is essentially the displacement of the outer electron clouds with respect to the inner positive atomic cores.
- **Atomic or ionic polarization**: The electric field causes the atoms or ions of a polyatomic molecule to be displaced relative to each other. This is essentially the distortion of the normal lattice vibration, and this is why it is sometimes referred to as vibrational polarization.
- **Orientational polarization**: This polarization occurs only in materials consisting of molecules or particles with a permanent dipole moment. The electric field causes the reorientation of the dipoles toward the direction of the field.

Both the electronic polarization and the atomic polarization are due mainly to the elastic displacement of electron clouds and lattice vibration within the atoms or molecules. Their interaction is an intramolecular phenomenon, and the restoring force against the displacement is relatively insensitive to temperature, so electronic and atomic (or ionic) polarization processes are only slightly dependent on temperature. However, orientational polarization is a rotational process, which encounters not only the resistance due to thermal agitation, but also that due to the inertia resistance of the surrounding molecules, giving rise to mechanical friction. The rotation of a dipole in a material is like a small ball, or any body, rotating in a viscous fluid. Under an external force, it tends to change from its original equilibrium state to a new, dynamic equilibrium state, and when the force is removed, it then relaxes back to its original equilibrium state. This process is generally referred to as the relaxation process. This

polarization involves the inelastic movement of particles, and its interaction is an intermolecular phenomenon; hence, orientational polarizability is strongly temperature-dependent.

At higher fields, carrier injection becomes important. For materials consisting of a high concentration of charge carriers (i.e., with a high conductivity), polarization due to the migration of charge carriers to form space charges at interfaces or grain boundaries becomes important. This type of polarization is called space charge polarization.

Thus, the total polarizability of material α comprises four components

$$\alpha = \alpha_e + \alpha_i + \alpha_o + \alpha_d \quad (2-70)$$

where α_e , α_i , α_o , and α_d are the polarizabilities due to electronic, atomic, orientational, and space charge polarizations, respectively. For ferroelectric materials, there is also a component, called spontaneous polarization, under certain conditions. The following sections discuss the mechanisms responsible for these polarizations.

2.3.1 Electronic Polarization (Also Called Optical Polarization)

There are two ways to calculate electronic polarizability: the classical approach and the quantum-mechanical approach. We will first discuss the classical approach.

Classical Approach

The polarizability α_e depends on the atomic number of the atom and predominantly on the number of electrons in the outermost shell. The noble gas atoms, such as He, Ne, Ar, Kr, Xe, and Ra, whose shells are completely filled (effectively screening the nucleus from the influence of the applied fields), have the lowest polarizabilities for their atomic numbers. The Group I elements, such as H, Li, Na, K, Rb, and Cs, with only one electron in the outermost shell, have the highest polarizabilities for their atomic numbers, possibly due to the ease of polarization of a single electron in the outermost orbit.^{15,16} Electronic polarizability can be

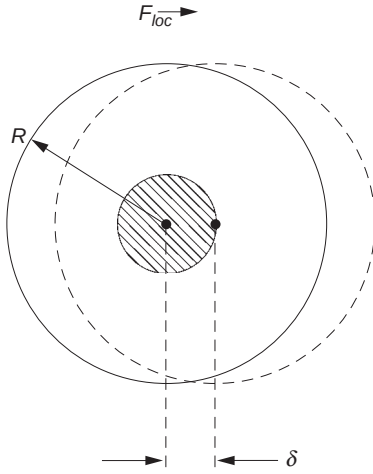


Figure 2-10 The displacement of the electron cloud relative to the nucleus of positive charges Zq due to the polarization by the local field F_{loc} .

measured in monoatomic gases. Suppose that the electron cloud of charges $-Zq$ is uniformly distributed in a sphere of radius R and that its center of gravity originally coincided with that of the nucleus, and suppose that it is displaced by the field to a distance δ from the center of the nucleus, as shown in Figure 2-10. There is a coulombic force tending to bring the electron cloud back to its original position.

According to Gauss's law, the coulombic force is only exerted on the electron cloud that does not surround the positive nucleus charges $+Zq$. This portion of the electron cloud is contained in the sphere of radius δ and it is

$$Q_\delta = \frac{Zq(4\pi\delta^3/3)}{4\pi R^3/3} = -Zq\delta^3/R^3 \quad (2-71)$$

Thus, the coulombic force is

$$\mathcal{F} = \frac{Zq \cdot Zq\delta^3/R^3}{4\pi\epsilon_0\delta^2} = (Zq)^2\delta/4\pi\epsilon_0R^3 \quad (2-72)$$

This force must balance the displacement force, which is

$$\mathcal{F}_d = ZqF_{loc} \quad (2-73)$$

By setting Equation 2-72 equal to Equation 2-73, we obtain

$$\delta = 4\pi\epsilon_0R^3F_{loc}/Zq \quad (2-74)$$

The dipole moment is given by

$$u_e = \alpha_e F_{loc} = Zq\delta = 4\pi\epsilon_0R^3F_{loc} \quad (2-75)$$

Therefore, the electronic polarizability is

$$\alpha_e = 4\pi\epsilon_0R^3 = 3\epsilon_0V_a \quad (2-76)$$

where V_a is the volume of the atom.

α_e is proportional to the volume of the atom. For the hydrogen atom, R is about $0.50 \text{ \AA}$, $\alpha_e = 1.57 \times 10^{-24} \epsilon_0 \text{ cm}^3$. For $F_{loc} = 10^4 \text{ V cm}^{-1}$, $\delta = 10^{-14} \text{ cm}$. The displacement distance δ is extremely small, only of the order of two-millionths of the radius of the atom. The susceptibility $\chi_e = N\langle u \rangle / \epsilon_0 F_{loc} = N4\pi R^3 = 1.57 \times 10^{-24} N$. For the ideal gas under normal conditions (0°C and 760 mm Hg), the number of particles per unit volume is $2.687 \times 10^{19} \text{ cm}^{-3}$, which is also known as the Loschmidt number, and $\chi_e = 1.57 \times 10^{-24} \times 2.687 \times 10^{19} = 4.22 \times 10^{-5}$. These numerical data give us a feeling for the orders of magnitude involved in the microscopic parameters. There is another classical method that can be used for the calculation of α_e , i.e., using Bohr's model with the electrons revolving around a circular orbit instead of using a uniform electron cloud density up to radius R as just shown.

This displacement of the electrons Δx , resulting from electronic polarization, always generates an elastic restoring force tending to bring the electrons back to their equilibrium position. This restoring force is proportional to the displacement Δx , resulting from the interaction of the electron with the bare nucleus and with other electrons screening the nucleus. The electron orbiting the nucleus is like a harmonic oscillation with a natural frequency ω_0 . Thus, the motion of the electrons is governed by the following equation:

$$m \frac{d^2 \Delta x}{dt^2} = -\gamma \Delta x - ZqF_{loc} \quad (2-77)$$

where m is the electron mass, q is the electronic charge, Z is the number electrons involved, F_{loc} is the local field acting on the atoms, and γ is the force constant. By assuming, for simplicity, that when the centroid of the negatively charged electrons is displaced by Δx from that of the positively charged nucleus, the

coulumbic force is the restoring force, which can be expressed as

$$\frac{Zq}{4\pi\epsilon_o(\Delta x)^2} \bullet Zq\left(\frac{\Delta x}{R}\right)^3 = \frac{(Zq)^2}{4\pi\epsilon_o R^3} \Delta x = \gamma \Delta x \quad (2-78)$$

In this case, the restoring force constant is

$$\gamma = \frac{(Zq)^2}{4\pi\epsilon_o R^3} \quad (2-79)$$

The restoring force can also be expressed in terms of the natural oscillation frequency ω_o by the following relation:

$$\gamma \Delta x = m\omega_o^2 \Delta x \quad (2-80)$$

Thus,

$$\omega_o = \left(\frac{\gamma}{m}\right)^{1/2} \quad (2-81)$$

However, there must be some loss or damping mechanisms to limit the forced oscillation. The oscillation of an electron is equivalent to an oscillating dipole and would radiate energy according to the electromagnetic theory of radiation. Thus, radiation can be considered one of the damping mechanisms. Therefore, a term $\beta \frac{dx}{dt}$, representing the retarding force, should be included in Equation 2-77. By including this term and substituting Equation 2-81 into Equation 2-77, we obtain

$$m \frac{d^2 \Delta x}{dt^2} + m\omega_o^2 \Delta x = -\beta \frac{dx}{dt} - ZqF_{loc} \quad (2-82)$$

where β is the damping coefficient. Damping mechanisms will be discussed further in Section 2.6. If the field is an alternating field with the frequency ω expressed as

$$F_{loc} = F_{o(loc)} e^{-j\omega t} \quad (2-83)$$

then the solution of Equation 2-82 gives the net displacement of the electrons by the applied field, which is

$$\Delta x = \frac{ZqF_{loc}}{m(\omega_o^2 - \omega^2) + j\beta\omega} \quad (2-84)$$

The induced dipole moment is

$$\vec{u} = \frac{(Zq)^2 F_{loc}}{m(\omega_o^2 - \omega^2) + j\beta\omega} = \alpha_e F_{loc} = \alpha_{e(eff)} F \quad (2-85)$$

and hence the electronic polarizability is

$$\alpha_e = \frac{(Zq)^2}{m(\omega_o^2 - \omega^2) + j\beta\omega} \quad (2-86)$$

In static fields, $\omega = 0$, and the static electronic polarizability becomes

$$\alpha_e = \frac{(Zq)^2}{m\omega_o^2} \quad (2-87)$$

Based on Bohr's model of an electron revolving around a circular orbit about the nucleus, the potential energy of the electron is given by

$$E = \hbar\omega_o = \frac{mq^4}{(4\pi\epsilon_o)^2 \hbar^2} \quad (2-88)$$

and the radius of the ground-state orbit of Bohr's atom is given by

$$R = \frac{(4\pi\epsilon_o)\hbar^2}{mq^2} \quad (2-89)$$

where $\hbar = h/2\pi$ and h is Planck's constant. When $Z = 1$, by expressing the parameters m and ω_o in terms of R and ϵ_o , and substituting them into Equation 2-87, we obtain

$$\alpha_e = \frac{q^2}{m\omega_o^2} = 4\pi\epsilon_o R^3 \quad (2-90)$$

Amazingly, Equation 2-90 turns out to be the same as Equation 2-76, although the two equations are derived on the basis of different assumptions. For the hydrogen atom, $\alpha_e = 1.57 \times 10^{-24} \epsilon_o \text{ cm}^3$. Taking $q = 1.6 \times 10^{-19} \text{ C}$ and $m = 9.11 \times 10^{-31} \text{ kg}$, we have estimated the value of ω_o from Equation 2-90; it is $4.5 \times 10^{16} \text{ radians sec}^{-1}$.

The average electronic polarizability is obtained by the summation of the contributions of all atoms divided by the number of atoms. We can consider only the contributions of the outermost electrons of each atom, and ignore the negligibly small contribution of inner electrons. From Equation 2-66, the electronic susceptibility is given by

$$\chi_e = \frac{N\alpha_e}{\epsilon_o} = \frac{N}{\epsilon_o} \left[\frac{(Zq)^2}{m\omega_o^2} \right] \quad (2-91)$$

and the relative permittivity, or dielectric constant, is

$$\epsilon_r = 1 + \chi_e = 1 + \frac{N (Zq)^2}{\epsilon_0 m \omega_0^2} \quad (2-92)$$

the natural frequency ω_0 should vary from atom to atom. However, for gases, the separation between atoms is large enough, and it is quite acceptable to assume that the local field F_{loc} , at which the atom is polarized, is equal to the applied field F and that the interaction of the outermost electrons between atoms is negligibly small and can be completely ignored. So, in this case, we can assume that each atom has the same ω_0 and hence, the same α_e . Equation 2-92 indicates that the dielectric constant increases with the increase of the density of atoms. For a hydrogen atomic gas, N is about 2.7×10^{19} atoms cm^{-3} at the standard temperature and pressure. Thus, the dielectric constant for the hydrogen atomic gas is $\epsilon_r = 1 + 2.7 \times 10^{19} \times 1.57 \times 10^{-24} = 1 + 4.24 \times 10^{-5}$. N of gases is much smaller than that of solids, which is of the order of 10^{22} to 10^{23}cm^{-3} . Furthermore, in solids, the interaction of outermost electrons between atoms cannot be ignored.

It should be noted that only in inert gases, such as He, A, etc., the quantum states of the outermost shell are completely filled, so the inert gases are atomic gases. In other gases, including hydrogen, oxygen, etc., the gases are formed by molecules such as H_2 , and O_2 instead of H or O. In this case, the particles are no longer spherical in shape, but prolate ellipsoidal or spheroidal, and we shall see that geometric shape does play an important role in determining α_e .

Single atoms in gases may be considered spherical in shape, but molecules, each comprising two or more atoms, are generally not spherical. In this case, the geometric shape does play an important role in determining α_e . We will choose a diatomic molecule with a prolate spheroidal shape to demonstrate the calculation of the electronic polarizability of this molecule. We assume that the two atoms are identical; each has an electron polarizability α_e , which is proportional to its volume, as given in Equation 2-76. There are two extreme possible arrangements, as shown in Figure 2-11(a) and (b): the molecular axis ab is parallel to the applied field ($ab \parallel F$), or the axis ab is perpendicular to F ($ab \perp F$). For the case $ab \parallel F$, the dipole of each atom consists of two components, one due to the applied field F and the other due to the field produced by the other polarized atom, which is in the same direction as F . Thus, we have

$$u_{//} = \alpha_e \left(F + \frac{u_{//}}{2\pi\epsilon_0 ab^2} \right) \quad (2-93)$$

Rearrangement of Equation 2-93 gives

$$u_{//} = \frac{\alpha_e}{1 - \alpha_e / 2\pi\epsilon_0 ab^2} F = \alpha'_e F \quad (2-94)$$

where ab is the distance between the two atoms and α'_e is the effective polarizability of each atom. Thus, the electronic polarizability of this diatomic molecule is

$$\alpha_{//} = 2\alpha'_e = \frac{2\alpha_e}{1 - 2(R/ab)^3} > 2\alpha_e \quad (2-95)$$

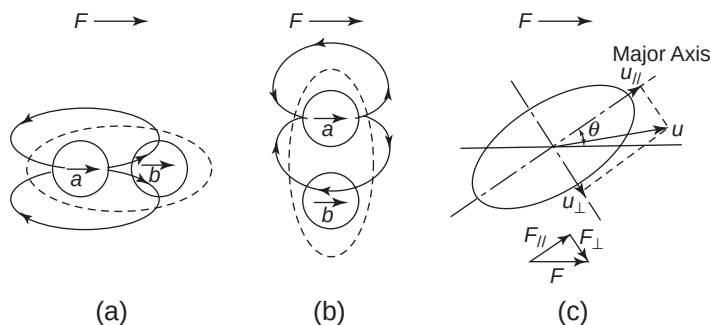


Figure 2-11 (a) The major axis ab of a spheroidal molecule parallel to the applied field F , (b) the major axis ab of a spheroidal molecule perpendicular to F , and (c) the major axis ab of a spheroidal molecule at an angle θ with respect to F .

In this case, $\alpha_{//}$ is larger than twice the polarizability of a single atom.

For the case $ab \perp F$, the dipole of each atom also consists of two components, one due to the applied field F , and the other due to the field produced by the other polarized atom, but this field is in the opposite direction to F . Thus, we have

$$u_{\perp} = \alpha_e \left(F - \frac{u_{\perp}}{4\pi\epsilon_0 ab^3} \right) \quad (2-96)$$

Rearrangement of Equation 2-96 gives

$$u_{\perp} = \frac{\alpha_e}{1 + \alpha_e/4\pi\epsilon_0 ab^3} F = \alpha_e'' F \quad (2-97)$$

where α_e'' is the effective polarizability of each atom. Thus, the electronic polarizability is

$$\alpha_{\perp} = 2\alpha_e'' = \frac{2\alpha_e}{1 + (R/\overline{ab})^3} < 2\alpha_e \quad (2-98)$$

In this case, $\alpha_{\perp}$ is smaller than twice the polarizability of a single atom. Obviously, $\alpha_{//} > \alpha_{\perp}$.

If the molecular axis is neither parallel nor perpendicular to F , but has an angle θ with respect to F , as shown in Figure 2-11(c), the total dipole moment is

$$\begin{aligned} \vec{u} &= \vec{u}_{//} + \vec{u}_{\perp} \\ &= \alpha_{//} \vec{F}_{//} + \alpha_{\perp} \vec{F}_{\perp} \end{aligned} \quad (2-99)$$

and the dipole moment in the direction of $\vec{F}$ is

$$\begin{aligned} \vec{u}F &= \vec{u}_{//} \cos \theta + \vec{u}_{\perp} \sin \theta \\ &= (\alpha_{//} \cos^2 \theta + \alpha_{\perp} \sin^2 \theta) F \end{aligned} \quad (2-100)$$

The potential energy of a dipole induced by the field is therefore

$$\begin{aligned} W &= -\frac{1}{2} (\alpha_{//} F_{//}^2 + \alpha_{\perp} F_{\perp}^2) \\ &= -\frac{1}{2} (\alpha_{//} \cos^2 \theta + \alpha_{\perp} \sin^2 \theta) F^2 \\ &= -\frac{1}{2} [\alpha_{//} - (\alpha_{//} - \alpha_{\perp}) \sin^2 \theta] F^2 \end{aligned} \quad (2-101)$$

Equation 2-101 indicates that when $\theta = 0$ or π , the potential energy is minimal, corresponding to the most stable position; when $\theta = \pi/2$, W is maximal, corresponding to the most unstable condition. If the molecules are free to rotate, as

in liquids or gases, they will tend to align their major axes with the applied field. It should be noted that the induced dipoles in the molecules due to electronic polarization also involve the orientation of the molecules in a manner similar to the orientation of the permanent dipoles.¹⁷

The field strength produced by the nucleus charges and experienced by the electrons is $Zq/4\pi\epsilon_0 R^2$ is larger than 10^9 V/cm. This field strength is far larger than the applied field. This may be the reason why the induced dipole moment for electronic polarization is proportional to the applied field and independent of frequency, because the time required for this polarization to occur is of the order of 10^{-15} sec (ω_0^{-1}). This type of polarization is also called optical polarization, because it does not vary until the frequencies are in the optical range (visible region).

Equations 2-76 and 2-87 are, of course, only a very rough approximation, because we have used a model that ignores the complications involving quantum mechanics. We have mentioned that in gases we can assume quite reasonably that the interaction between outermost electrons of atoms can be ignored and that the local field is equal to the applied field. However, in a condensed phase (i.e., in liquids or solids) the interaction between electrons of different atoms becomes very important. According to Pauli's exclusion principle, there should be many values of ω_0 , since each electron is allowed to occupy only its own state. Therefore, the electric susceptibility should be obtained by the summation of the contributions of all outermost electrons of all atoms in a piece of dielectric solid. The electric susceptibility can be written as

$$\chi_e = \sum_{i=1}^N \frac{1}{\epsilon_0} \left(\frac{(Zq)^2}{m\omega_{0i}^2} \right) \left(\frac{F_{loc}}{F} \right) \quad (2-102)$$

where N denotes the total number of atoms per unit volume. Furthermore, because the surrounding polarized atoms will modify the applied field, in this case each atom is polarized in the so-called local field, which is generally larger than the applied field. We shall discuss the local field in Section 2.5.

Quantum Mechanical Approach

The motion of electrons in an atom is governed by quantum mechanical rules. On the basis of the quantum mechanical approach, electron cloud density is assumed to vary from zero at the center of the atom to zero at far distance, with a peak around the Bohr radius R . The electronic polarizability, derived by van Vleck in 1932, is given by

$$\alpha_e(\omega) = \frac{q^2}{m} \left(\frac{f_{10}}{\omega_{10}^2 - \omega^2} \right) \quad (2-103)$$

where $\omega_{10} = (E_1 - E_0)/\hbar$ is the proper oscillation frequency for two energy levels, E_0 denotes the energy level of the ground state, E_1 denotes the energy level of one of the allowed excited states, and f_{10} represents the strength of the oscillator associated with the coupling between the wave function in the ground state and that in the allowed excited state.^{18,19} Its value is of the order of unity. It is amazing that Equation 2-103, based on the quantum mechanical approach, is similar to Equation 2-86, based on the classical approach. By setting $\omega = 0$, we have the static electronic polarizability

$$\alpha_e(o) = \frac{q^2}{m} \frac{f_{10}}{\omega_{10}^2} \quad (2-104)$$

which is again similar to Equation 2-87. If an atom has several excited levels, then Equation 2-103 must be written in generalized form, as

$$\alpha_e(\omega) = \frac{q^2}{m} \sum_{j \neq 0} \frac{f_{j0}}{\omega_{j0}^2 - \omega^2} \quad (2-105)$$

and

$$\alpha_e(o) = \frac{q^2}{m} \sum_{j \neq 0} \frac{f_{j0}}{\omega_{j0}^2} \quad (2-106)$$

where $\omega_{j0} = (E_j - E_0)/\hbar$ and j refers to the j th excited level.

Nonpolar solids are generally elemental solids, and they consist of only one kind of atom, such as diamond (C), silicon (Si), and germanium (Ge). In these materials, there are no permanent dipoles or ions. The only polarization they have is electronic polarization. In general, Equation 2-105 for α_e , derived on the basis of the displacement of electron clouds in

an isolated atom, may be applicable to gases because a gas may be considered an assembly of isolated atoms with negligible interaction among them. Equation 2-105 is not applicable to solids, where the binding force between atoms affects the movement of valence electrons, and also the energy levels are no longer discrete but form continuous bands. However, ω_{j0} is directly related to $E_g = E_c - E_v$, where E_c and E_v are, respectively, the energy levels of the conduction and the valence band edges, and E_g is the forbidden energy gap. Qualitatively, we can think from Equation 2-106 and would expect that the smaller the forbidden gap, the larger is the electron polarizability and hence, the static dielectric constant, as shown in the following table^{20,21}:

Crystal	C	Si	Ge
E_g (eV)	5.2	1.1	0.7
ϵ_{sr}	5.68	12	16

It should be noted that in crystalline solids with covalent bonds, such as Si and Ge, the calculation for α_e is quite involved,^{18,19} since the wave functions in the crystal are quite different from those in isolated atoms because of the involvement of valence electrons in the binding of atoms to form a solid. However, in ionic crystals, such as NaCl, the valence electrons are strongly localized in each ion. In this case, the electronic polarizability can be calculated simply by $\sum_i N_i \alpha_{ei}$, where N_i and α_{ei} are, respectively, the concentration of i th ions and the electronic polarizability of each i th ion. For example, for NaCl ionic crystal, the energy band gap $E_g = 7$ eV, the values of electronic polarizability^{21,22} for Na^+ and Cl^- are, respectively, 0.20×10^{-40} and 2.65×10^{-40} farad m^2 . From Equation 2-66, we have calculated the relative permittivity (dielectric constant) to be about 2.28 by assuming the concentration of NaCl molecules $N = 4 \times 10^{28} \text{m}^{-3}$ and $\alpha = (0.20 + 2.65) \times 10^{-40}$ farad m^2 . This value is very close to the experimental value of 2.25. Note that the static dielectric constant of NaCl

crystals is 5.62. The difference between 5.62 and 2.28 is due to the contribution of ionic polarizability, which will be discussed in the following section. Similarly, in molecular solids where the atoms or molecules are bonded by weak van der Waals forces, each particle (atom or molecule) can be considered a quasi-isolated particle. In this case, the electronic polarizability of the solid can also be calculated by $\sum_i N_i \alpha_{ei}$, where N_i is the concentration of i th particles and α_{ei} is the electronic polarizability of each i th particle. For condensed matter (liquids or solids) made of single inert (or noble) atoms, such as He, neon, or argon, the total electronic polarizability can be simply calculated by $N\alpha_e$. In this case, N is just the concentration of atoms in the material.

2.3.2 Atomic or Ionic Polarization (Vibrational Polarization)

A dielectric material consisting of polyatomic molecules usually has electronic polarization and ionic polarization, and in some cases, also orientational polarization when it is subjected to an electric field. In general, there are two groups of ionic solids. One group does not possess permanent dipoles, such as NaCl, which forms a simple cubic lattice so that the lattice symmetry and the overall charge neutrality ensure that electric dipoles formed by each ion pair everywhere cancel each other. The other group possesses permanent dipoles, because the crystal lattice in this case is less symmetrical, as with HCl. In fact, the internal field at the positive ion sites is generally different from that at the negative ion sites. In general, most ionic solids are asymmetrical and the electronegativities of both ions are different. They possess permanent dipoles, but these dipoles are foreseen in the solid state and cannot be aligned by an electric field. This is why in most ionic solids belonging to this group, the permanent dipole moments do not contribute to the polarizability in the solid state although the materials possess them.

Quantitative analysis of the ionic polarization is quite involved.²²⁻²⁴ In this section, we

will consider only a simple case to illustrate the basic concept of atomic polarization. Suppose we have a molecule consisting of two atoms: A and B . Atom A is ready to give part of its valence electrons to atom B in order to make the outermost shells of both atoms more fully completed. This tendency is, in fact, the force that produces the so-called ionic bond. In this case, atom A is more electropositive and atom B more electronegative. NaCl is a good example, in which atom A is Na and atom B is Cl. The one electron in the outermost shell of Na is not completely given to Cl (Na gives, on average, about 78% of the electron to Cl). However, Cl receives some of the valence electron from Na and becomes a negative ion; Na becomes a positive ion.

Let us consider a linear chain of ions A and ions B placed at equal distances along the x direction, as shown in Figure 2-12(a). In thermal equilibrium and in the absence of an electric field, the positive ions A at x_{2n} , x_{2n+2} , x_{2n-2} , etc., and the negative ions B at x_{2n+1} , x_{2n+3} , x_{2n-1} , etc., will always undergo lattice vibrations, but their interatomic distance is $x_{2n+1} - x_{2n} = a$, on average. Now if a step-function electric field is applied in the x direction, the electron clouds immediately shift to the left, and it takes only about 10^{-15} sec to produce electronic polarization, as shown in Figure 2-12(b). In about 10^{-13} sec after the application of the field, the positive ion at x_{2n} and the negative ion at x_{2n+1} tend to attract and move toward each other, making the interatomic distance $\Delta x_1 = x_{2n+1} - x_{2n} < a$ and $\Delta x_2 = x_{2n} - x_{2n-1} > a$. The same tendency prevails in other ions. The displacement in both the electron clouds and the ions themselves produces electronic polarization, as well as ionic polarization, as shown in Figure 2-12(c).

Displacement of the atoms (i.e., ions) from their equilibrium sites by Δx will generate a force that tends to bring them back to their original thermal equilibrium sites. Within the approximation of harmonic oscillation, the elastic restoring force is proportional to the difference between displacements of neighboring ions. Thus, from Figure 2-12(c), we can write

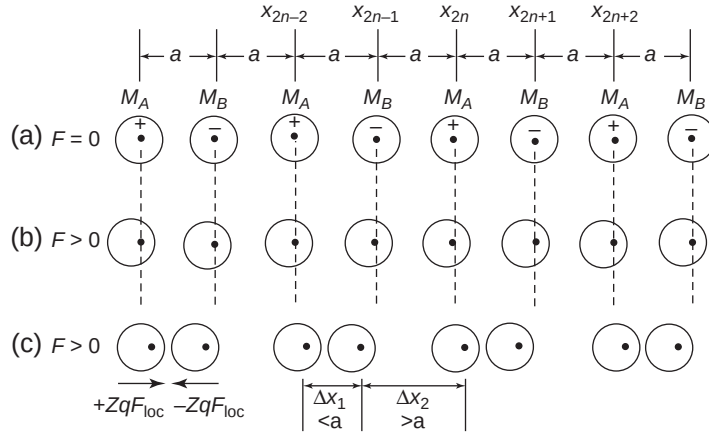


Figure 2-12 Illustrating a linear chain of atoms A with mass M_A and atoms B with mass M_B placed at equal spacing a along the x direction. (a) In the absence of an electric field $F = 0$, (b) in the presence of an electric field $F > 0$, only electron clouds are shifted to the left, and (c) also in the presence of an electric field $F > 0$, but in this case the positive ions and the negative ions tend to move toward each other, resulting in ionic polarization, which involves both the displacement of electron clouds and the movement of ions.

the equations of motion for ions M_A and M_B as follows:

$$M_A \frac{d^2 x_{2n}}{dt^2} = \gamma_i [(x_{2n+1} - x_{2n}) - a] - \gamma_i [(x_{2n} - x_{2n-1}) - a] = \gamma_i [x_{2n+1} + x_{2n-1} - 2x_{2n}] \quad (2-107)$$

$$M_B \frac{d^2 x_{2n-1}}{dt^2} = \gamma_i [(x_{2n} - x_{2n-1}) - a] - \gamma_i [(x_{2n-1} - x_{2n-2}) - a] = \gamma_i [x_{2n} + x_{2n-2} - 2x_{2n-1}] \quad (2-108)$$

where γ_i is the restoring force constant, which is different from γ in Equation 2-79. γ_i can be calculated on the basis of only the electrostatic interaction between the ion and its nearest neighbors. Suppose, for simplicity, the positive and the negative ions have net charges $Z_A q$ and $-Z_B q$, respectively, and the net displacement from their equilibrium sites is Δx , then the coulombic force between the two ions is

$$\frac{Z_A Z_B q^2}{4\pi\epsilon_o (a - \Delta x)^2} \approx \frac{Z_A Z_B q^2}{4\pi\epsilon_o a^2} + \frac{2Z_A Z_B q^2}{4\pi\epsilon_o a^3} \Delta x \quad (2-109)$$

since $\Delta x \ll a$. We retain only the first two terms. The first term is the force of interaction

corresponding to the equilibrium sites, and the second term is the elastic restoring force

$$\frac{2Z_A Z_B q^2}{4\pi\epsilon_o a^3} \bullet \Delta x = \gamma_i \Delta x \quad (2-110)$$

In general, $Z_A = Z_B = Z$. For NaCl, $Z = 0.78$. So the force constant is

$$\gamma_i = \frac{2Z^2 q^2}{4\pi\epsilon_o a^3} \quad (2-111)$$

Equations 2-107 and 2-108 are equations of phase waves and therefore have the solutions of phase wave form

$$x_{2n} = x_I \exp[jk(2n)a] \quad (2-112)$$

$$x_{2n-1} = x_{II} \exp[jk(2n-1)a] \quad (2-113)$$

where x_I and x_{II} are periodic functions of time and the exponential is a phase factor. These equations indicate that the phase of the elastic wave changes by $2ka$ from one ion to the nearest ion of the same kind, such as from M_A at x_{2n-2} to another M_A at x_{2n} . k is the wave number, which may be taken as $2\pi/\lambda$ with λ as the wavelength of the elastic wave. For lattice vibrations, in the visible and infrared regions, λ is of the order of 3×10^{-5} cm, which is considerably larger than the interatomic distance of

about 3×10^{-8} cm. So we can say that ka is extremely small and is much smaller than unity for the frequency range associated with the lattice waves. Substituting Equations 2-112 and (2-113 into Equations 2-107 and 2-108, and bearing in mind that $\exp(jka)$ is equal to $\cos ka$ (which can be practically considered as $\cos ka \approx 1$), we obtain

$$M_A \frac{d^2 x_I}{dt^2} = -2\gamma_i (x_I - x_{II}) \quad (2-114)$$

$$M_B \frac{d^2 x_{II}}{dt^2} = -2\gamma_i (x_{II} - x_I) \quad (2-115)$$

Under an electric field, the equations of motion for ions should include both the elastic force and the electric force, as well as the damping or retarding force. Therefore, the above equations must be modified to

$$M_A \frac{d^2 x_I}{dt^2} = -2\gamma_i (x_I - x_{II}) - \beta \frac{d(x_I - x_{II})}{dt} + ZqF_{loc} \quad (2-116)$$

$$M_B \frac{d^2 x_{II}}{dt^2} = -2\gamma_i (x_{II} - x_I) - \beta \frac{d(x_{II} - x_I)}{dt} - ZqF_{loc} \quad (2-117)$$

Equation 2-117 $\times M_A$ minus Equation 2-116 $\times M_B$ yields

$$M_r \frac{d^2 \Delta x}{dt^2} = -2\gamma_i \Delta x - \beta \frac{d\Delta x}{dt} - ZqF_{loc} \quad (2-118)$$

where $M_r = M_A M_B / (M_A + M_B)$ is the reduced mass and $\Delta x = \Delta x_{II} - \Delta x_I$ is the relative displacement of positive and negative ions. Following the same principle for Equation 2-81, we can express γ_i in terms of the natural oscillation (or lattice vibration) frequency ω_o as $2\gamma_i = M_r \omega_o^2$. If the applied electric field is an alternating field with the frequency ω , as given by Equation 2-83, then the solution of Equation 2-118 gives

$$\Delta x = \frac{ZqF_{loc}}{M_r(\omega_o^2 - \omega^2) + j\beta\omega} \quad (2-119)$$

The induced ionic dipole moment is

$$\begin{aligned} \mu_i &= Zq\Delta x = \frac{(Zq)^2 F_{loc}}{M_r(\omega_o^2 - \omega^2) + j\beta\omega} \\ &= \alpha_i F_{loc} = \alpha_{i(eff)} F \end{aligned} \quad (2-120)$$

Thus, the ionic (or atomic) polarizability is

$$\alpha_i = \frac{(Zq)^2}{M_r(\omega_o^2 - \omega^2) + j\beta\omega} \quad (2-121)$$

In static fields, $\omega = 0$ the static ionic polarizability becomes

$$\alpha_i = \frac{(Zq)^2}{M_r \omega_o^2} \quad (2-122)$$

The expression for α_i is similar to that for α_e . It should be noted that the distinction between electron and ionic polarizations is not sharp, because a displacement of ions is always accomplished by a displacement of electrons. Generally, it is much easier to measure the electronic polarizability than the ionic polarizability. Usually, the electronic polarizability is determined by measuring the refractive index in the visible or ultraviolet region, and the ionic polarizability is then determined by extrapolating the refractive index spectrum to frequencies much lower than the visible region—generally, in the infrared region, in which both α_e and α_i are the dominant polarizabilities. The time required for electronic polarization is about 10^{-15} sec, and that required for the ionic polarization is about 10^{-13} sec, simply because ions are heavier than electrons by more than 10^3 times. This is why the resonances for these two polarizations occur in different frequency regions.

2.3.3 Orientational Polarization

To begin, it may be desirable to show why some molecules possess permanent dipole moments and some do not. For a molecule with two atoms A and B , atom A gives some of its valence electrons to atom B , then atom A becomes a positive ion and atom B becomes a negative ion. This ionic bond molecule obviously possesses a permanent dipole moment, which is the product of the charge of the portion of the valence electrons transferred from atom A to atom B and the interatomic distance. If a

great many such molecules form an ionic crystal, such as an NaCl crystal, the vector sum of all the dipole moments vanishes, because they tend to cancel each other. So, the crystal itself is not a dipolar material. For some materials, a molecule consists of three atoms in the form AB_2 or A_2B . For example, CO_2 is in the form of AB_2 . In this molecule, the bonding structure is symmetrical, so the centroids of the positive and the negative charges are coincident, thus the resultant dipole moment is zero, as shown in Figure 2-13(a). Similarly, molecules consisting of similar atoms, such as H_2 , O_2 , and N_2 , carry no permanent dipole moments. However, H_2O is in the form of A_2B , but in this case, the bonding structure is asymmetrical, so the centroid of the negative charge is not coincident with that of the positive charge, thus resulting in a net permanent dipole moment, i.e., the vector sum of the two individual dipole moments gives a resultant dipole moment, as shown in Figure 2-13(b).

In the presence of an electric field F , the molecules carrying a permanent dipole moment

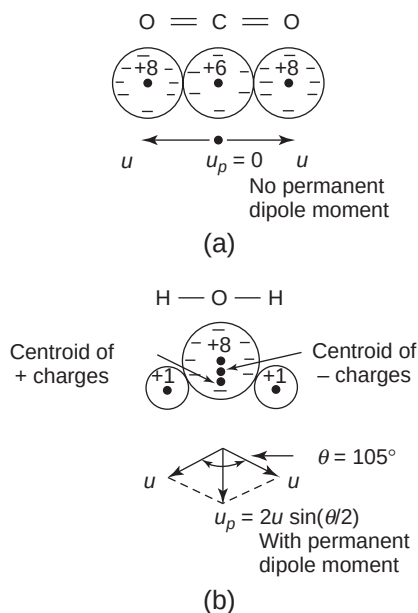


Figure 2-13 The formation of dipole moments for two typical molecules: (a) molecule CO_2 having no net permanent dipole moment and (b) molecule H_2O having net permanent dipole moment.

will orient to align the permanent dipole moments along the direction of F . This process is referred to as orientational polarization, which occurs only in dipolar materials possessing permanent dipoles. It is obvious that orientation of a molecule involves the energy required to overcome the resistance of the surrounding molecules, so the orientation process is strongly temperature dependent.

In general, nonmagnetic dielectric materials can be considered paraelectric materials, the counterpart to paramagnetic materials, in which the induced electric susceptibility $\chi > 0$ is always positive. There is no counterpart to diamagnetic materials in dielectric materials, in which χ should be negative. In paraelectric materials, α_e and α_i can be considered practically independent of temperature for the normal temperature range in most applications, because the electronic structure does not change in the normal temperature range. However, the orientational polarizability α_o is strongly temperature dependent. Figure 2-14 shows the total polarization of some vapors as a function of temperature. The data are from the references.²⁵⁻²⁸ It can be seen that C_2H_2 and molecules with a similar symmetrical and linear structure, such as CH_4 , CCl_4 , and CO_2 , do not possess permanent dipole moments and therefore are not dipolar. Those with an asymmetrical structure are dipolar, such as H_2O , in which two OH bonds make an angle of 105° , resulting in the formation of a permanent dipole moment. This is why the total polarization in which the orientational polarization is dominant decreases with increasing temperature, as shown in Figure 2-14. It should be noted that for gases at normal temperature and pressure, susceptibility is of the order of 10^{-4} to 10^{-3} . The difference in susceptibility reflects the difference in the number of atoms per unit volume. However, for solids, susceptibility is generally much higher than unity.

Molecules having a permanent dipole moment experience a torque in an electric field, tending to orient themselves to the field direction, but thermal agitation tends to put them back into random orientation. However, the ensemble of molecules will gradually attain a

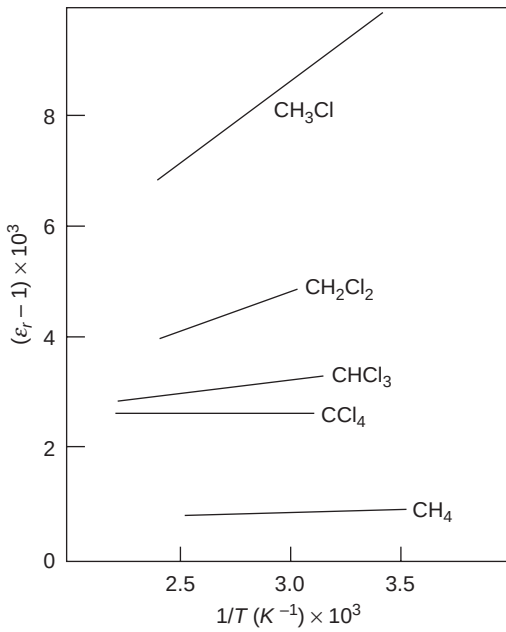


Figure 2-14 The susceptibility ($\epsilon_r - 1$) as a function of temperature for some polar and nonpolar gases at the normal pressure of 1 atmosphere.

statistical quasi-equilibrium. The method of calculating orientational polarizability was first developed for permanent magnetic moments in paramagnetic materials by Langevin²⁹ and later applied to permanent dipole moments in dielectric materials by Debye.^{30,31}

Supposing that the permanent dipole moment u_o of the molecule is not affected by the applied electric field and temperature, and that the density of molecules in the gas medium as an ensemble is so small that the dipole-dipole interaction is very small compared to the thermal equilibrium energy kT ; the moment of the permanent dipole in the direction of the applied field F , as shown in Figure 2-15, can be written as

$$\langle u_F \rangle = u_o \langle \cos \theta \rangle \quad (2-123)$$

where θ is the angle between the dipole moment and the applied field, and $\langle \rangle$ represents the statistical ensemble average.

At zero applied field, the number of dipoles having an inclination of their axes to the axes

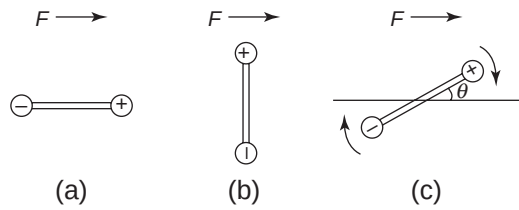


Figure 2-15 The orientation of a dipole: (a) stable position, (b) unstable position, and (c) orienting to the field direction.

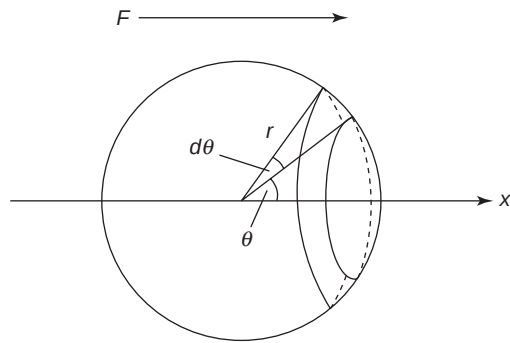


Figure 2-16 The calculation of the number of dipoles having the inclination of their axes to the x-axis between θ and $\theta + d\theta$.

located between θ and $\theta + d\theta$, as shown in Figure 2-16, is

$$dN = N \frac{2\pi r \sin \theta r d\theta}{4\pi r^2} = \frac{1}{2} N \sin \theta d\theta \quad (2-124)$$

With an applied field F and the effect of temperature, Equation 2-124 is changed to the following form, according to Boltzmann's statistics,

$$dN = A \exp(-U/kT) \frac{1}{2} \sin \theta d\theta \quad (2-125)$$

where U is the potential energy of the dipole at θ , which is given by

$$U = u_o F \cos \theta \quad (2-126)$$

Thus, the average dipole moment in the field direction is

$$\begin{aligned}
\langle u_F \rangle &= \frac{\int u_o \cos \theta dN}{\int dN} \\
&= u_o \frac{\int_0^\pi \exp(u_o F \cos \theta / kT) \cos \theta \sin \theta d\theta}{\int_0^\pi \exp(u_o F \cos \theta / kT) \sin \theta d\theta} \\
&= u_o \langle \cos \theta \rangle
\end{aligned}
\tag{2-127}$$

By introducing

$$y = \cos \theta \tag{2-128}$$

$$z = u_o F / kT \tag{2-129}$$

we obtain

$$\begin{aligned}
\langle \cos \theta \rangle &= \frac{\int_{-1}^{+1} \exp(yz) y dy}{\int_{-1}^{+1} \exp(yz) dy} \\
&= \coth z - 1/z = L(z)
\end{aligned}
\tag{2-130}$$

The function $L(z)$ is called the Langevin function, which is shown in Figure 2-17 as a function of $z = u_o F / kT$.

At low values of z , $L(z)$ varies almost linearly, with z following the relation

$$L(z) = \langle \cos \theta \rangle = u_o F / 3kT \tag{2-131}$$

For example, a dipole moment $u_o = 10^{-8}$ qC – cm at a field of 10^6 V/cm and a temperature of 300 K gives $u_o F / kT \approx 0.4$, which is smaller than 1. There is no possibility for $\langle \cos \theta \rangle$ to reach unity at room temperature. Only at very low temperatures may it approach unity at high fields. For $z \ll 1$, Equation 2-131 is reasonably accurate, and for $z \gg 1$, $L(z)$ can be estimated by

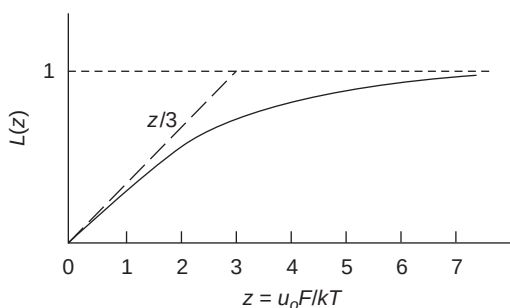


Figure 2-17 The Langevin function.

$$L(z) = \langle \cos \theta \rangle = 1 - 1/z \tag{2-132}$$

However, as F/T increases, the dipole of the molecule becomes more parallel to the applied field. In practical and most cases, $u_o F / kT \ll 1$. Thus, substitution of Equation 2-131 into Equation 2-123 gives

$$\langle u_F \rangle = \frac{u_o^2 F}{3kT} \tag{2-133}$$

and the orientational polarizability is

$$\alpha_o = \frac{u_o^2}{3kT} \tag{2-134}$$

In general, orientational polarizability is much larger than electronic and atomic polarizabilities at normal conditions, since α_e and α_i are practically independent of temperature (i.e., their temperature dependence at normal conditions is not significant) but α_o is strongly temperature dependent. Thus, α_o can be easily distinguished from α_e and α_i by the temperature dependence measurement of ϵ_r .

There are two factors that have not been taken into account in the derivation of α_o . One is the effect of electron spins, which cause spin paramagnetism and may affect the results. To take this effect into account, the quantum mechanical approach must be used. The other factor is that the permanent dipole moment u_o for multiatomic molecules is not independent of temperature because in this case, the permanent dipole moment is the result of several moments, and their internal orientation is dependent on their individual activation energies and hence on the temperature.

In solids, the dipoles do not rotate freely as do dipoles in liquids or gases. The rotation of dipoles in solids is constrained to a few discrete orientations, which are influenced by the crystalline field determined by the interaction of the dipole with neighboring ones. Thus, to describe the dielectric constant in terms of the dipole moments u_o of individual dipoles (molecules), it is necessary to know the crystal structure of the material in the solid state.

The potential energy of a dipole in a crystalline solid depends on the direction of the dipole relative to the crystal axes. In other

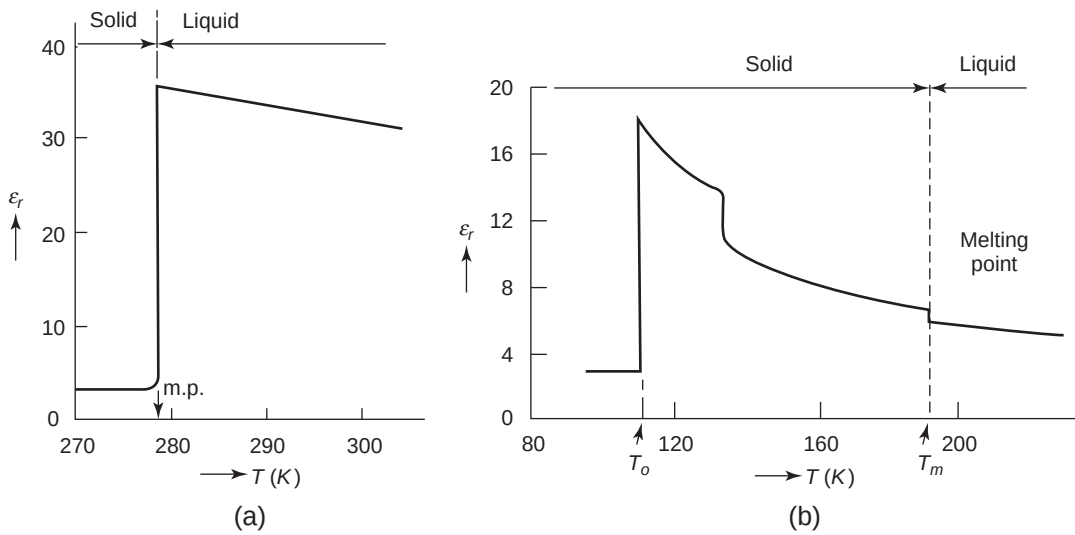


Figure 2-18 Variation of dielectric constant (measured at low frequency 5 kHz) with temperature for (a) nitrobenzene ($\text{C}_6\text{H}_5\text{CO}_3$), and (b) hydrogen sulfide (H_2S).

words, the potential energy is directly related to the crystalline field acting on the dipole and is therefore temperature dependent. In some solids, dipoles at temperatures below the melting point are frozen, and their rotation is inhibited even under an electric field. In such solids, the contribution of α_o to the total polarization is zero, i.e., $\alpha_o = 0$, as in nitrobenzene ($\text{C}_6\text{H}_5\text{CO}_3$), shown in Figure 2-18(a). The dielectric constant decreases abruptly from the value of 35 (in liquid state) to a value of about 3 (in solid state) at 278 K, the melting point of this material. For $T < 278$ K, only α_e and α_i contribute to ϵ_r . For $T > 278$ K, the material is in the liquid state; α_e , α_i , and α_o all contribute to ϵ_r , and ϵ_r decreases with increasing temperature following Equation 2-134. However, in some other solids, the dielectric constant is still increasing with decreasing temperature at temperatures below the melting point, i.e., in the solid state. This rise in dielectric constant continues until a critical temperature T_o , at and below which the dipoles become frozen. H_2S is a typical example of this kind of material. Figure 2-18(b) shows that, for temperatures between the melting point T_m (188 K) and the critical temperature T_o (103 K), the dielectric

constant of H_2S in the solid state increases with decreasing temperature, but at $T_o = 103$ K, ϵ_r drops abruptly from about 20 to about 3, indicating that the dipoles at T_o are frozen and become immobile, resulting in $\alpha_o \rightarrow 0$. The experimental results are from Smyth and Hitchcock.^{32,33}

To explain why H_2S and some other materials, such as HCl , still have the temperature dependence of α_o in the solid state within a certain range of temperatures between T_m and T_o , we can assume that the average potential energy of a dipole due to short-range interaction, which creates the crystalline field, has a profile similar to that shown in Figure 2-19. The bottom of the potential wells represents the equilibrium state of the dipoles in which the potential energy of the dipoles reaches the minimum value. In the absence of an electric field, a dipole has the same probability of orienting to the right or to the left, so there is no net polarization. But when the material is present in an electric field, the well at $\theta = 0$ is lowered by an amount of $u_o F$, and that, at $\theta = \pi$, is deepened by the same amount of $u_o F$, as shown by the dashed line in Figure 2-19. The direction of the dipoles orienting to the right

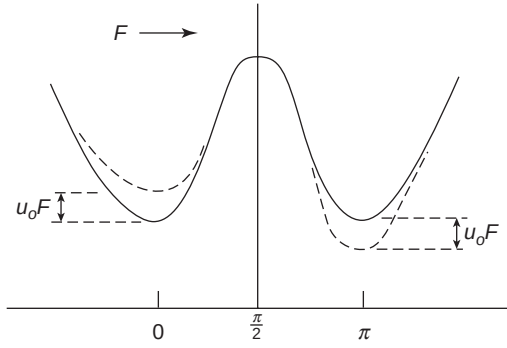


Figure 2-19 Schematic diagram showing the potential energy of a dipole as a function of the orientation angle θ in the solid state. Solid curve—in the absence of an electric field and dashed curve—in the presence of an electric field.

corresponds to the orientation of the dipoles toward the direction of the applied field, and that to the left corresponds to the orientation of the dipoles opposite to the field. Denoting the probability of the dipoles at $\theta = \pi$ orienting to the left as w , then $1 - w$ is the probability of the dipoles at $\theta = 0$ orienting to the right, and according to the Boltzmann statistics, we can write

$$\frac{w}{1-w} = \exp[-2u_o F/kT] \quad (2-135)$$

This leads to

$$w = \frac{\exp(-2u_o F/kT)}{1 + \exp(-2u_o F/kT)} \quad (2-136)$$

In general, $u_o F$ is much smaller than kT under most practical conditions. For $u_o F \ll kT$, Equation 2-136 can be simplified to

$$w = \frac{1}{2} \exp[-2u_o F/kT] \quad (2-137)$$

Based on this model, the percentage of a dipole moment u_o in the direction opposite to the applied field is w , than the percentage of the dipole moment in the direction parallel to the field is $1 - w$. Thus, we can write

$$\begin{aligned} \vec{u}_o(\text{parallel to } \vec{F}) &= u_o(1-w) - u_o(w) \\ &= u_o(1-2w) \\ &= u_o[1 - \exp(-2u_o F/kT)] \end{aligned} \quad (2-138)$$

The term $\exp(-2u_o F/kT)$ can be expanded into a series because $u_o F \ll kT$. This leads to

$$\begin{aligned} \exp(-2u_o F/kT) &= 1 - \frac{2u_o F}{kT} + \frac{1}{2} \left(\frac{2u_o F}{kT} \right)^2 \\ &\quad - \frac{1}{3!} \left(\frac{2u_o F}{kT} \right)^3 + \dots \end{aligned}$$

By keeping only the first two terms, we obtain

$$\vec{u}_o(\text{parallel to } \vec{F}) = \frac{2u_o^2 F}{kT} \quad (2-139)$$

This leads to the polarizability

$$\alpha_o = \frac{\vec{u}}{F} = \frac{2u_o^2}{kT} \quad (2-140)$$

which is in the same form as that of Equation 2-134 derived on the assumption that the dipoles rotate freely in the material. However, the difference between these two equations explains that at $T = T_m$, α_o in the solid state is larger than in the liquid state, as shown in Figure 2-18(b).

The transition from one form of the solid state at $T_o < T < T_m$ to the form at $T < T_o$ can be considered the transition from a disordered state to an ordered state. At $T = 0$, all dipoles are in a completely ordered state. That at $T = 0$, $\alpha_o = 0$ implies that the number of dipoles pointing in one direction is equal to that pointing in the opposite direction, so the net polarization vanishes. As the temperature increases, the number of dipoles in the ordered state will tend to move to the disordered state. Considering that the dipoles have their lowest potential energy in the completely ordered state, there exists a potential difference $U(T)$ between two equilibrium states, as shown in Figure 2-20(a).

It can be imagined that the probability for a dipole having its direction at $\theta = 0$, termed the right direction, is w , then a lower probability (zero at $T = 0$) for the dipole having its direction at $\theta = \pi$, termed the wrong direction, is $1 - w$, as shown in Figure 2-20(a). Similarly, another dipole may behave oppositely; its direction at $\theta = 0$ is the wrong direction, and that at $\theta = \pi$ is the right direction,^{28,34,25} as shown in Figure 2-20(a). We can use the same approach for the polarization at $T_o < T < T_m$; the probability for a dipole in the wrong direction

is $1 - w$, which is governed by the following relation:

$$\frac{w}{1-w} = \exp[-U(T)/kT] \quad (2-141)$$

Hence, we have

$$w = \frac{\exp[-U(T)/kT]}{1 + \exp[-U(T)/kT]} \quad (2-142)$$

The calculation of $U(T)$ is very involved, even by means of approximation methods.^{28,36} We shall not give details about the calculation here.

A qualitative picture of $U(T)$ as a function of T , as shown in Figure 2-20(b), is sufficient to explain the temperature dependence of ϵ_r at $T < T_o$. Since $U(T) \gg kT$, the energy of the dipole due to the applied field $u_o F$ is negligibly small compared to $U(T)$; therefore, the field F does not affect the order-disorder transition for $T \leq T_o$. According to Equation 2-142 and Figure 2-20(c), w is very small for $T \leq T_o$ and becomes $1/2$ at $T = T_o$, at which $U(T) = 0$, as shown in Figure 2-20(b). This means that $w = 1/2$ for $T \geq T_o$, implying that the probability for the dipoles pointing to the right or to the left is

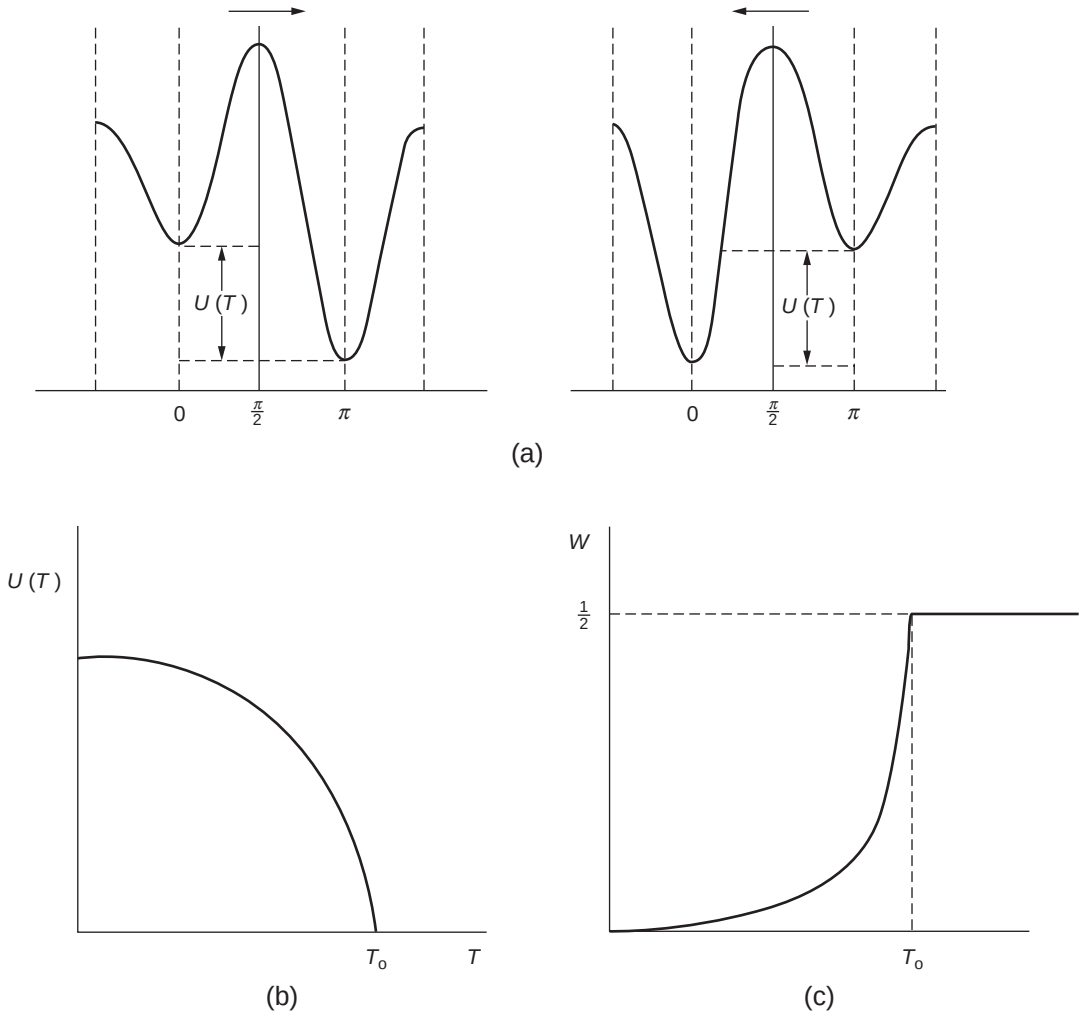


Figure 2-20 (a) Potential energy profiles of two different types of dipoles in the lattice, (b) $U(T)$ as a function of temperature, and (c) the probability w as a function of temperature.

equal. Under this situation, the applied field will start to play an important role in causing polarization, as discussed earlier for T between T_o and T_m .

2.3.4 Spontaneous Polarization

Electronic polarization is always present in atoms or molecules in all kinds of materials; ionic polarization is present only in materials made of two or more different kinds of atoms that form ions due to the sharing of the valence electrons of one or more atoms with the others. For both electronic and ionic polarizations, the dipole moments are induced by electric fields, so they are classified as induced dipole moments. In short, both electronic and ionic polarizations are due to the translation (or deformation) of the valence electron clouds from their original thermal equilibrium state to a new equilibrium state. These types of polarization are only slightly dependent on temperature because they are intramolecular phenomena. However, orientational polarization occurs only in the materials composed of molecules with an asymmetrical structure in which the centroid of the negative charge (mainly electrons) and that of the positive charge (mainly nuclei) are not coincident, so they possess permanent dipole moments in the absence of external fields. The directions of these permanent dipole moments are randomly distributed in the material. An electric field will cause them to reorient toward the direction of the field, resulting in orientational polarization. The net polarization will return to zero after the removal of the external field because thermal agitation tends to randomize the alignment. This is why polarization decreases with increasing temperature.

However, there is another kind of polarization, called spontaneous polarization. By analogy to magnetization, electric polarization can be grouped into two major polarizations:

- Paraelectric polarization, which includes mainly electronic, ionic, and orientational polarizations, with χ always positive
- Ferroelectric polarization, with χ very large, similar to ferromagnetization

There is no counterpart to diamagnetization in dielectric materials, with χ in negative values. Spontaneous polarization occurs in materials whose crystalline structure exhibits electrical order. This implies that spontaneous polarization occurs only in single crystals or crystallites in polycrystalline materials with a noncentrosymmetric structure, because only in a noncentrosymmetric structure does the centroid of the negative charges not coincide with that of the positive charges. In ferroelectric materials, electric polarization occurs spontaneously due to a phase transition at a critical temperature called the Curie temperature, T_c , without the help of an external electric field. At and below T_c , the crystal undergoes a phase transition, usually from a nonpolar cubic structure to a polar structure. BaTiO_3 is a typical example of a ferroelectric crystal. At $T > T_c$, the BaTiO_3 crystal assumes a cubic structure in which the centroid of the negative charges and that of the positive charges coincide, so the molecule does not form a dipole moment. However, at $T \leq T_c$, the cubic structure is slightly distorted, resulting in a slight displacement of Ba^{2+} and Ti^{4+} ions. This displacement, though very small (only about $0.15\text{\AA}$), is enough to cause the centroid of the negative charges to be different from that of the positive charges, thus forming an electric dipole moment. Each unit cell carries a reversible electric dipole moment spontaneously oriented parallel to the dipole moment of neighboring cells. This chain-reaction process is referred to as spontaneous polarization.

The build-up of such dipole moments pointing along one crystal axis will gradually form a domain and will gradually increase the free energy of the system. This process cannot continue; the domain will stop growing when it has reached a certain size, and another domain with dipole moments pointing in the opposite direction will be formed in order to reduce the free energy of the system. In a single crystal or a crystallite, there are many domains with moments pointing in various directions. But the vector sum of the dipole moments of all domains vanishes. Each domain can be considered as a large dipole. Under an external

electric field, all of these randomly arranged domains tend to move toward the direction of the field, resulting in a net total spontaneous polarization. Upon the removal of the field, spontaneous polarization does not vanish but remains inside the material. The field-polarization relation forms a hysteresis loop similar to the hysteresis loop for ferromagnetic materials. This topic will be discussed in more detail in Ferroelectric Phenomena in Chapter 4.

2.3.5 Space Charge Polarization

The induced, orientational, and spontaneous polarizations discussed previously are due to the bound positive and negative charges within the atom or the molecule itself, which are linked intimately to each other and which normally cannot be separated. However, electric polarization may also be associated with mobile and trapped charges. This polarization is generally referred to as space charge polarization, P_d . This occurs mainly in amorphous or polycrystalline solids or in materials consisting of traps. Charge carriers (electrons, holes, or ions), which may be injected from electrical contacts, may be trapped in the bulk or at the interfaces, or may be impeded to be discharged or replaced at the electrical contacts. In this case, space charges will be formed, field distribution will be distorted, and hence, the average dielectric constant will be affected. In the following sections, we shall consider two possible ways in which space charge polarization may result.

Hopping Polarization

In a dielectric material, localized charges (ions and vacancies, or electrons and holes) can hop from one site to the neighboring site, creating so-called hopping polarization.¹⁶ These charges are capable of moving freely from one site to another site for a short time, then becoming trapped in localized states and spending most of their time there. Occasionally, these charges make a jump surmounting a potential barrier to other sites. In fact, the movement of ions or vacancies in ionic crystals and the movement of electrons and holes in glasses and amorphous

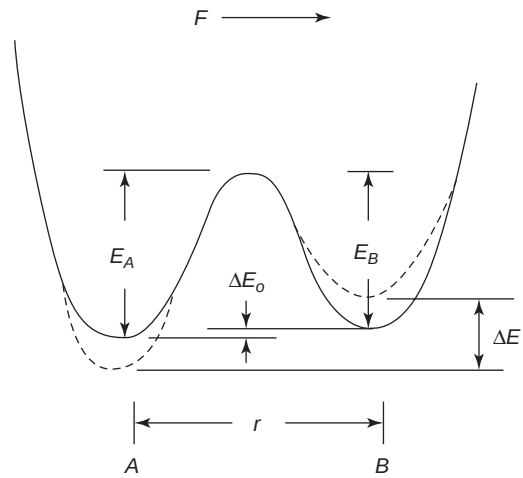


Figure 2-21 A double potential well illustrating hopping polarization due to the hopping of the charged particle over the barrier from one well to the other.

semiconductors are essentially due to the hopping process.³⁷ Depending on the width and the height of the potential barrier, a charged particle on one site may hop or tunnel to the other site. A simple, double-well potential barrier is shown in Figure 2-21.

In thermal equilibrium, the time-averaged probability for a charged particle (for example, a negatively charged particle) to hop from site A to site B, leaving a positive charge in site A and creating a negative charge in site B to form a dipole, is

$$P_{o(A \rightarrow B)} = C \exp(-E_A/kT) \quad (2-143)$$

where C is a constant and E_A is the activation energy for the hopping transition. Similarly, the probability for the charged particle to hop in the reverse direction from site B to site A is

$$P_{o(B \rightarrow A)} = C \exp\left(\frac{-E_A - \Delta E_o}{kT}\right) \quad (2-144)$$

where ΔE_o is the difference between E_A and E_B , which is positive when $E_A > E_B$, zero when $E_A = E_B$, or negative when $E_A < E_B$. An applied field changes the potential barrier profile, making the hopping of the particle from site B to site A much easier than from site A to site B (for the case shown in Figure 2-21). Thus, the probabilities under an applied field become

$$P_{(A \rightarrow B)} = C \exp[-(E_A + a\Delta E')/kT] \quad (2-145)$$

and

$$P_{(B \rightarrow A)} = C \exp\{-[E_A - (1-a)\Delta E']/kT\} \quad (2-146)$$

where $\Delta E' = \Delta E - \Delta E_o$ is the potential barrier height difference created by the applied field F , $a(\Delta E')$ is the increase in potential barrier height in site A , and $(1-a)\Delta E'$ is the decrease in potential barrier height in site B . Obviously, the magnitude of $\Delta E'$ depends on the direction of the vector r (the axis joining site A and site B) relative to F . Therefore, $\Delta E'$ can be written as

$$\Delta E' = \Delta E - \Delta E_o = qFr \cos \theta \quad (2-147)$$

This implies that $\Delta E' = qFr$ when $\theta = 0$, i.e., the vector r is in the field direction; $\Delta E' = 0$ when $\theta = \pi/2$, i.e., the vector r is perpendicular to the field direction. In this case, the applied field does not modify the original potential barrier height. In terms of probabilities, we can write

$$P_{(A \rightarrow B)} = P_{o(A \rightarrow B)} - p' \quad (2-148)$$

and

$$P_{(B \rightarrow A)} = P_{o(B \rightarrow A)} + p' \quad (2-149)$$

The applied field causes a decrease in hopping probability from site A to site B by an amount p' and an increase in hopping probability from site B to site A by p' . From Equations 2-143 and 2-147 we have

$$\frac{P_{o(B \rightarrow A)}}{P_{o(A \rightarrow B)}} = \exp(\Delta E_o/kT) \quad (2-150)$$

$$\frac{P_{(B \rightarrow A)}}{P_{(A \rightarrow B)}} = \exp(\Delta E/kT) = \exp[(\Delta E_o + \Delta E')/kT] \quad (2-151)$$

Assuming that the particle must be located either in site A or in site B , then

$$P_{o(A \rightarrow B)} + P_{o(B \rightarrow A)} = P_{(A \rightarrow B)} + P_{(B \rightarrow A)} = 1 \quad (2-152)$$

From Equations 2-148 and 2-152, we obtain

$$p' = P_{o(A \rightarrow B)} P_{o(B \rightarrow A)} \times \left\{ \frac{1 - \exp(qrF \cos \theta/kT)}{1 - P_{o(B \rightarrow A)} [\exp(qrF \cos \theta/kT) - 1]} \right\} \quad (2-153)$$

If $qrF/kT \ll 1$, p' can be approximated to

$$p' \approx P_{o(A \rightarrow B)} P_{o(B \rightarrow A)} (qrF \cos \theta/kT) \quad (2-154)$$

the random hopping dipole moment is

$$u_h = qr \quad (2-155)$$

The hopping dipole moment may be considered similar phenomenally to the orientational dipole moment, but in nature, they are different. The orientational dipole moment refers to the permanent dipole moment formed by bound charges within the particle, while the hopping dipole moment is the moment formed by the transition of a separate charged particle from one potential well to another potential well.

Following the same concept in treating the orientation of permanent dipoles (Equation 2-123), the hopping dipole moment in the direction of the applied field can be written as

$$\langle u_{hF} \rangle = u_h \langle p' \cos \theta \rangle \quad (2-156)$$

where the factor p' means that it is the field-induced excess probability p' that produces the hopping polarization. Using Equation 2-127, we have

$$\begin{aligned} \langle p' \cos \theta \rangle &= \frac{\int p' \cos \theta dN}{\int dN} \\ &= \frac{\int_0^\pi P_{o(A \rightarrow B)} P_{o(B \rightarrow A)} (qrF \cos \theta/kT) \times (\cos \theta) \frac{1}{2} N \sin \theta d\theta}{\int_0^\pi \frac{1}{2} N \sin \theta d\theta} \\ &= \frac{P_{o(A \rightarrow B)} P_{o(B \rightarrow A)} (qrF)}{3kT} \end{aligned} \quad (2-157)$$

Substitution of Equation 2-157 into Equation 2-156 gives

$$\langle u_{hF} \rangle = \frac{q^2 r^2 F}{3kT} P_{o(A \rightarrow B)} P_{o(B \rightarrow A)} \quad (2-158)$$

Thus, the hopping polarizability is

$$\alpha_h = \frac{q^2 r^2}{3kT} P_{o(A \rightarrow B)} P_{o(B \rightarrow A)} \quad (2-159)$$

where $P_{o(A \rightarrow B)} P_{o(B \rightarrow A)}$ denotes the ensemble average of the product of these two probabilities, because E_A and E_B may vary from site to site.

Interfacial Polarization

The space charge, or interfacial polarization, is produced by the separation of mobile positively and negatively charged particles under an applied field, which form positive and negative space charges in the bulk of the material or at the interfaces between different materials. These space charges, in turn, modify the field distribution.

Suppose that we have a composite dielectric specimen comprising two parallel sheets of different materials, and that this dielectric specimen is inserted into the space between two parallel metallic plates of unit area, as shown in Figure 2-22. In this case, under AC fields the admittance is

$$Y = \frac{Y_1 Y_2}{Y_1 + Y_2} \quad (2-160)$$

in which

$$Y_1 = (\sigma_1 + j\omega\epsilon_{1r}\epsilon_0)/d_1 \quad (2-161)$$

$$Y_2 = (\sigma_2 + j\omega\epsilon_{2r}\epsilon_0)/d_2 \quad (2-162)$$

where ϵ_{1r} and ϵ_{2r} are the dielectric constants, σ_1 and σ_2 are the conductivities, and d_1 and d_2 are the thicknesses of sheets 1 and 2, respectively. Substitution of Equations 2-161 and 2-162 into Equation 2-160 gives

$$Y = \frac{\sigma_1 \sigma_2}{\sigma_1 d_2 + \sigma_2 d_1} \left[\frac{\begin{pmatrix} 1 - \omega^2 \tau_1 \tau_2 \\ + \omega^2 \tau_1 \tau \\ + \omega^2 \tau_2 \tau \end{pmatrix} + j \begin{pmatrix} \omega \tau_1 + \omega \tau_2 \\ + \omega^3 \tau_1 \tau_2 \tau \\ - \omega \tau \end{pmatrix}}{1 + \omega^2 \tau^2} \right] \quad (2-163)$$

where

$$\tau_1 = \epsilon_{1r} \epsilon_0 / \sigma_1 \quad (2-164)$$

$$\tau_2 = \epsilon_{2r} \epsilon_0 / \sigma_2 \quad (2-165)$$

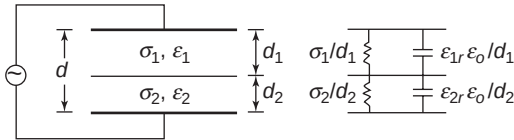


Figure 2-22 A composite dielectric material between two parallel metal plates of unit area.

$$\tau = (\epsilon_{1r} d_2 + \epsilon_{2r} d_1) \epsilon_0 / (\sigma_1 d_2 + \sigma_2 d_1) \quad (2-166)$$

Y can also be written as

$$Y = \sigma/d + j\omega\epsilon_r\epsilon_0/d \quad (2-167)$$

By comparing Equation 2-167 with Equation 2-163, we obtain the overall dielectric constant and conductivity of this composite dielectric as follows:

$$\epsilon_r = \frac{1}{\epsilon_0} \left(\frac{\sigma_1 \sigma_2 d}{\sigma_1 d_2 + \sigma_2 d_1} \right) \frac{\tau_1 + \tau_2 + \omega^2 \tau_1 \tau_2 \tau - \tau}{1 + \omega^2 \tau^2} \quad (2-168)$$

$$\sigma = \left(\frac{\sigma_1 \sigma_2 d}{\sigma_1 d_2 + \sigma_2 d_1} \right) \frac{1 - \omega^2 \tau_1 \tau_2 + \omega^2 \tau_1 \tau + \omega^2 \tau_2 \tau}{1 + \omega^2 \tau^2} \quad (2-169)$$

Under static DC fields, $\omega = 0$. Thus, the static overall dielectric constant and conductivity are

$$\epsilon_{rs} = \left(\frac{\sigma_1 \sigma_2 d}{\sigma_1 d_2 + \sigma_2 d_1} \right) \left(\frac{\tau_1 + \tau_2 - \tau}{\epsilon_0} \right) \quad (2-170)$$

and

$$\sigma_s = \frac{\sigma_1 \sigma_2 d}{\sigma_1 d_2 + \sigma_2 d_1} \quad (2-171)$$

At high frequencies, the space charges cannot follow the change of the field and hence do not produce space charge polarization. By setting $\omega \rightarrow \infty$, we obtain

$$\epsilon_{r\infty} = \frac{\epsilon_{1r} \epsilon_{2r} d}{\epsilon_{1r} d_2 + \epsilon_{2r} d_1} \quad (2-172)$$

$$\sigma_\infty = \sigma_s (\tau_1 \tau_2 + \tau_1 \tau + \tau_2 \tau) / \tau_2 \quad (2-173)$$

We can consider $\epsilon_{r\infty}$ to be mainly the contribution of electronic and atomic polarizations. Thus, the total polarization is

$$P = (\epsilon_{rs} - 1) \epsilon_0 F = N \alpha F \quad (2-174)$$

and the polarization due to electronic, atomic, and orientational contributions is

$$P' = (\epsilon_{r\infty} + \epsilon_{ro} - 1) \epsilon_0 F = N(\alpha_\infty + \alpha_o) F \quad (2-175)$$

Therefore, $P - P'$ is the polarization due to the space charge contribution. Thus,

$$\begin{aligned} P_c = P - P' &= (\epsilon_{rs} - \epsilon_{r\infty} - \epsilon_{ro}) \epsilon_0 F \\ &= N(\alpha - \alpha_\infty - \alpha_o) F \end{aligned} \quad (2-176)$$

and the space charge polarizability is

$$\alpha_c = \alpha - \alpha_\infty - \alpha_o = (\epsilon_{rs} - \epsilon_{r\infty} - \epsilon_{ro})\epsilon_o / N \quad (2-177)$$

where N is the number of molecules per unit volume of the composite dielectric. It should be noted that if this composite dielectric also involves hopping polarization, then the space charge polarizability $\alpha_d = \alpha_h + \alpha_c$ is the combination of both the hopping and the interfacial polarizabilities. Since these two types of polarization involve the movement of charged particles, there is no easy experimental method to separate these two different mechanisms.

2.4 Classification of Dielectric Materials

Dielectric materials may be classified into two major categories: nonferroelectric (or normal dielectric or paraelectric) materials and ferroelectric materials. The former may be divided into three classes, as the following sections show.

2.4.1 Nonferroelectric Materials (Normal Dielectric or Paraelectric Materials)

In materials of this category, electric polarization is actuated by external electric fields. Based on the mechanisms of electric polarization, we have three classes: nonpolar, polar, and dipolar materials.

Nonpolar Materials

In materials of this class, an electric field can cause only elastic displacement of electron clouds (mainly valence electron clouds), so they have only electronic polarization. Such materials, which are generally referred to as elemental materials, consist of a single kind of atom, such as silicon (Si), diamonds (C), inert elements in gas, liquid, or solid phase. For these materials, the appreciable absorption occurs at the resonance frequency ω_o , which is in the visible-to-ultraviolet region. For frequencies lower than the resonance frequency, the

dielectric constant should be independent of frequency and equal to the static dielectric constant. According to Maxwell's relation, the refractive index of such materials can be written as

$$n = (\epsilon_{rs})^{1/2} \quad (2-178)$$

The total polarizability is

$$\alpha = \alpha_e \quad (2-179)$$

Polar Materials

In materials of this class, an electric field will cause elastic displacement of the valence electron clouds, as well as elastic displacement of the relative positions of ions, so such materials have both electronic and ionic polarization. The material may be composed of molecules, and each of the molecules is made of more than one kind of atom without permanent dipole moments. Examples of such materials are ionic crystals, including alkali halides, some oxides, paraffins, benzene, carbon tetrachloride, etc. The appreciable absorption occurs at two resonant frequencies: one in the optical frequency region (corresponding to electronic polarization) and the other in the lower resonance frequency—the infrared region corresponding to ionic polarization). In this case, the total polarizability is

$$\alpha = \alpha_e + \alpha_i \quad (2-180)$$

Dipolar Materials

The materials of this class have all three fundamental polarizations: electronic, ionic, and orientational. Thus, the total polarizability is

$$\alpha = \alpha_e + \alpha_i + \alpha_o \quad (2-181)$$

Materials whose molecules possess a permanent dipole moment belong to this class. An electric field will cause spatial orientation of the permanent dipoles, resulting in orientational polarization. The orientation process occurs predominantly in liquids and gases, and in some solids within a certain range of temperatures, such as solid hydrochloric and sulfuric acids. However, in the solid state, there exists

a critical temperature at and below which all dipoles are frozen in and lose their capability to contribute orientational polarizability. But for most materials, the dipoles are frozen in below the melting point of the material.

It should be noted that dielectric materials, in general, are not single crystals; they are either amorphous or polycrystalline, containing a large quantity of various traps. Furthermore, they are not nonconductive and always involve charge carriers (electrons, holes, or both) injected from electrical contacts. Therefore, in this case, the total polarizability should include the space charge polarizability

$$\alpha = \alpha_e + \alpha_i + \alpha_o + \alpha_d \quad (2-182)$$

where α_d in the space charge polarizability includes $\alpha_n + \alpha_c$.

2.4.2 Ferroelectric Materials

A ferroelectric material is normally in single crystalline or polycrystalline form and possesses a reversible spontaneous polarization over a certain temperature range. There is a critical temperature, called the Curie temperature, which marks the transition from the ordered to the disordered phase. At this temperature, the dielectric constant may reach values three to four orders of magnitude higher than in the disordered phase. The order–disorder phase transition involves the displacement of atoms so that crystals or crystallites exhibiting ferroelectric phenomena must be noncentrosymmetric. This implies that a phase transition will induce a mechanical strain, tending to change not only the volume and the shape of the material body, but also the optical refractive index. Thus, ferroelectric materials exhibit not only ferroelectric phenomena, but also piezoelectric, pyroelectric, and electro-optic effects, which can be used for many technological applications. In general, ferroelectric materials also have electrically induced polarizations, but these are negligibly small compared to spontaneous polarization. For most practical cases dealing with ferroelectric phenomena, electrically induced polarization can be ignored. More details about the properties and applications of

ferroelectric materials are given in Ferroelectric Phenomena in Chapter 4.

2.5 Internal Fields

Only in gases or in dilute phases, in which the interaction between atoms or molecules can be neglected, the internal or local field, F_{loc} , acting on the atomic or molecular sites, can be assumed to be the same as the applied field F . However, the local field in condensed phases (solids and liquids) is higher than F because of the polarization of the surrounding particles. In this section, we shall briefly discuss local fields for nondipolar and dipolar materials.

Local Fields for Nondipolar Materials

To understand the concept of local fields, imagine a small sphere of dielectric material removed from around the site of the atom or molecule to form a cavity, as shown in Figure 2-23. The local field F_{loc} is composed of four components

$$F_{loc} = F_0 + F_1 + F_2 + F_3 \quad (2-183)$$

where F_0 is the externally applied field in the cavity, which is assumed to be a vacuum, acting on A and is related to the applied field in the bulk F by

$$D = \epsilon_0 F + P = \epsilon_0 F_0 \quad (2-184)$$

or

$$F_0 = F + P/\epsilon_0$$

F_1 is the depolarization field resulting from polarization charges (bound charges) on the surface of the specimen, which is given by

$$F_1 = -P/\epsilon_0 \quad (2-185)$$

In fact, $F_0 + F_1$ is equal to the applied field F . F_2 is the Lorentz field due to polarization charges on the inside surface of the spherical cavity,^{6,27} which is given by

$$F_2 = \frac{(\epsilon_r - 1)}{3} F \quad (2-186)$$

F_3 is the field of the adjacent dipoles due to the molecules inside the spherical cavity. In

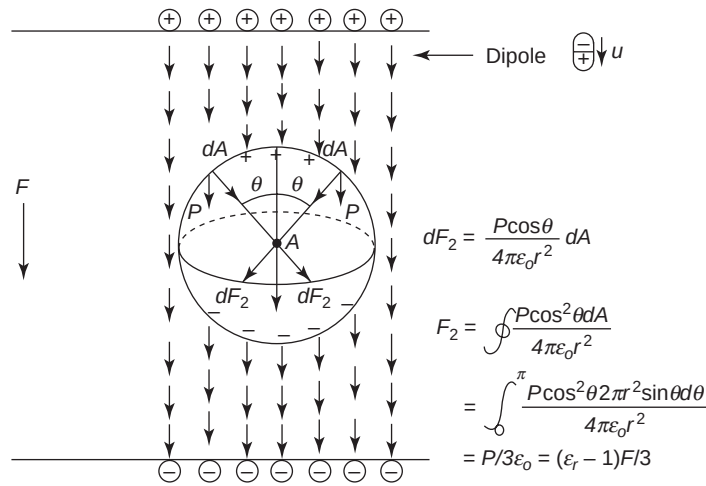


Figure 2-23 The model for the calculation of the Lorentz field.

general, we assume that the cavity is ideally formed without disturbing the state of polarization of the surrounding material and that the fields contributed by all dipoles inside the cavity tend to compensate each other. This assumption leads to

$$F_3 = 0 \quad (2-187)$$

Thus, the local field can be written as

$$F_{loc} = F_0 + F_1 + F_2 + F_3 = \frac{\epsilon_r + 2}{3} F \quad (2-188)$$

The Clausius-Mossotti Equation

From Equations 2-63, 2-64, and 2-188, we obtain

$$\frac{\epsilon_r - 1}{\epsilon_r + 2} = \frac{N\alpha}{3\epsilon_0} \quad (2-189)$$

This is the well known Clausius-Mossotti equation. The number of atoms or molecules per unit volume N is related to the number of atoms or molecules per mole (Avogadro's number N_0) by

$$N_0 = \frac{NM}{\rho} = 6.023 \times 10^{23} \text{ per mole} \quad (2-190)$$

where M and ρ are, respectively, the atomic or molecular weight and the density of the mate-

rial. Substitution of Equation 2-190 into Equation 2-189 gives the molar polarization in terms of polarization per mole

$$\frac{\epsilon_r - 1}{\epsilon_r + 2} \frac{M}{\rho} = \frac{N_0\alpha}{3\epsilon_0} \quad (2-191)$$

This is the Clausius-Mossotti equation expressed in terms of M and ρ . Using either Equation 2-189 or Equation 2-191, α can be determined because all other parameters are generally known or can easily be measured. It should be noted that the Clausius-Mossotti equation can be used only for nonpolar materials because F_3 cannot be assumed to be zero for polar materials, which will be discussed in later sections. It should also be noted that when $\frac{N\alpha}{3\epsilon_0}$

approaches unity, ϵ_r approaches infinity, and when ρ is larger than a certain critical value, ϵ_r tends to be negative. This phenomenon is generally called the Clausius-Mossotti catastrophe, indicating that Equations 2-189 and 2-191 are reasonable only for a low α , a low ρ , and nondipolar materials.

The Lorentz-Lorenz Equation

The Clausius-Mossotti equation can be used for dipolar materials in high-frequency AC

fields in which orientational polarization cannot follow the time-varying fields but electronic and atomic polarizations are still present. For optical frequencies, only electronic polarization is dominant, and in this case $\epsilon_r = n^2$, where n is the refractive index. In the optical frequency range, Equations 2-189 and 2-191 can be rewritten as

$$\frac{n^2 - 1}{n^2 + 2} = \frac{N\alpha}{3\epsilon_0} \quad (2-192)$$

$$\frac{n^2 - 1}{n^2 + 2} \frac{M}{\rho} = \frac{N_0\alpha}{3\epsilon_0} \quad (2-193)$$

These equations are sometimes called the Lorentz–Lorenz equation because they were first derived independently by H.A. Lorentz in Holland and L. Lorenz in Denmark in 1880.²⁷

At high frequencies, which are so high that no orientational polarization can occur but still low enough to include both the electronic and atomic polarizations, ϵ_r is termed $\epsilon_{r\infty}$. In this case, $\epsilon_{r\infty} - n^2$ is the contribution of atomic polarization. We can also estimate the orientational contribution. The electronic and atomic polarizability can be estimated from

$$\frac{\epsilon_{r\infty} - 1}{\epsilon_{r\infty} + 2} = \frac{N(\alpha_e + \alpha_i)}{3\epsilon_0} \quad (2-194)$$

The total polarizability can be obtained by measuring ϵ_r under static DC fields. Thus, subtracting Equation 2-194 from Equation 2-189 gives the polarizability contributed by orientational polarization if both the hopping and the space charge polarizations are absent. This gives

$$\frac{(\epsilon_{rs} - \epsilon_{r\infty})}{(\epsilon_{rs} + 2)(\epsilon_{r\infty} + 2)} = \frac{N\alpha_o}{9\epsilon_0} \quad (2-195)$$

where ϵ_{rs} denotes the static dielectric constant.

2.5.2 The Reaction Field for Dipolar Materials

In Section 2.5.1, we derived an expression for the local field based on the assumption that there is no contribution to the field from the molecules inside the spherical cavity. If a dipolar molecule is located at the center of the

cavity of radius a , it produces a field that polarizes the cavity's surrounding molecules. The reaction field F_r is the field created inside the cavity by the charges of these polarized molecules on the cavity surface. Thus, in the absence of the applied field $F = 0$, the effective dipole moment of a dipolar molecule in a condensed dipolar material is

$$u_{eff} = u_0 + \alpha F_r \quad (2-196)$$

where u_0 is the permanent dipole moment of an individual molecule and α is the total polarizability of the material. The permanent dipole moment and F_r are in the same direction. This reaction field is given by

$$F_r = \frac{2u_{eff}}{4\pi\epsilon_0 a^3} \left(\frac{\epsilon_r - 1}{2\epsilon_r + 1} \right) = f u_{eff} \quad (2-197)$$

where

$$f = \frac{1}{2\pi\epsilon_0 a^3} \left(\frac{\epsilon_r - 1}{2\epsilon_r + 1} \right) = \frac{2N}{3\epsilon_0} \left(\frac{\epsilon_r - 1}{2\epsilon_r + 1} \right) \quad (2-198)$$

since $N = (4\pi a^3/3)^{-1}$.⁶ Then, from Equations 2-196 and 2-197, we obtain

$$u_{eff} = \frac{u_0}{1 - \alpha f} \quad (2-199)$$

With an applied field F , the field inside the empty spherical cavity is given by

$$F_i = \frac{3\epsilon_r}{2\epsilon_r + 1} F = hF \quad (2-200)$$

where

$$h = \frac{3\epsilon_r}{2\epsilon_r + 1} \quad (2-201)$$

Thus, the effective induced dipole moment of a dipolar molecule inside the cavity due to the cavity field F_i and the reaction field F_r in the direction of the applied field F , as shown in Figure 2-24, becomes

$$u'_{eff} = \alpha(\vec{F}_i + f u'_{eff})$$

or

$$u'_{eff} = \frac{\alpha h}{1 - \alpha f} F \quad (2-202)$$

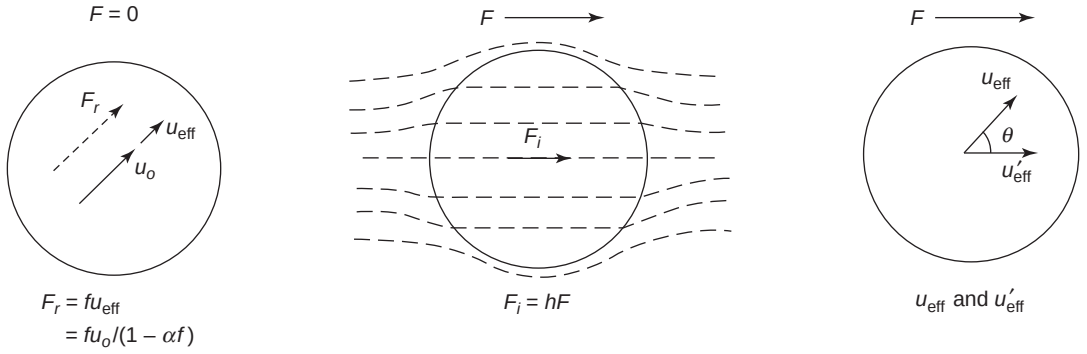


Figure 2-24 Schematic diagrams illustrating the concept of the reaction field F_r , the cavity internal field F_i , the effective dipole moment of a polar molecule u_{eff} , and the effective induced dipole moment of a polar molecule u'_{eff} in a condensed polar material.

The total dipole moment in the cavity is the sum of the permanent dipole moment and the induced dipole moment in the direction of F

$$u_F = u_{\text{eff}} \cos \theta + u'_{\text{eff}} \quad (2-203)$$

Polarized by the reaction field at $F = 0$ Polarized by the cavity field at applied field F

To obtain the mean value of u_F , we use the Boltzmann statistics with a weight factor $\exp\left[-\frac{W(\theta)}{kT}\right]$, in which $W(\theta)$ is the activation energy for the rotation of a dipole toward the field direction and is given by

$$W(\theta) = \vec{u}_{\text{eff}} \cdot \vec{F} = -\frac{hu_{\text{eff}}F \cos \theta}{kT} \quad (2-204)$$

Since only u_{eff} must rotate but not u'_{eff} , which is already in parallel to F , we have

$$\langle u_F \rangle = \frac{\int (u_{\text{eff}} \cos \theta + u'_{\text{eff}}) \exp(hu_{\text{eff}}F \cos \theta / kT) d\Omega}{\int \exp(hu_{\text{eff}}F \cos \theta / kT) d\Omega} \quad (2-205)$$

where $d\Omega = 2\pi \sin \theta d\theta$ is the solid angle corresponding to $d\theta$. Following the same approach for $L(z)$ (Equation 2-127) and substituting Equations 2-199 and 2-202 into Equation 2-205, we obtain

$$\begin{aligned} \langle u_F \rangle &= \frac{\int \left[\left(\frac{u_0}{1 - \alpha_f} \right) \cos \theta + \frac{\alpha h}{1 - \alpha_f} F \right] \exp \left[\left(\frac{hu_0}{1 - \alpha_f} \right) F \cos \theta / kT \right] 2\pi \sin \theta d\theta}{\int \exp \left[\left(\frac{hu_0}{1 - \alpha_f} \right) F \cos \theta / kT \right] 2\pi \sin \theta d\theta} \\ &= \frac{h}{1 - \alpha_f} \left[\alpha + \frac{u_0^2}{3(1 - \alpha_f)kT} \right] F \end{aligned} \quad (2-206)$$

The Debye Equation

Assuming that $\alpha_h = \alpha_c = 0$ (or $\alpha_d = 0$) and expressing the total polarizability as

$$\alpha = \alpha_e + \alpha_i + \frac{u_0^2}{3kT} \quad (2-207)$$

substitution of Equation 2-207 into Equation 2-189 gives

$$\frac{\epsilon_r - 1}{\epsilon_r + 2} = \frac{N}{3\epsilon_0} \left(\alpha_e + \alpha_i + \frac{u_0^2}{3kT} \right) \quad (2-208)$$

This is called the Debye equation. The molar polarization can be expressed in a form similar to Equation 2-191

$$\pi = \left(\frac{\epsilon_r - 1}{\epsilon_r + 2} \right) \frac{M}{\rho} = \frac{N_0}{3\epsilon_0} \left(\alpha_e + \alpha_i + \frac{u_0^2}{3kT} \right) \quad (2-209)$$

In fact, the Debye equation is similar to the Clausius–Mossotti equation; the only differ-

ence is that the former introduces $\alpha_o = u_o^2/3kT$ into the equation. Both equations were derived on the basis of the internal local field rather than the reaction field. The Debye equation is in poor agreement with experiments for dipolar liquids and solids, possibly due in part to the fact that for dipolar liquids and solids, the Lorentz internal field is not applicable. The Onsager equation improves on the Debye equation by using the reaction field instead of the Lorentz internal field.

Theoretically, the molar polarization deduced from the Debye equation should be a constant at a fixed temperature and independent of pressure. This means that Equation 2-191 should have the same value for the material whether in the gaseous phase or in the liquid phase. Experimentally, this is not true for dipolar liquids. However, for gases in which $\epsilon_r - 1 \ll 1$, the Debye equation gives a reasonably accurate prediction. Based on the Debye equation, $(\alpha_e + \alpha_i)$ and u_o can be determined by measuring ϵ_r at different temperatures. Furthermore, the Debye equation gives reasonable results for diluted solutions of dipolar molecules in nondipolar solvents.

In general, dipolar materials can be classified into three groups based on the permanent dipole moments of their molecules:

- Weakly dipolar materials in which $u_o < 0.5$ debye
- Medium dipolar materials in which 0.5 debye $< u_o < 1.5$ debye
- Strongly dipolar materials in which $u_o > 1.5$ debye

The Onsager Equation

In the static DC or low-frequency fields, we have

$$\begin{aligned} h \rightarrow h_s &= \frac{3\epsilon_{rs}}{2\epsilon_{rs} + 1} \\ f \rightarrow f_s &= \frac{2N}{3\epsilon_0} \left(\frac{\epsilon_{rs} - 1}{2\epsilon_{rs} + 1} \right) \end{aligned} \quad (2-210)$$

On the basis of the definition of the polarization, we obtain

$$P = (\epsilon_{rs} - 1)\epsilon_0 F = N\langle u_F \rangle \quad (2-211)$$

At zero and low frequencies, all types of polarization are taking place, so $\alpha = \alpha_e + \alpha_o + \alpha_i$ (hopping and space charge polarizations are assumed to be absent). However, in the high-frequency fields, in which orientational polarization does not occur, $u_o = 0$. Then we have

$$\begin{aligned} h \rightarrow h_\infty &= \frac{3\epsilon_{r\infty}}{2\epsilon_{r\infty} + 1} \\ f \rightarrow f_\infty &= \frac{2N}{3\epsilon_0} \left(\frac{\epsilon_{r\infty} - 1}{2\epsilon_{r\infty} + 1} \right) \end{aligned} \quad (2-212)$$

The polarization becomes

$$P_\infty = (\epsilon_{r\infty} - 1)\epsilon_0 F = N\langle u_{F\infty} \rangle \quad (2-213)$$

Hence, we obtain

$$\frac{\epsilon_{r\infty} - 1}{\epsilon_{r\infty} + 2} = \frac{N(\alpha_e + \alpha_i)}{3\epsilon_0} \quad (2-214)$$

This equation is, in fact, the same as the Clausius-Mossotti equation. By subtracting Equation 2-213 from Equation 2-211, we obtain

$$\frac{(\epsilon_{rs} - \epsilon_{r\infty})(2\epsilon_{rs} + \epsilon_{r\infty})}{\epsilon_{rs}(\epsilon_{r\infty} + 2)^2} = \frac{Nu_o^2}{9\epsilon_0 kT} \quad (2-215)$$

This is the well known Onsager equation.³⁸ This equation enables the computation of u_o of a dipolar molecule from ϵ_{rs} of a pure dipolar liquid or solution, if its density and $\epsilon_{r\infty}$ are known. The Onsager equation was first derived on the assumption that the molecules are spherical in shape and that there are no interactions between the molecules (or between the dipoles). However, u_o of nonassociating organic compounds (with weak interactions between molecules) agrees reasonably with those computed from Onsager equation. For associating organic compounds (with strong interactions between molecules), the difference between the measured and the computed values of u_o is large, indicating that Onsager's model is inadequate for associating organic compounds.³⁹ This is mainly due to the following reasons:

- The model assumes that the surrounding medium of any molecule is a continuum. This means that it does not take into account the interactions between the molecule and its neighbors. In other words, only long-

range interaction is considered; short-range interaction energy has been neglected.

- The model does not take into consideration the electrostatic interaction between nearest dipoles.
- The model does not take into account the nonpolar molecular interactions, which may be significant in some liquids and solids. Several investigators^{6,24,28} have attempted to modify the Onsager equation by using an ellipsoid instead of a spheroid in the shape of the cavity, and also by taking into account the anisotropy of the polarizabilities of the molecules. However, their modified models do not improve much. The improvement must take into account the intermolecular interactions, and to this end, we must use the statistical-mechanical approach.

Statistical-Mechanical Approaches and the Kirkwood Equation

Both Debye and Onsager used a statistical method to solve only part of the polarization problem by considering the whole dielectric medium as a continuum where there is no correlation between molecules (molecular interactions) in deformational (or induced) polarization and in orientational polarization. Statistical-mechanical approaches to the polarization problem, taking into account the interaction between nearest dipoles, were first used by Kirkwood.⁴⁰ He developed a general method of solving the problem that had some influence on all the subsequent polarization theories. Kirkwood had taken into account the short-range correlation in orientational polarization, but he ignored the correlation in deformational polarization between molecules. In Kirkwood's final equation, the terms due to deformational polarization were added empirically. Later, Fröhlich²⁸ developed a more rigorous expression, taking into account the correlation between molecules in both orientational and deformational polarizations by means of a statistical-mechanical approach. The equations derived by these and other investigators include a correlation factor g , which is related to the effects of molecular interactions. When $g = 1$, no correlation between molecules is taken

into account. This is true only for systems of nonassociating molecules. Therefore, the Clausius-Mossotti, Debye, and Onsager equations can apply only to systems of nonassociating molecules and, as a result, they are not satisfactory for condensed systems, such as polymeric materials.

Kirkwood⁴⁰ developed his theory of static polarization by taking into account the short-range interaction between molecules in the liquid state. The Kirkwood equation is

$$\frac{(\epsilon_{rs} - 1)(2\epsilon_{rs} + 1)}{9\epsilon_{rs}} = \frac{N}{9\epsilon_0} \left(\alpha_e + \alpha_i + \frac{gu_o^2}{3kT} \right) \quad (2-216)$$

where g is the correlation factor. The correlation factor is a measure of the local ordering in the material. $g = 1$ indicates that the average dipole moment of a finite spherical region around one reference molecule, which is held fixed in an infinite size of the material specimen, is about equal to the moment of the fixed molecule. This implies that the location of one dipole does not influence the positions of the other dipoles. If a fixed dipole tends to make the neighboring dipoles line up in a parallel direction, then the effective moment is larger than u_o and hence $g > 1$. Conversely, if the fixing dipole tends to line up with the neighboring dipoles in an antiparallel direction, then $g < 1$.³⁴ If the reference molecule (the fixing dipole) is surrounded by z neighboring molecules and $\cos\Phi$ is the average of $\cos\Phi$ made between the reference molecule and one of its z nearest neighbors, then g may be expressed as^{39,40}

$$g = 1 + z \overline{\cos\phi} \quad (2-217)$$

The correlation factor is supposed to characterize the molecular interaction and the short-range structure. Up to the present, no good method is available for determining the value of g accurately. In general, it is roughly estimated from the measured value of the dielectric constant or the dipole moment. This is one of the major shortcomings of the Kirkwood equation. Furthermore, Kirkwood used the approximate formula for the internal field, which also makes Kirkwood's equation inaccurate.

The Frohlich Equation

In the model of Frohlich,²⁸ both the short-range interaction between molecules and the deformational polarization are taken into account. Frohlich also considered a spherical region within an infinite, homogeneous dielectric continuum of static dielectric constant ϵ_{rs} . This region contains dipoles whose behavior is governed by statistical-mechanical laws. Frohlich also treated the internal and reaction fields in a macroscopic manner, which is more realistic than the Onsager model. The general Frohlich equation is

$$\frac{(\epsilon_{rs} - \epsilon_{r\infty})(2\epsilon_{rs} + \epsilon_{r\infty})}{\epsilon_{rs}(\epsilon_{r\infty} + 2)^2} = \frac{Ngu^2}{9\epsilon_0 kT} \quad (2-218)$$

where

$$g = 1 + \sum_{\substack{j \\ j \neq 1}}^{N_v} \overline{\cos \phi_{ij}} \quad (2-219)$$

N_v is the number of the dipoles within the spherical region of volume V , and thus, $N = N_v/V$ is the number of dipoles per unit volume, and $\cos \phi_{ij}$ is the average of $\cos \Phi_{ij}$ made between the reference molecule i and the j th nearest neighbor.

Equation 2-218 becomes identical with the Onsager equation (Equation 2-215) when $g = 1$. However, the definition of u in the Frohlich equation differs from that in the Onsager equation.²⁸ If $\epsilon_{rs} \gg \epsilon_{r\infty}$, the Kirkwood equation and the Frohlich equation do not differ significantly.

In polymeric liquids or solids, the polymer chains are entangled. The appropriate way to deal with such molecules is to choose a small, repeating unit of a chain as a basic dipole unit; each of these basic units contributes equally to the average polarization of a macroscopic sphere of the dielectric in an applied field. Based on this consideration, the Frohlich equation can be used for polymers, provided that g is replaced by

$$g' = 1 + \sum_{\substack{j \\ j \neq 1}}^x \overline{\cos \phi'_{ij}} + \sum_k^x \overline{\cos \phi''_{ik}} \quad (2-220)$$

where $\overline{\cos \phi'_{ij}}$ is the average of $\cos \phi'_{ij}$ made between the reference unit i and the j th unit

within the same polymer chain, and $\overline{\cos \phi''_{ik}}$ is the average of $\cos \phi''_{ik}$ made between the reference unit i and a k th unit in the polymer chains that do not contain the reference unit i .³⁹ It is also assumed that the polymer chain contains x repeating dipole units and the polymer has a high molecular weight, so the contribution to the overall polarization made by the end groups can be neglected.

In contrast to Frohlich's approach, Harris and Adler⁴¹ considered a macroscopic sphere immersed not in an infinite macroscopic dielectric continuum but in a vacuum, in order to simplify the evaluation of the energy of the interaction between the specimen and the field. Their results predicted dielectric constants larger by a factor roughly equal to $\frac{(2\epsilon_{rs} + 1)(\epsilon_{r\infty} + 2)}{3(2\epsilon_{rs} + \epsilon_{r\infty})}$. However, because of many shortcomings to their approach, several investigators⁴²⁻⁴⁶ have proposed different approaches in order to correct them.

Buckingham⁴⁵ has pointed out that the Onsager relation between u_{eff} and u_o is a sufficiently good approximation for many polar liquids. By taking into account the discrete nature of the particles surrounding the reference molecule and the short-range interactions, Buckingham, in his derivation for the dielectric constants, arrived at the same expression as the Frohlich equation (Equation 2-218).

As mentioned earlier, the statistical theories of polarization may enable the determination of the effective dipole moment and the correlation factor, which characterize the molecular interactions. It has been shown that the value of the effective dipole moment in polymers depends on the internal rotation in polymer chains, which is greatly hindered due to the presence of strong molecular interactions.⁴⁷ A study of the dielectric polarization of polymers certainly will provide useful information about molecular interaction in polymers. As regards polymers, it is not possible to eliminate the interaction between polar groups belonging to the same chain; even polymers are dissolved in a nonpolar solvent with an infinite dilution. In alcohol, g is large ($g > 3$ for butyl alcohol at 0°C) because of the high degree of correlation between the alcohol molecules due to the

existence of hydrogen bonds, which leads to the parallel orientation of dipole moments. Replacing the OH radical with a halogen in butyl alcohol leads to a big change in g . For example, $g = 0.76$ for butyl bromide and $g = 0.85$ for butyl chloride at 0°C . The value of $g < 1$ indicates that the correlation in orientation makes the dipole moments antiparallel and cancel one another. The study of g as a function of temperature and pressure in polar substances, and as a function of the concentration in polar solution, is important to understanding molecular interactions with respect to structural changes.

2.6 Electric Polarization and Relaxation in Time-Varying Electric Fields

Atoms, molecules, and ions are so small that even a macroscopically tiny region in a solid contains very large numbers of such particles. The discrete nature of matter, and the behavior and interaction of those particles, can be manifested through their response to time-varying electric fields with wavelengths comparable to the distances between particles (interparticle distances). In condensed matter—solids and liquids—the interparticle distances are of the order of a few angstroms; electric fields with wavelengths of this order would be in the region of x-rays, which have energies capable of ionizing the particles. In this section, we will deal with the dynamic response of the dielectric material involving electric fields with wavelengths much larger than the interparticle distances, so the dielectric material can be treated as a continuum.

2.6.1 The Time-Domain Approach and the Frequency-Domain Approach

In the previous section, we discussed the static response of dielectric materials under static electric fields with $\omega = 0$. However, the dynamic response under time-varying electric fields provides much more information about the structure of the material and its dielectric behavior for fundamental studies, as well as for technological applications. To measure the

dynamic response, we can use either the time-domain approach or the frequency-domain approach. These two approaches are equally powerful methods for studying dielectric phenomena. From the viewpoint of measuring techniques, the time-domain approach is simpler than the frequency-domain approach, but from the viewpoint of data analyses, the time-domain approach is more complex. In the time-domain approach, we measure the time-dependent polarization immediately after the application of a step-function electric field, or we measure the decay of the polarization from an initial steady state value to zero after the sudden removal of an initial polarizing field. This decay is generally referred to as dielectric relaxation. In the frequency-domain approach, we mainly measure the dielectric constants at various frequencies of alternating excitation fields. Both approaches should be intimately connected and should yield, in principle, the same results.

2.6.2 Complex Permittivity

When a time-varying electric field is applied across a parallel-plate capacitor with the plate area of one unit and a separation of d between the plates, then the total current is given by

$$J_T = J + \frac{dD}{dt} = J + \epsilon^* \frac{dF}{dt} \quad (2-221)$$

where J is the conduction current and ϵ^* is defined as the complex permittivity, which is introduced to allow for dielectric losses due to the friction accompanying polarization and orientation of electric dipoles. This may be written as

$$\epsilon^* = \epsilon - j\epsilon' = (\epsilon_r - j\epsilon'_r)\epsilon_0 \quad (2-222)$$

in which ϵ_r is called the dielectric constant and ϵ'_r the loss factor. An arbitrary time-varying field can be resolved into sinusoidal AC fields of a spectrum of frequencies by means of the Fourier transformation. For simplicity, we consider the applied field to be monochromatic and a sinusoidal function of angular frequency ω , which can be expressed as

$$F = F_m \exp(j\omega t) \quad (2-223)$$

Substitution of Equations 2-222 and 2-223 into Equation 2-216 yields

$$J_T = \sigma F + j\omega(\epsilon - j\epsilon')F = (\sigma + \omega\epsilon')F + j\omega\epsilon F \quad (2-224)$$

where σ is the electric conductivity of the material. The first term on the right is a loss component due to the inelastic scattering of conducting charge carriers with scatterers during their migration, which is present at all frequencies, including $\omega = 0$ (DC fields); the second term is also a loss component due to the friction in the polarization processes, which disappears when $\omega = 0$ and increases with ω ; and the third term is a lossless component which is, in fact, the displacement current. For dielectric polymers, σ is normally extremely small, and in most cases the contribution of the first term can be neglected. By ignoring the first term, the $\tan \delta$ is given by

$$\tan \delta = \epsilon'_r / \epsilon_r \quad (2-225)$$

which is generally called the loss tangent, as shown in Figure 2-25. If $\epsilon'_r / \epsilon_r \ll 1$ then

$$\tan \delta = \epsilon'_r / \epsilon_r \approx \delta \quad (2-226)$$

This is generally called the loss angle.

The instantaneous energy absorbed per second per cm^3 by the material is given by $J_T(t)F(t)$. Thus, on average, the amount of energy per cm^3 per second absorbed by the material is

$$\begin{aligned} W &= \frac{1}{2\pi} \int_0^{2\pi} J_T(t)F(t)d(\omega t) \\ &= \omega\epsilon'_r\epsilon_0 F_m^2 / 2 \end{aligned} \quad (2-227)$$

2.6.3 Time-Dependent Electric Polarization

In general, the time required for electronic and atomic polarization and depolarization is very short ($<10^{-12}$ sec). This deformational polarization process is also referred to as the resonance process because it involves vibrating modes. Resonance of a vibrating system occurs when an excitation field oscillates at a frequency close to the natural frequency of the system. The time required for orientational, hopping, or space charge polarization and depolarization is quite long and varies in a wide range, depending on the dielectric systems; such polarization processes are sometimes referred to as relaxation processes because they involve a relaxation time. A relaxation phenomenon occurs when restoring action tends to bring the excited system back to its original equilibrium state.

Ignoring, for simplicity, hopping and space charge polarization, the total polarization of an arbitrary dielectric system is

$$P = P_e + P_i + P_o \quad (2-228)$$

Since the response time for electronic and atomic polarization is so short and can be assumed to be practically constant for all frequencies from 0 to about 10^{12} Hz, we can group these two polarizations as $P_\infty = P_e + P_i = (\epsilon_{r\infty} - 1)\epsilon_0 F$. These two types of polarization can be considered to follow instantaneously the excitation field F without time lag; in other words, P_∞ and F can be considered to be in phase. Thus, we can write

$$P = P_\infty + P_o = (\epsilon_{r\infty} - 1)\epsilon_0 F + (\epsilon_{rs} - \epsilon_{r\infty})\epsilon_0 F \quad (2-229)$$

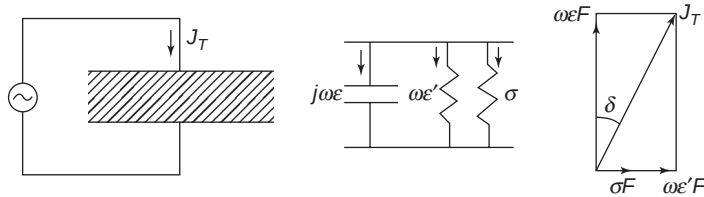


Figure 2-25 The response of a parallel plate capacitor with the plate area of one unit and its equivalent circuit under an alternating electric field.

It is P_o that has a time lag with F ; therefore, there is a phase difference between P_o and F .

Supposing that the orientational polarization takes time to respond to the applied excitation field F , and that at the removal of the excitation field making $F = 0$ at $t = 0$, the polarization will decay at a rate proportional to its change from its equilibrium state (as shown in Figure 2-26), then we can write

$$\frac{dP_o(t)}{dt} = -\frac{P_o(t)}{\tau_o} \quad (2-230)$$

where τ_o is the macroscopic relaxation time. Using the boundary condition that at $t = 0$, $P_o = (\epsilon_{rs} - \epsilon_{r\infty})\epsilon_o F$, the solution of Equation 2-230 is

$$P_o(t) = (\epsilon_{rs} - \epsilon_{r\infty})\epsilon_o F \exp(-t/\tau_o) \quad (2-226)$$

So, $dP_o(t)/dt$ gives the depolarization rate for this case, in which the excitation field is a step-function, $F = \text{constant}$ up to $t = 0$. Similarly, if a step-function excitation field F is applied to the dielectric system at $t = 0$, then $P_o(t) = 0$ at $t = 0$, and $P_o(t)$ increases with time. Following the same method, we obtain

$$P_o(t) = (\epsilon_{rs} - \epsilon_{r\infty})\epsilon_o F [1 - \exp(-t/\tau_o)] \quad (2-232)$$

In this case, $dP_o(t)/dt$ gives the polarization rate. The approximate time required for polarization is shown in Figure 2-27.

Supposing that during the time interval between u and $u + du$, an excitation field $F(u)$ is applied to the dielectric system, and that $F = 0$ for $t < u$ and $t > u + du$ (as shown in Figure 2-28), then $P_o(t)$ will take time to respond and will change for $t > u$. As soon as $P_o(t)$ reaches the value of $P_o(u + du)$ at $u = t + du$, the polarization will decay gradually. During the polarization period $u < t < u + du$, the change of the polarization can be written as

$$\begin{aligned} dP_o(t-u) \\ = (\epsilon_{rs} - \epsilon_{r\infty})\epsilon_o \left[1 - \exp\left(-\frac{t-u}{\tau_o}\right) \right] dF(u) \end{aligned} \quad (2-233)$$

in which $1 - \exp[-(t - u)/\tau_o]$ is a response function. The total P consists of two parts: P_∞ , which can follow the excitation field immediately, and $P_o(t)$, which is governed by Equation 2-233. Thus, we can write

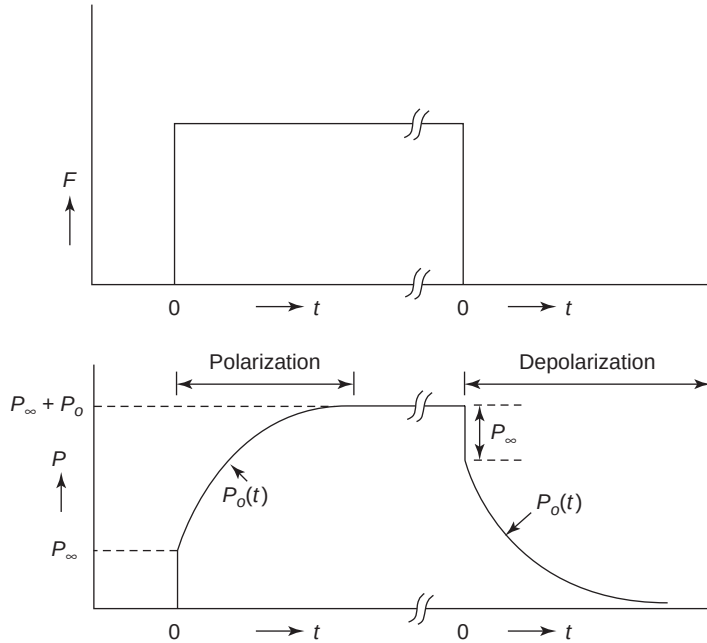


Figure 2-26 Time dependence of polarization and depolarization under a step-function electric field F .

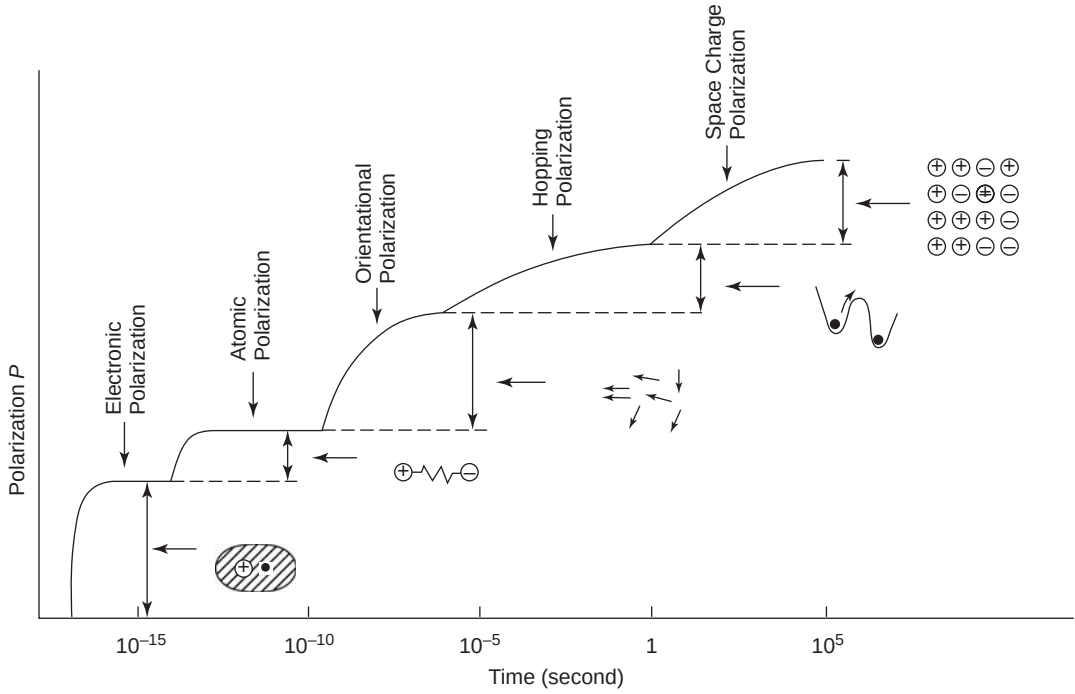


Figure 2-27 The variation of different types of polarization with time under a step-function electric field F .

$$\begin{aligned}
 dP(t-u) &= dP_{\infty}(t-u) + dP_o(t-u) \\
 &= (\epsilon_{r\infty} - 1)\epsilon_o dF(u) + (\epsilon_{rs} - \epsilon_{r\infty}) \\
 &\quad \times \epsilon_o \left[1 - \exp\left(-\frac{t-u}{\tau_o}\right) \right] dF(u)
 \end{aligned}
 \quad (2-234)$$

According to the superposition principle, the total polarization at time t is a superposition of all increments dP , so we have

$$\begin{aligned}
 P &= (\epsilon_{r\infty} - 1)\epsilon_o F(t) \\
 &\quad + (\epsilon_{rs} - \epsilon_{r\infty})\epsilon_o \int_0^t \left[1 - \exp\left(-\frac{t-u}{\tau_o}\right) \right] dF(u)
 \end{aligned}
 \quad (2-235)$$

Integrating by part, we obtain

$$\begin{aligned}
 P &= (\epsilon_{r\infty} - 1)\epsilon_o F(t) \\
 &\quad + (\epsilon_{rs} - \epsilon_{r\infty})\epsilon_o \int_0^t \frac{F(u)}{\tau_o} \exp[-(t-u)/\tau_o] du
 \end{aligned}
 \quad (2-236)$$

where $\exp[-(t-u)/\tau_o]$ is a decay function that tends to approach zero at $t \rightarrow \infty$.

In the following cases, we shall discuss two general cases.

Case A

If F is a step-function electric field with $F = 0$ at $t = 0^-$ and $F = F$ at $t = 0^+$, then from Equation 2-236 we have

$$\begin{aligned}
 P &= (\epsilon_{r\infty} - 1)\epsilon_o F \\
 &\quad + (\epsilon_{rs} - \epsilon_{r\infty})\epsilon_o F [1 - \exp(-t/\tau_o)]
 \end{aligned}
 \quad (2-237)$$

The variation of P with time t is shown in Figure 2-26.

Case B

If F is a sinusoidal AC field with

$$F = F_m \cos \omega t = \text{Re}[F_m \exp(j\omega t)] \quad (2-238)$$

where Re refers to the real part of $F_m \exp(j\omega t)$, P_{∞} can follow the field immediately, but P_o has a time lag, i.e., P_o has a phase shift. To analyze this case, we must assume that P has reached its dynamic steady state at $t = 0$, so that the bottom limit of the integral in Equation 2-236 must be changed to $-\infty$, where $P = 0$. From Equation 2-236 we obtain

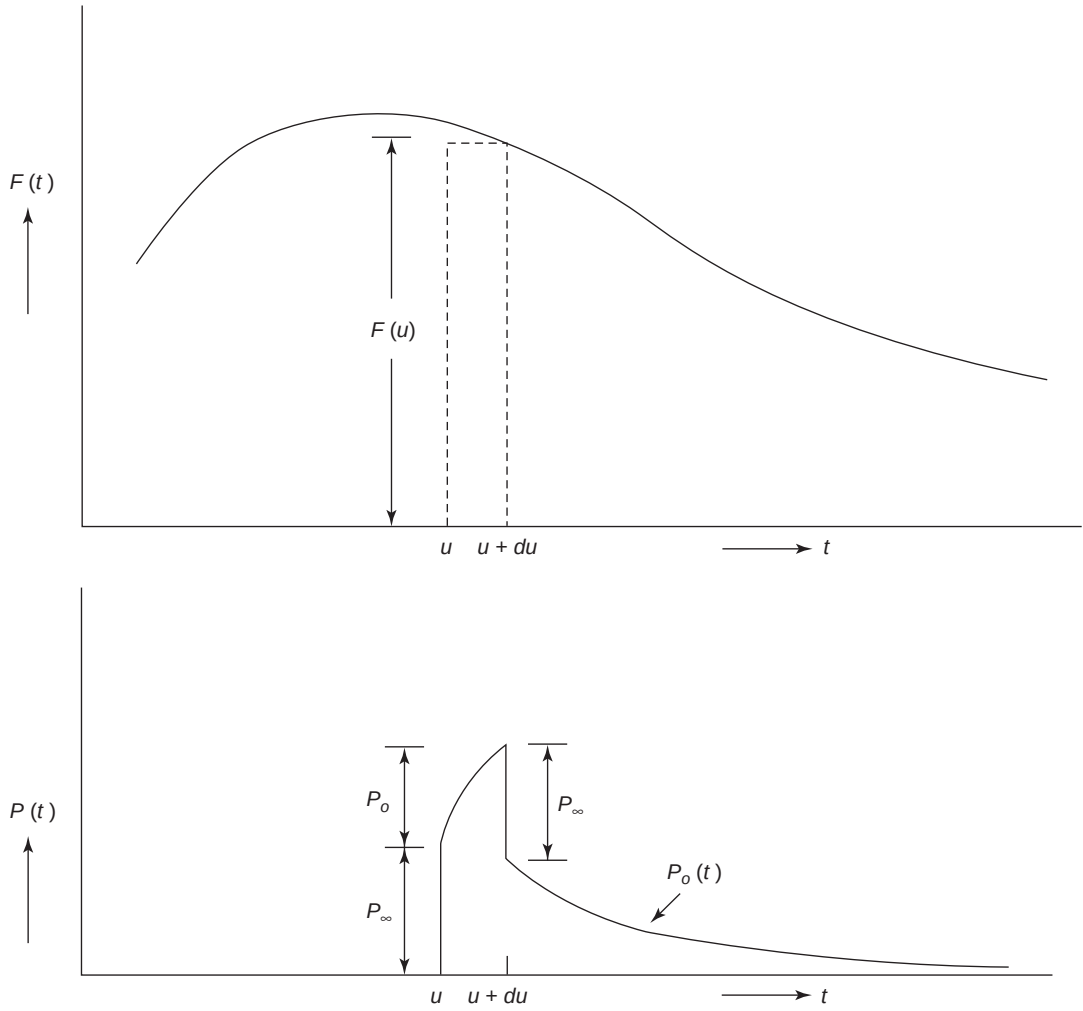


Figure 2-28 The time response of $P(t)$ to a delta function electric field $F(u)$ of strength $F(u)$ within the time period of $u \leq t \leq u + du$.

$$\begin{aligned}
 P &= (\epsilon_{r\infty} - 1)\epsilon_0 F_m \cos \omega t + (\epsilon_{rs} - \epsilon_{r\infty})\epsilon_0 \frac{F_m}{\tau_o} \operatorname{Re} \left[\int_{-\infty}^t \exp(j\omega t) \exp[-(t-u)/\tau_o] du \right] \\
 &= (\epsilon_{r\infty} - 1)\epsilon_0 F_m \cos \omega t + (\epsilon_{rs} - \epsilon_{r\infty})\epsilon_0 F_m \operatorname{Re} \left[\frac{\exp(j\omega t)}{1 - j\omega \tau_o} \right] \\
 &= (\epsilon_{r\infty} - 1)\epsilon_0 F_m \cos \omega t + \frac{(\epsilon_{rs} - \epsilon_{r\infty})\epsilon_0 F_m \cos \omega t}{1 + \omega^2 \tau_o^2} + \frac{(\epsilon_{rs} - \epsilon_{r\infty})\omega \tau_o \epsilon_0 F_m \sin \omega t}{1 + \omega^2 \tau_o^2}
 \end{aligned} \tag{2-239}$$

In phase with the applied field
Lag by $\pi/2$ from the applied field

The component in phase with the applied field is the lossless component, while the component with $\pi/2$ out of phase with the applied field is the loss component. It is obvious that for

$\omega \gg 1/\tau_o$ the dipoles cannot follow the field variation; hence, the polarization gradually decreases to zero as ω increases. The loss component, in fact, represents the dielectric losses

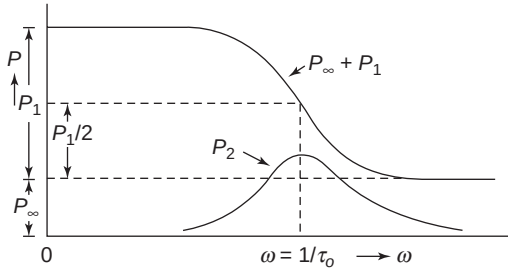


Figure 2-29 The variation of P_∞ , P_1 , and P_2 with frequency.

in the form of energy absorption. This term is maximum at $\omega_0 = 1/\tau_0$ and decreases gradually as ω increases or decreases from the point $\omega_0 = 1/\tau_0$. Using P_∞ , P_1 , and P_2 to denote the three components, Equation 2-234 can be written as

$$P = (P_\infty + P_1)\cos\omega t + P_2\sin\omega t \quad (2-240)$$

These components as functions of frequency are shown in Figure 2-29.

The variation of the dielectric parameters, such as P and ϵ^* , with frequency is known as the dispersion of the dielectric material, and it is one of the most important properties of all materials.

2.6.4 Kramers–Kronig Relations

Equation 2-231 can be used to derive the relations between the dielectric constant $\epsilon_r(\omega)$ and the loss factor $\epsilon_r'(\omega)$. These relations are known as the Kramers–Kronig relations.^{48,49} By writing the decay function in a more generalized form $\Gamma(t - u)$ to replace $\frac{1}{\tau_0}\exp[-(t-u)/\tau_0]$, Equation 2-231 can be written as

$$P = (\epsilon_{r\infty} - 1)\epsilon_0 F(t) + \int_0^t (\epsilon_{rs} - \epsilon_{r\infty})\epsilon_0 F(u)\Gamma(t-u)du \quad (2-241)$$

Since $D = D_0 + P = \epsilon_0 F(t) + P$, we have

$$D = \epsilon_{r\infty}\epsilon_0 F(t) + (\epsilon_{rs} - \epsilon_{r\infty})\epsilon_0 \int_0^t F(u)\Gamma(t-u)du \quad (2-242)$$

By introducing a new variable $x = t - u$ and assuming that the material has been subjected to the AC field $F(t) = F_m \cos \omega t$ for a sufficiently

long time that D has been settled to be a periodic function of time, we can write

$$\begin{aligned} D - \epsilon_{r\infty}\epsilon_0 F_m \cos \omega t \\ = (\epsilon_{rs} - \epsilon_{r\infty})\epsilon_0 F_m \left[\int_0^\infty \Gamma(x) \cos \omega t \cos \omega x dx \right. \\ \left. + \int_0^\infty \Gamma(x) \sin \omega t \sin \omega x dx \right] \end{aligned} \quad (2-243)$$

By expressing the AC field in the form of Equation 2-238, D can be considered to be the real part of the product ϵ^*F . Thus,

$$\begin{aligned} D &= \text{Re}[\epsilon_r^* \epsilon_0 F] = \text{Re}[(\epsilon_r - j\epsilon_r')\epsilon_0 F_m \exp(j\omega t)] \\ &= \epsilon_0 F_m [\epsilon_r \cos \omega t + \epsilon_r' \sin \omega t] \end{aligned} \quad (2-244)$$

Comparing Equation 2-243 with Equation 2-244, it follows that

$$(\epsilon_r - \epsilon_{r\infty}) = (\epsilon_{rs} - \epsilon_{r\infty}) \int_0^\infty \Gamma(x) \cos \omega x dx \quad (2-245)$$

$$\epsilon_r' = (\epsilon_{rs} - \epsilon_{r\infty}) \int_0^\infty \Gamma(x) \sin \omega x dx \quad (2-246)$$

These two equations indicate that the real and imaginary parts of complex permittivity depend on the same decay function $\Gamma(x)$. According to the Fourier theorem, $\Gamma(x)$ can be expressed in the following form:

$$\Gamma(x) = \frac{2}{\pi} \int_0^\infty \frac{\epsilon_r(u) - \epsilon_{r\infty}}{\epsilon_{rs} - \epsilon_{r\infty}} \cos ux dx \quad (2-247)$$

or

$$\Gamma(x) = \frac{2}{\pi} \int_0^\infty \frac{\epsilon_r'(u)}{\epsilon_{rs} - \epsilon_{r\infty}} \sin ux dx \quad (2-248)$$

These two equations must be equal, so ϵ_r must be related to ϵ_r' . Substituting Equation 2-248 into Equation 2-245, we obtain

$$\begin{aligned} \epsilon_r(\omega) - \epsilon_{r\infty} &= \frac{2}{\pi} \int_0^\infty \left[\int_0^\infty \epsilon_r'(u) \sin ux du \right] \cos \omega x dx \\ &= \frac{2}{\pi} \int_0^\infty [\sin ux \cos \omega x] \epsilon_r'(u) du \\ &= \frac{2}{\pi} \int_0^\infty \frac{\epsilon_r'(u)u}{u^2 - \omega^2} du \end{aligned} \quad (2-249)$$

Similarly, substitution of Equation 2-247 into Equation 2-246 yields

$$\epsilon'_r(\omega) = \frac{2}{\pi} \int_0^\infty [\epsilon_r(u) - \epsilon_{r\infty}] \frac{\omega}{u^2 - \omega^2} du \quad (2-250)$$

Equations 2-249 and 2-250 are called the Kramers–Kronig relations. These relations are very useful under certain situations because they enable the determination of the value of one parameter (ϵ_r or ϵ'_r) from the other parameter (ϵ'_r or ϵ_r) at any frequency. This is particularly important when, for some reason, one of the parameters cannot be measured. These relations are valid for any type of polarization, and from them a complete spectrum of ϵ_r and ϵ'_r over a wide range of frequencies can be obtained—provided that one of the parameters (ϵ_r or ϵ'_r) can be measured throughout the same frequency range.

If we set $\omega = 0$ for the case of DC fields, Equation 2-249 becomes

$$\begin{aligned} \epsilon_{rs} - \epsilon_{r\infty} &= \frac{2}{\pi} \int_0^\infty \epsilon'_r(u) \frac{du}{u} = \frac{2}{\pi} \int_0^\infty \epsilon'_r(\omega) \frac{d\omega}{\omega} \\ &= \frac{2}{\pi} \int_0^\infty \epsilon'_r(\omega) d(\ln\omega) \end{aligned}$$

or

$$\int_0^\infty \epsilon'_r(\omega) d(\ln\omega) = \frac{\pi}{2} (\epsilon_{rs} - \epsilon_{r\infty}) \quad (2-251)$$

This implies that the total area under the curve of ϵ'_r vs $\ln\omega$ is simply related to $\epsilon_{rs} - \epsilon_{r\infty}$, the extreme values of the dielectric constants, and is independent of their dispersion mechanisms. We shall return to this interesting feature in the Section 2.7, which deals with the distribution of relaxation times.

2.6.5 Debye Equations, Absorption, and Dispersion for Dynamic Polarizations

Strictly speaking, no material is free of dielectric losses, and therefore, no material is free of absorption and dispersion. This implies that there is no material having frequency-independent ϵ_r and ϵ'_r . In fact, the dispersion in ϵ_r and ϵ'_r is an intrinsic property of all dielectric materials, and all other properties have to coexist with it. In this section, we shall first show the derivation of the Debye equations and then discuss the absorption and dispersion phenomena.

By expressing $F_m \cos \omega t = F_m \exp(j\omega t)$, $F_m \sin \omega t$ can be expressed as

$$\begin{aligned} F_m \sin \omega t &= F_m \exp[j(\omega t - \pi/2)] \\ &= -jF_m \exp(j\omega t) \end{aligned} \quad (2-252)$$

and Equation 2-239 can be rewritten as

$$\begin{aligned} P &= \left[\epsilon_{r\infty} - 1 + \frac{(\epsilon_{rs} - \epsilon_{r\infty})}{1 + \omega^2 \tau_o^2} \right] \epsilon_o F_m \exp(j\omega t) \\ &\quad - j \left[\frac{(\epsilon_{rs} - \epsilon_{r\infty}) \omega \tau_o}{1 + \omega^2 \tau_o^2} \right] \epsilon_o F_m \exp(j\omega t) \end{aligned} \quad (2-253)$$

From Equations 2-68 and 2-222, P can also be expressed as

$$\begin{aligned} P &= (\epsilon_r^* - 1) \epsilon_o F \\ &= [(\epsilon_r - 1) - j\epsilon'_r] \epsilon_o F_m \exp(j\omega t) \end{aligned} \quad (2-254)$$

By comparing Equation 2-253 with Equation 2-254, we have

$$\epsilon_r^* = \epsilon_r - j\epsilon'_r = \epsilon_{r\infty} + \frac{\epsilon_{rs} - \epsilon_{r\infty}}{1 + j\omega \tau_o} \quad (2-255)$$

$$\epsilon_r = \epsilon_{r\infty} + \frac{\epsilon_{rs} + \epsilon_{r\infty}}{1 + \omega^2 \tau_o^2} \quad (2-256)$$

$$\epsilon'_r = \frac{(\epsilon_{rs} - \epsilon_{r\infty}) \omega \tau_o}{1 + \omega^2 \tau_o^2} \quad (2-257)$$

and

$$\tan \delta = \frac{\epsilon'_r}{\epsilon_r} = \frac{(\epsilon_{rs} - \epsilon_{r\infty}) \omega \tau_o}{\epsilon_{rs} + \epsilon_{r\infty} \omega^2 \tau_o^2} \quad (2-258)$$

Equations 2-255 through 2-257 are generally called the Debye equations for dynamic polarization with only one relaxation time τ_o .^{31,50,51} These equations are based on the assumption that the decay function is exponential. Similar to the situation for static polarization, Debye equations are satisfactory only for the condition $\epsilon_{rs} - \epsilon_{r\infty} < 1$, which can be fulfilled only in dilute solutions because the derivation of Equations 2-255 through 2-257 has not taken into account the interaction between particles. ϵ_r and ϵ'_r are temperature dependent through the temperature dependence of $(\epsilon_{rs} - \epsilon_{r\infty})$ and τ_o .

We can easily find the frequency at which ϵ'_r and $\tan \delta$ are maximal by setting $\frac{d\epsilon'_r}{d\omega} = 0$ and

$\frac{d(\tan \delta)}{d\omega} = 0$. The maximum value of ε'_r occurs at ω_0 when

$$\omega_0 \tau_0 = 1 \quad (2-259)$$

At this frequency, ε_r , ε'_r and $\tan \delta$ are given by

$$\varepsilon_r|_{\omega=\omega_0} = \frac{\varepsilon_{rs} + \varepsilon_{r\infty}}{2} \quad (2-260)$$

$$\varepsilon'_r|_{\omega=\omega_0} = \varepsilon'_{r(\max)} = \frac{\varepsilon_{rs} - \varepsilon_{r\infty}}{2} \quad (2-261)$$

$$\tan \delta|_{\omega=\omega_0} = \frac{\varepsilon_{rs} - \varepsilon_{r\infty}}{\varepsilon_{rs} + \varepsilon_{r\infty}} \quad (2-262)$$

However, the value of $\tan \delta$ at ω_0 is not maximal. The maximum value of $\tan \delta$ occurs at ω_δ when

$$\omega_\delta \tau_0 = (\varepsilon_{rs}/\varepsilon_{r\infty})^{1/2} > 1 \quad (2-263)$$

At this frequency, $\tan \delta$ is given by

$$\tan \delta|_{\omega=\omega_\delta} = \tan \delta_{(\max)} = \frac{\varepsilon_{rs} - \varepsilon_{r\infty}}{2(\varepsilon_{rs}\varepsilon_{r\infty})^{1/2}} \quad (2-264)$$

The values of ε_r , ε'_r , and $\tan \delta$ as functions of ω are shown in Figure 2-30.

Equations 2-256 and 2-257 can be written as

$$\frac{\varepsilon_r - \varepsilon_{r\infty}}{\varepsilon_{rs} - \varepsilon_{r\infty}} = \frac{1}{1 + \omega^2 \tau_0^2} \quad (2-265)$$

$$\frac{\varepsilon'_r}{\varepsilon_{rs} - \varepsilon_{r\infty}} = \frac{\omega \tau_0}{1 + \omega^2 \tau_0^2} \quad (2-266)$$

These two equations are the parametric equations of a circle in the $\varepsilon_r - \varepsilon'_r$ plane. By

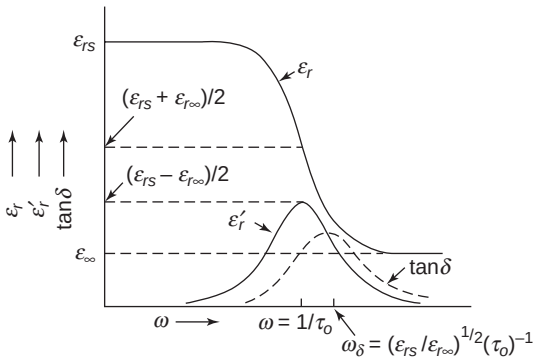


Figure 2-30 ε_r , ε'_r and $\tan \delta$ as functions of ω for cases with negligible contribution of σ due to carrier migration.

eliminating $\omega \tau_0$ from these two equations, we obtain

$$\left(\varepsilon_r - \frac{\varepsilon_{rs} + \varepsilon_{r\infty}}{2} \right)^2 + \varepsilon_r'^2 = \left(\frac{\varepsilon_{rs} - \varepsilon_{r\infty}}{2} \right)^2 \quad (2-267)$$

Of course, only the semicircle of Equation 2-267, over which ε'_r is positive, as shown in Figure 2-31, has physical significance.

The major disadvantage of this method for finding ε_r or ε'_r is that the frequency, which is the independent variable and is the parameter most accurately measured, is not explicitly shown in the Argand diagram. It should also be noted that the derivation of Equations 2-255 through 2-257 is based on the following assumptions for simplicity: The local field is the same as the applied field F ; the conductivity of the materials is negligible; and all dipoles have only one identical relaxation time τ_0 . The effects of these assumptions on ε_r and ε'_r will be discussed later in this section.

Debye equations emerge directly from the Kramers–Kronig relations. They show explicitly the frequency dependence of ε_r and ε'_r . The type of dispersion shown in Figure 2-30 is generally considered the normal dispersion. All types of polarization may be grouped into two major regimes: the resonance regime and the relaxation regime. Polarizations associated with vibrations of electrons (i.e., electronic or optical polarization) or with vibrations of atoms or ions (i.e., atomic or ionic polarization) belong to the resonance regime because in the

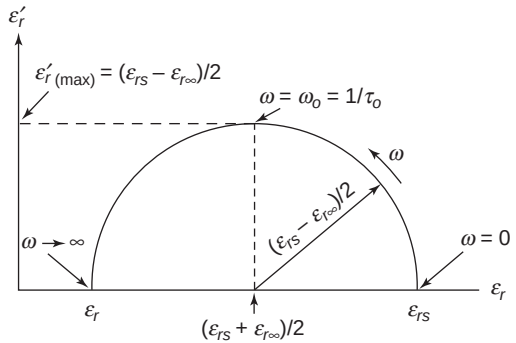


Figure 2-31 Argand diagram of $\varepsilon_r - \varepsilon'_r$ relation for cases with only one relaxation time τ_0 .

polarization a resonance will occur when the frequency of the excitation field is close to the natural frequency of the vibration or oscillation system. Polarizations involving the movements of charges either by orientation (i.e., orientational polarization) or through the migration of charge carriers (i.e., hopping or space charge polarization) belong to the relaxation regime because during the polarization or depolarization processes, a relaxation phenomenon occurs due to the time required for the charge carriers to overcome the inertia arising from the surrounding medium in order to proceed in their movement. The variation of ϵ_r and ϵ'_r with frequency shown in Figure 2-30 illustrates schematically the typical dispersion behavior for polarizations in the relaxation regime.

However, in the resonance regime, the optical constants, i.e., the complex refractive index n^* , which consists of the refractive index n and the extinction coefficient (or called the absorption index) k , can be expressed in terms of the complex dielectric constant as

$$\begin{aligned} n^* &= n - jk \\ &= (\epsilon_r^*)^{1/2} = (\epsilon_r - j\epsilon'_r)^{1/2} \end{aligned} \quad (2-268)$$

Rearrangement of Equation 2-268 leads to

$$\epsilon_r = n^2 - k^2 \quad (2-269)$$

$$\epsilon'_r = 2nk \quad (2-270)$$

The frequency dependence of ϵ_r and ϵ'_r in the optical frequency range, i.e., $f > 10^{11}$ Hz, and the dispersion of ϵ_r and ϵ'_r (or n and k) are anomalous. So this type of dispersion is sometimes referred to as anomalous dispersion. Figure 3-32 in Chapter 3 shows schematically this anomalous dispersion phenomenon, and Figure 3-33 shows the variation of ϵ_r and ϵ'_r with frequency for various types of polarization in both the resonance and the relaxation regimes. For more details, see Absorption and Dispersion in Chapter 3.

The Effects of the Local Field

So far in Section 2.6, all equations derived are based on the assumption that the local field is the same as the applied field. This, of course, is

valid in dilute gases. In Section 2.5, we discussed the internal field or the Lorentz field (Equation 2-188). By taking into account the local field based on the Lorentz field, Equations 2-255 through 2-258 and other equations are still valid if τ_o is replaced with τ'_o , which is given by¹⁷

$$\tau'_o = \frac{\epsilon_{rs} + 2}{\epsilon_{r\infty} + 2} \tau_o \quad (2-271)$$

In general, the measured relaxation time based on the Cole–Cole plot is higher than the real relaxation time, possibly due to the local field effect. Note that the local field given by Equation 2-193 is not applicable to condensed polar polymer materials, as explained in Section 2.5.2.

The Effects of DC Conductivity

If the DC conductivity σ is not negligibly small, then σ will contribute to the imaginary part of complex permittivity. The total complex permittivity becomes

$$\epsilon^* = \epsilon - j\left(\epsilon' + \frac{\sigma}{\omega}\right) = \left[\epsilon_r - j\left(\epsilon'_r + \frac{\sigma}{\omega\epsilon_o}\right)\right]\epsilon_o$$

or

$$\epsilon_r^* = \epsilon_r - j\epsilon'_r - j\frac{\sigma}{\omega} \quad (2-272)$$

Taking the DC conductivity into account, the Debye equation (Equation 2-255), for example, becomes

$$\epsilon_r^* - \epsilon_{r\infty} = \frac{\epsilon_{rs} - \epsilon_{r\infty}}{1 + j\omega\tau_o} - j\frac{\sigma}{\omega\epsilon_o} \quad (2-273)$$

and Equations 2-257 and 2-258 become

$$\epsilon'_r = \frac{(\epsilon_{rs} - \epsilon_{r\infty})\omega\tau_o}{1 + \omega^2\tau_o^2} + \frac{\sigma}{\omega\epsilon_o} \quad (2-274)$$

$$\tan \delta = \frac{\omega\epsilon_o(\epsilon_{rs} - \epsilon_{r\infty})\omega\tau_o + (1 + \omega^2\tau_o^2)\sigma}{\omega\epsilon_o(\epsilon_{rs} + \epsilon_{r\infty}\omega^2\tau_o^2)} \quad (2-275)$$

When $\omega\tau_o \ll 1$, Equations 2-274 and 2-275 reduce to

$$\epsilon'_r = \frac{\sigma}{\omega\epsilon_o} \quad (2-276)$$

$$\tan \delta = \frac{\sigma}{\omega \epsilon_0 \epsilon_{rs}} \quad (2-277)$$

When $\omega \tau_0 \approx 1$, Equations 2-274 and 2-275 reduce to

$$\epsilon'_r = \frac{\epsilon_{rs} - \epsilon_{r\infty}}{2} + \frac{\sigma \tau_0}{\epsilon_0} \quad (2-278)$$

$$\tan \delta = \frac{\epsilon_0(\epsilon_{rs} - \epsilon_{r\infty}) + 2\sigma \tau_0}{\epsilon_0(\epsilon_{rs} + \epsilon_{r\infty})} \quad (2-279)$$

When $\omega \tau_0 \gg 1$, Equations 2-274 and 2-275 reduce to

$$\epsilon'_r = \left(\frac{\epsilon_{rs} - \epsilon_{r\infty}}{\omega \tau_0} \right) \quad (2-280)$$

$$\tan \delta = \frac{\epsilon_0(\epsilon_{rs} - \epsilon_{r\infty}) + \sigma \tau_0}{\omega \tau_0 \epsilon_0 \epsilon_{r\infty}} \quad (2-281)$$

The variation of ϵ'_r and $\tan \delta$ with ω , including the effect of DC conductivity, is shown in Figure 2-32, and the Argand diagram of $\epsilon_r - \epsilon'_r$ relation showing this effect in Figure 2-33.

2.6.6 The Cole-Cole Plot

Debye equations (Equations 2-256 and 2-257) based on a single relaxation time do not suffice to describe the relaxation phenomena for most dielectric materials, e.g., polymers. In this case, a distribution of relaxation times is necessary to interpret the experimental data. In Section 2.7, we shall discuss the approach based on a distribution of relaxation times. However, to take into account the effect of a distribution of relaxation times, Cole and Cole⁵² have proposed that an Argand diagram, in which ϵ'_r is plotted as a function of ϵ_r , can be constructed following the empirical relation

$$\epsilon_r^* - \epsilon_{r\infty} = \epsilon_r - \epsilon_{r\infty} - j\epsilon'_r = \frac{\epsilon_{rs} - \epsilon_{r\infty}}{1 + (j\omega \tau_0)^{1-\alpha}} \quad (2-282)$$

where α is the parameters with $0 < \alpha < 1$. In fact, many materials, particularly polymers with long chains, have a broader $\epsilon'_r - \ell n \omega$ dispersion curve and a lower maximum loss than would be expected from the Debye semicircle shown in Figure 2-31.³⁹ In such cases, the $\epsilon_r - \epsilon'_r$ arc will fall inside the Debye semicircle. On

the basis of the suggestion of Cole and Cole, Equation 2-282 can be written as

$$\frac{\epsilon_r - \epsilon_{r\infty}}{\epsilon_{rs} - \epsilon_{r\infty}} = \frac{1 + (\omega \tau)^{1-\alpha} \sin \alpha \pi / 2}{1 + 2(\omega \tau)^{1-\alpha} \sin \alpha \pi / 2 + (\omega \tau)^{2(1-\alpha)}} \quad (2-283)$$

$$\frac{\epsilon'_r}{\epsilon_{rs} - \epsilon_{r\infty}} = \frac{(\omega \tau)^{1-\alpha} \cos \alpha \pi / 2}{1 + 2(\omega \tau)^{1-\alpha} \sin \alpha \pi / 2 + (\omega \tau)^{2(1-\alpha)}} \quad (2-284)$$

When $\alpha = 0$, Equations 2-283 and 2-284 reduce back to the Debye equations. The dispersion curve with $\alpha > 0$ is broader than that for a single relaxation time but still symmetrical about $\omega \tau = 1$. By setting $\frac{d\epsilon'_r}{d\omega} = 0$, we can find that the maximum loss also occurs at $\omega \tau = 1$. By eliminating $\omega \tau$ from Equations 2-283 and 2-284, we obtain

$$\left[\epsilon_r - \frac{(\epsilon_{rs} + \epsilon_{r\infty})}{2} \right]^2 + \left[\epsilon'_r + \frac{(\epsilon_{rs} - \epsilon_{r\infty})}{2} \tan \frac{\alpha \pi}{2} \right]^2 = \left[\frac{\epsilon_{rs} - \epsilon_{r\infty}}{2} \sec \frac{\alpha \pi}{2} \right]^2 \quad (2-285)$$

This equation represents a circle with the center at $\left[\frac{\epsilon_{rs} - \epsilon_{r\infty}}{2}, -\frac{(\epsilon_{rs} - \epsilon_{r\infty})}{2} \tan \frac{\alpha \pi}{2} \right]$ and the radius of $\left[\frac{\epsilon_{rs} - \epsilon_{r\infty}}{2} \sec \frac{\alpha \pi}{2} \right]$. This is shown in

Figure 2-34. The Cole-Cole plot exhibits a depressed semicircle because of $\alpha > 0$. Some experimental results agree well with Equations 2-283 and 2-284 with $\alpha > 0$.³⁹

Debye's $\epsilon_r - \epsilon'_r$ semicircle is based on $\alpha = 0$, which represents the materials' having only one single relaxation time. In general, there exists a distribution of relaxation times in all solid materials because there are always some nonuniformities in local domains, which would alter the individual dipoles or charges in addition to existing dipoles in the material. If the relaxation times are due to several different mechanisms, then the $\epsilon_r - \epsilon'_r$ arc will be asymmetrical.^{53,54} Several investigators⁵³⁻⁵⁶ have put forward suggestions to modify the Cole-Cole empirical equations. Some of the equations are listed below for comparison purposes:

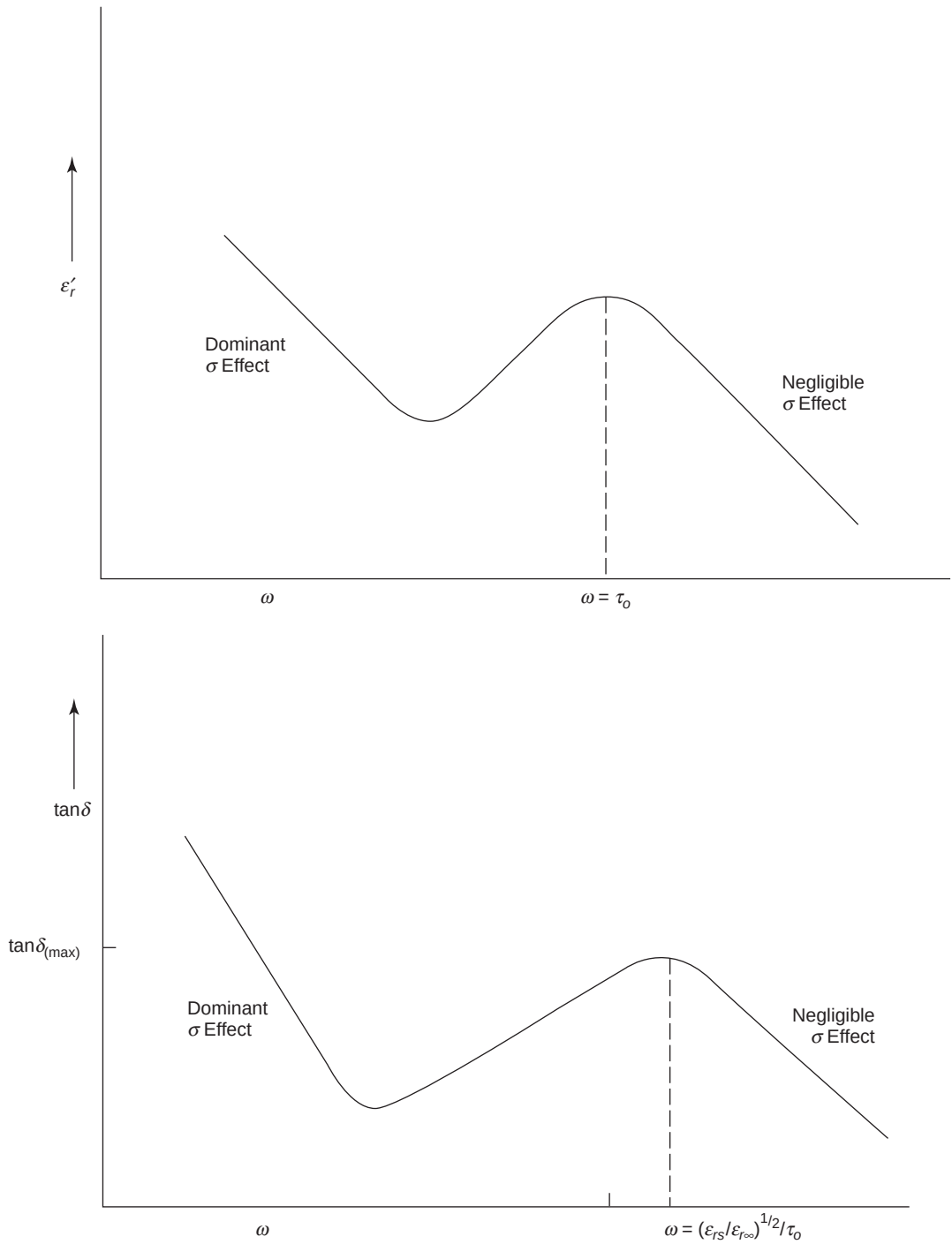


Figure 2-32 ϵ'_r and $\tan \delta$ as functions of ω , taking into account the effect of DC conductivity.

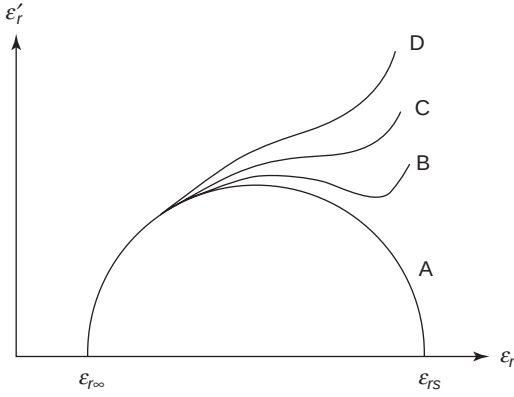


Figure 2-33 The effect of DC conductivity on the $\epsilon_r - \epsilon'_r$ arc. (A) $\sigma = 0$, (B) $\sigma = \sigma_1 > 0$, (C) $\sigma_2 > \sigma_1$, (D) $\sigma_3 = \sigma_2$.

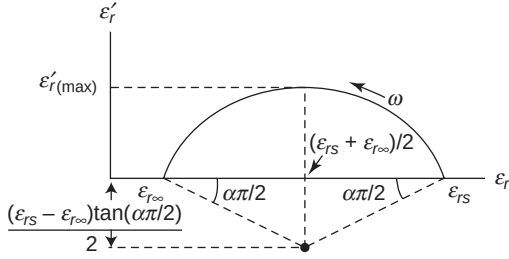


Figure 2-34 The Cole–Cole plot for dielectric material with a set of relaxation times.

Debye equation

$$\epsilon_r^* - \epsilon_{r\infty} = \frac{\epsilon_{rS} - \epsilon_{r\infty}}{1 + j\omega\tau} \quad (2-255)$$

Cole–Cole equation

$$\epsilon_r^* - \epsilon_{r\infty} = \frac{\epsilon_{rS} - \epsilon_{r\infty}}{1 + (j\omega\tau)^{1-\alpha}} \quad (2-282)$$

Davidson–Cole equation

$$\epsilon_r^* - \epsilon_{r\infty} = \frac{\epsilon_{rS} - \epsilon_{r\infty}}{(1 + j\omega\tau)^\beta} \quad (2-286)$$

Havriliak–Negami equation

$$\epsilon_r^* - \epsilon_{r\infty} = \frac{\epsilon_{rS} - \epsilon_{r\infty}}{[1 + (j\omega\tau)^{1-\alpha}]^\beta} \quad (2-287)$$

All of these equations depend on the values of α and β chosen within the ranges $0 < \alpha < 1$ and $0 < \beta < 1$. Obviously, the Davidson–Cole

equation will reduce to the Debye equation for $\beta = 1$. But, with $\beta < 1$, the $\epsilon_r - \epsilon'_r$ arc will become asymmetrical, as shown in Figure 2-35(a). It has been found that several materials obey the Davidson–Cole equation, such as glycerol triacetate⁵⁵ and Pyralene (a fluid mixture of chlorinated diphenols in chlorobenzene).¹⁷ However, the Havriliak–Negami equation gives a much better fit to most experimental results if α and β are properly chosen. The $\epsilon_r - \epsilon'_r$ arc for $\alpha = 2/3$ and $\beta = 1/2$ is shown in Figure 2-35(b). In the Havriliak–Negami equation, the parameters α and β are not based on the physics of the dielectric polarization, although the modification of the original Cole–Cole equation empirically may make the equation better fit experimental results. The modification does not lead to a better understanding of the physics behind the distribution of relaxation times.

Fuoss and Kirkwood⁵⁶ have also put forward a semi-empirical equation, called the Fuoss–Kirkwood equation, to relate only the imaginary part ϵ'_r of the complex dielectric constant with frequency, which is given by

$$\epsilon'_r(\omega) = \frac{2\epsilon'_{r(\max)}}{(\omega\tau)^\lambda + (\omega\tau)^{-\lambda}} \quad (2-288)$$

where λ is the parameter with the value $0 < \lambda < 1$, and $\epsilon'_{r(\max)}$ is the maximum value of the loss factor when $\omega\tau = 1$. On a logarithmic frequency scale, the $\epsilon'_r - (\ln\omega)$ relation exhibits a symmetrical bell-shaped curve about a central frequency $\omega = 1/\tau$, which is broader than the corresponding Debye curve (Equation 2-257), but very similar in shape to, although not identical with, that based on the Cole–Cole equation (Equation 2-284). The difference can be seen by comparing these two equations. Cole^{57,58} has discussed the correlation between these two equations in some detail.

2.6.7 Temperature Dependence of Complex Permittivity

Complex permittivity is a complex function of both the frequency and the temperature, apart from the dependence of other parameters such as pressure, etc. We have briefly discussed the frequency dependence of ϵ_r and ϵ'_r . The

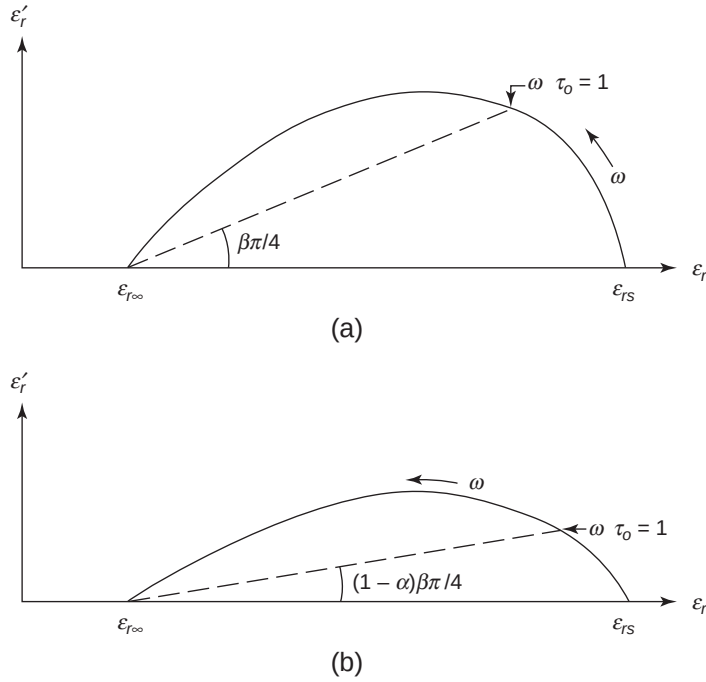


Figure 2-35 The $\epsilon_r - \epsilon'_r$ arcs based on (a) the Davidson-Cole equation for $\beta = 1/2$, and (b) the Havriliak-Negami equation for $\alpha = 1/3$ and $\beta = 1/2$.

temperature dependence of ϵ_r and ϵ'_r is mainly through the temperature dependence of the relaxation time which can, for a single relaxation time, be expressed as

$$\tau = \tau_h \exp(H/kT) \quad (2-289)$$

where τ_h is a pre-exponential factor and H is the activation energy.²⁸ By replacing τ_0 with a temperature-dependent τ , the Debye equations (Equations 2-256 and 2-257) can be written as

$$\epsilon_r = \epsilon_{r\infty}^T + \frac{\epsilon_{rs}^T - \epsilon_{r\infty}^T}{1 + \omega^2 \tau_h^2 \exp(2H/kT)} \quad (2-290)$$

$$\epsilon'_r = (\epsilon_{rs}^T - \epsilon_{r\infty}^T) \frac{\omega \tau_h \exp(H/kT)}{1 + \omega^2 \tau_h^2 \exp(2H/kT)} \quad (2-291)$$

where ϵ_{rs}^T and $\epsilon_{r\infty}^T$ are the value of ϵ_{rs} and $\epsilon_{r\infty}$ at temperature T . It can be seen from these two equations that the dielectric constant ϵ_r and the loss peak $\epsilon'_{r(\max)}$ decrease with increasing temperature, and the loss peak shifts toward higher temperatures, as shown in Figure 2-36. If

$(\epsilon_{rs} - \epsilon_{r\infty})$ for nondipolar materials is not temperature dependent, then the area beneath the ϵ'_r vs $1/T$ curve depends only on the activation energy H and does not depend explicitly on frequency, based on the following equation:

$$\int_0^\infty \epsilon'_r d\left(\frac{1}{T}\right) = (\epsilon_{rs} - \epsilon_{r\infty}) \frac{k}{H} \left(\frac{\pi}{2} - \tan^{-1} \omega \tau \right) \quad (2-292)$$

For $\tan^{-1} \omega \tau < \pi/2$, the area beneath the ϵ'_r vs $1/T$ curve is practically independent of frequency.

However, for dipolar materials with dipolar molecules, $(\epsilon_{rs}^T - \epsilon_{r\infty}^T)$ is temperature dependent. According to the theories of Onsager, Frohlich, and Debye (see Section 2.5.2), we can write

$$(\epsilon_{rs}^T - \epsilon_{r\infty}^T) = \frac{A}{T} \quad (2-293)$$

where A is a constant. In this case, the area beneath the ϵ'_r vs $1/T$ curve is more complicated, and it is given by^{39,59}

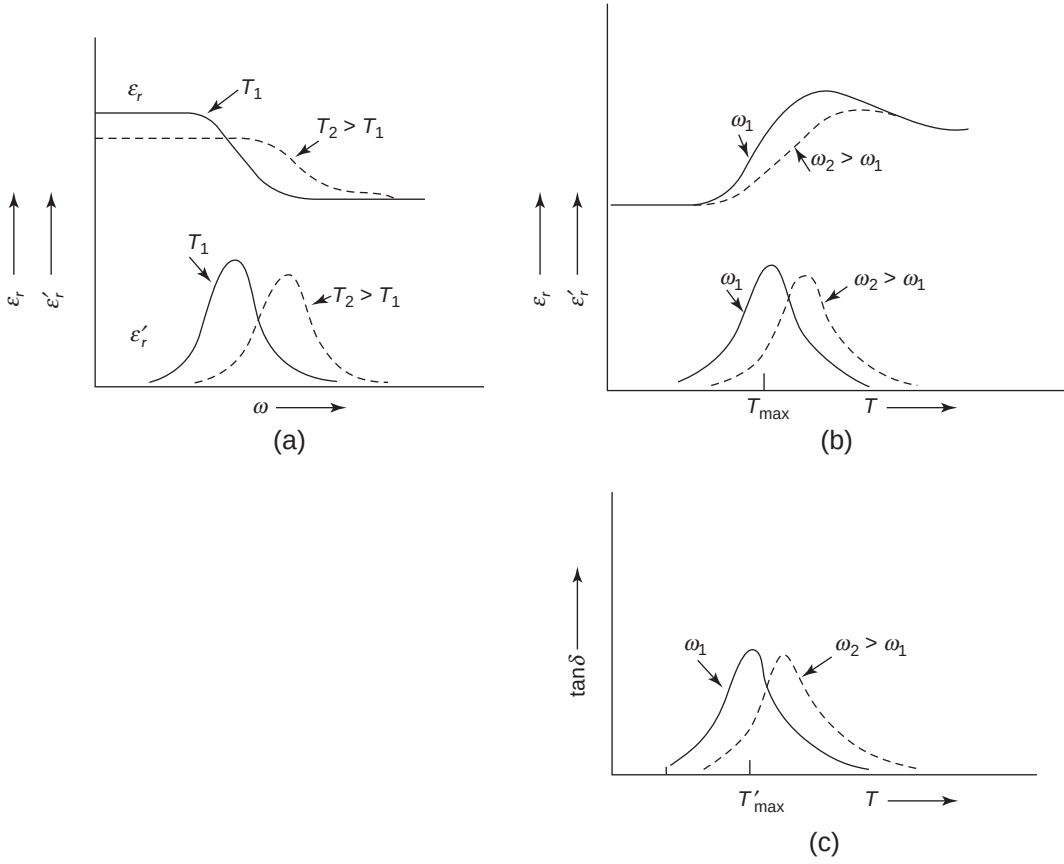


Figure 2-36 (a) ϵ_r and ϵ'_r as functions of ω for fixed temperatures T_1 and T_2 with $T_2 > T_1$, (b) ϵ_r and ϵ'_r as functions of temperature for fixed frequencies ω_1 and ω_2 with $\omega_2 > \omega_1$, and (c) $\tan \delta$ as a function of temperature for fixed frequencies ω_1 and ω_2 with $\omega_2 > \omega_1$. For the same frequency, $T_{\max}$ for $\epsilon'_{r(\max)}$ is higher than $T'_{\max}$ for $\tan \delta_{(\max)}$.

$$\int_0^\infty \epsilon'_r d\left(\frac{1}{T}\right) = \frac{Ak^2}{H} \left[-\frac{\pi n \omega \tau}{2} + \sum_{r=1}^\infty \frac{(-1)^r (\omega \tau)^{2r+1}}{(2r+1)^2} \right] \quad (2-294)$$

By denoting $T_{\max}$ as the temperature at which ϵ'_r is maximum, then

$$(\epsilon_{rs}^{T_{\max}} - \epsilon_{r\infty}^{T_{\max}}) = \frac{A}{T_{\max}} \quad (2-295)$$

$$\tau_{T_{\max}} = \tau_h \exp(H/kT_{\max}) \quad (2-296)$$

In general, $\omega \tau_{T_{\max}}$ differs little from unity.⁶⁰ Thus, for the frequency range normally encountered in practice ($\omega \tau < 1$), Equation 2-294 can be approximated to

$$\int_0^\infty \epsilon'_r d\left(\frac{1}{T}\right) = (\epsilon_{rs}^{T_{\max}} - \epsilon_{r\infty}^{T_{\max}}) \frac{\pi k}{2H} \quad (2-297)$$

For $\omega \tau < 1$, Equation 2-297 differs from Equation 2-292 only in the $(\epsilon_{rs}^T - \epsilon_{r\infty}^T)$ term. Thus, the area beneath the ϵ'_r vs $1/T$ curve is still practically independent of the frequency, as shown in Figure 2-36. It should be noted that for the same ω , $T'_{\max}$, for $\tan \delta$ to be maximum, is lower than $T_{\max}$ for ϵ'_r to be maximum.

The same concept for temperature dependence of ϵ_r and ϵ'_r can be applied to other empirical equations—the Cole–Cole equation, the Davidson–Cole equation, and the Havriliak–Negami equation. The measurement of ϵ'_r as a function of frequency at a fixed temperature is equivalent to the measurement of ϵ'_r as a

function of temperature at a fixed frequency. Both methods should provide similar information provided that there is no change in the specimen and the DC conductivity is negligible over the temperature range for the measurements.

2.6.8 Field Dependence of Complex Permittivity

From Equations 2-68 and 2-224, it can be seen that if the relationship between the polarization P and the electric field F is linear and the electrical conductivity σ is constant, then the dielectric constant ϵ_r and the loss factor ϵ_r' should be independent of electric field. This is true only when the total field during the measurement of these parameters is low. However, when the specimen is biased with a strong field, σ may not be constant or the P - F relationship may become nonlinear, then ϵ_r and ϵ_r' , measured with a small-signal AC voltage, will be field-dependent. Since the total current density in a material specimen under an AC field is given by

$$J_T = J + \epsilon^* \frac{dF}{dt} = \sigma F + j(\epsilon_r - j\epsilon_r')\epsilon_0 F \quad (2-298)$$

$$= (\sigma + \epsilon_r'\epsilon_0)F + j\epsilon_r\epsilon_0 F$$

the complex relative permittivity, including the electrical conductivity σ , can be expressed as

$$\epsilon_r^* = \epsilon_r - j\epsilon_r' = \epsilon_r - j\left(\epsilon_r' + \frac{\sigma}{\omega\epsilon_0}\right) \quad (2-299)$$

where J is the conduction current, ϵ_r' is the total loss factor, including the loss factor due to the loss involved in the polarization processes and the loss due to the electrical conduction involving the movement of charge carriers. In the following sections, we shall discuss the effects of strong electric fields on complex permittivity for three categories of materials: semiconducting, ferroelectric, and insulating.

Semiconducting Materials

In semiconductors, the concentration of free charge carriers and hence the conductivity are high. The value of the conductivity σ at

microwave fields of frequencies of the order of 10 GHz (10×10^9 Hz) or higher may become complex. Only at AC fields of frequencies low enough for the current capable of following the field, the conductivity would have the value appropriate to that at a DC field of the same magnitude. If the total field applied to a semiconductor specimen is

$$F = F_{DC} + F_{AC} \cos \omega t \quad (2-300)$$

with $F_{AC} \ll F_{DC}$. For AC fields not high enough to heat the charge carriers, the current may no longer follow the field at frequencies for which $\omega\tau$ is comparable to unity, where τ can be taken as the mean free time between collisions. Obviously, τ depends on both the temperature and the applied field magnitude; it decreases with increasing temperature and with increasing field magnitude. So we can describe the situation by expressing the AC conductivity as a complex conductivity $\sigma^*(\omega)$ with the real part $\sigma(\omega)$ representing the in-phase conductivity, which means the current capable of following the field, and the imaginary part $\sigma'(\omega)$ representing the out-of-phase conductivity ($\pi/2$ lagging the field). For a simple band structure, the AC conductivity at low AC fields is given by

$$\sigma^*(\omega) = \frac{nq^2}{m^*} \left[\frac{\langle \Delta E_{ave} \tau_m \rangle}{\langle \Delta E_{ave} (1 + \omega^2 \tau_m^2) \rangle} - j \frac{\omega \tau_m \langle \Delta E_{ave} \tau_m \rangle}{\langle \Delta E_{ave} (1 + \omega^2 \tau_m^2) \rangle} \right] \quad (2-301)$$

$$= \sigma(\omega) + j\sigma'(\omega)$$

where ΔE_{ave} is the average energy of the carrier, n is the carrier concentration, q is the electronic charge, m^* is the effective mass of the carrier, and τ_m is the momentum relaxation time, which is the average time required for a component of the carrier momentum in any direction to be reduced to $1/e$ of its value.^{61,62} The τ mentioned above can be considered to be identical to τ_m .⁶¹ The symbol $\langle \rangle$ indicates the average values to be taken over the carrier distribution in the specimen. The out-of-phase current ($\pi/2$ lagging F_{AC}) may be thought of as a contribution to the displacement current, so that we can write the dielectric constant and the loss factor

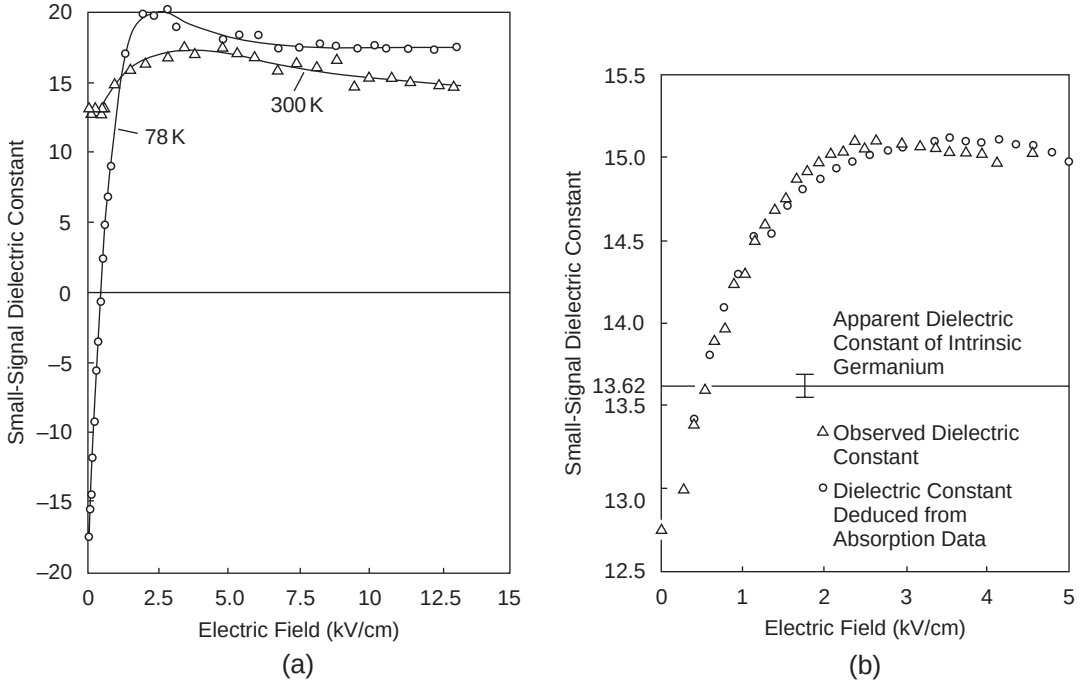


Figure 2-37 The variation of small-signal dielectric constant with DC bias field: (a) for *n*-Ge specimen of low-field resistivity of 1.9 ohm-cm at 300 K and small-signal frequency of 70 GHz, and (b) for *n*-Ge specimen of low field resistivity of 4.7 ohm-cm at 300 K and small signal frequency of 35 GHz.

for semiconductors at high frequency AC fields as

$$\epsilon_{rT}(\omega) = \epsilon_r(\omega) + \frac{\sigma^1(\omega)}{\omega\epsilon_o} \quad (2-302)$$

$$\epsilon'_{rT}(\omega) = \epsilon'_r(\omega) + \frac{\sigma(\omega)}{\omega\epsilon_o} \quad (2-303)$$

Since $\sigma'(\omega)$ is negative, the contribution of the carriers to the dielectric constant is also negative. This implies that $\sigma'(\omega)$ tends to decrease the dielectric constant, and it may even make the dielectric constant negative for high carrier concentrations.

Using a small-signal AC field $F_{AC} \cos \omega t$ superimposed on a DC bias field F_{DC} , as given by Equation 2-300, the field dependence of the dielectric constant for *n* type germanium specimens with low-field resistivity of 1.9 ohm-cm and the small-signal frequency of 70 GHz is shown in Figure 2-37(a). The results are from Conwell, Fowler, and Zucker.^{62,63} At $T = 300^\circ\text{K}$ and $F_{DC} = 0$, ϵ_{rT} does not lie much below ϵ_r , which is about 16, because of the small value

of $\omega\tau_m$ ($\omega\tau_m \leq 0.1$). However, at $T = 78^\circ\text{K}$ and $F_{DC} = 0$, τ_m becomes larger, making $\omega\tau_m$ close to unity. In this case, the carrier contribution becomes large enough to result in a large negative value of ϵ_{rT} . The DC bias field tends to reduce the value of τ_m and hence the value of $\omega\tau_m$, thus reducing the carrier contribution to the dielectric constant, as shown in Figure 2-37(a). Similar results have also been observed for *n*-type germanium specimens with low-field resistivity of 4.7 ohm-cm and the small-signal frequency of 35 GHz, as shown in Figure 2-37(b). The results are from Gibson, Granville, and Pagie.^{62,64} Obviously, the loss factor ϵ'_{rT} is also field dependent through the field dependence of τ_m , as indicated in Equations 2-301 and 2-303.

Ferroelectric Materials

For ferroelectric materials, the applied electric field has a marked effect on complex permittivity. In Thermodynamic Theory in Chapter 4, we mention that the application of a suitable

DC field can shift the Curie point temperature of BaTiO_3 to a higher temperature, converting the paraelectric state to the ferroelectric state; this converted ferroelectric state will return to the paraelectric state when the electric field is removed. However, dielectric constant is field dependent in the paraelectric state at temperatures above the Curie point temperature T_c . Typical results for triglycine sulphate (TGS) and potassium dihydrogen phosphate (KDP) are shown in Figure 2-38. In the measurements of the effects of an applied electric field, the electric field used to induce the polarization P was a 200Hz sinusoidal field, while for the measurements of the dielectric constants a small-signal AC field was used and kept constant and low at a fixed frequency of 50kHz. The results shown in Figure 2-38(a) for TGS are from Triebwasser,⁶⁵ and those in Figure 2-38(b) for KDP are from Baumpartner.⁶⁶

When no electric field is applied to the specimen, the polarization P is equal to zero. When a finite electric field F is applied, it induces a finite value of the polarization P . For $T > T_c$ and $F = 0$, the dielectric constant follows closely the Curie–Weiss relation (see Chapter 4)

$$\epsilon_r = \frac{C}{T - T_c} \quad (2-304)$$

where C is known as the Curie constant. In ferroelectric materials, there exists an intrinsic bias, which produces just the same effects on the dielectric constant as an applied bias.⁶⁷ Since the applied electric field along the ferroelectric axis is opposite in sign to the intrinsic bias, the net dielectric constant decreases with increasing applied field, as shown in Fig. 2.38.

Insulating Materials

Insulating materials generally refer to the materials with a large bandgap, a low concentration of free charge carriers, a small carrier mobility, and hence an extremely small conductivity. Therefore, for these materials we can ignore the effects of electrical conductivity σ because it would be negligibly small. The field dependence of the complex permittivity is associated mainly with the nonlinear field dependence of the polarization. Excluding the hopping and

the space charge polarizations for simplicity, the major types of the polarization are electronic polarization (P_e), atomic polarization (P_i), and orientational polarization (P_o). For P_e and P_i , the polarization-field relationship should remain linear even at high fields. But for P_o , the polarization-field relationship is linear only at low fields and becomes nonlinear at high fields.

Dipolar molecules in a material are similar to balls with a positive charge on one end and a negative charge of the same magnitude on the other end in a viscous fluid. They are in motion due to thermal agitation. When an electric field is applied across a material specimen between two electrodes, these dipolar molecules will orient toward the direction of the field, with the end of positive charge facing the negative electrode (cathode) and the end of negative charge facing the positive electrode (anode), resulting in orientational polarization. The magnitude of the polarization depends on the angle made between the dipole axis and the field direction, which, in turn, depends on the applied field strength and the temperature. The degree of polarization is governed by the Langevin function, as discussed in Section 2.3.3. It should be noted that molecules are generally not spherical in shape, but rather ellipsoidal. So, even in nondipolar materials, the molecules are not dipolar but are ellipsoidal in shape; the molecules will still orient under an applied electric field in order to reduce their potential energy. In fact, a molecule can be considered a configuration built up of a number of charges; it will be distorted in an elastic system under an external electric field. Including its distortion, any molecule can be represented by an ellipsoid with three principal polarizabilities related to the three perpendicular axes. If such a molecule is placed in a field, the axis of highest polarizability will orient toward the direction of the field to reduce its potential energy to a minimum. Therefore, there will exist an orientation effect in a manner similar to the dipolar molecules (see Classical Approach in Section 2.3.1).

It can be imagined that in a high electric field, all individual molecules will be largely oriented, so that there can be little extra polarization from the further orientation of dipolar or nondipolar

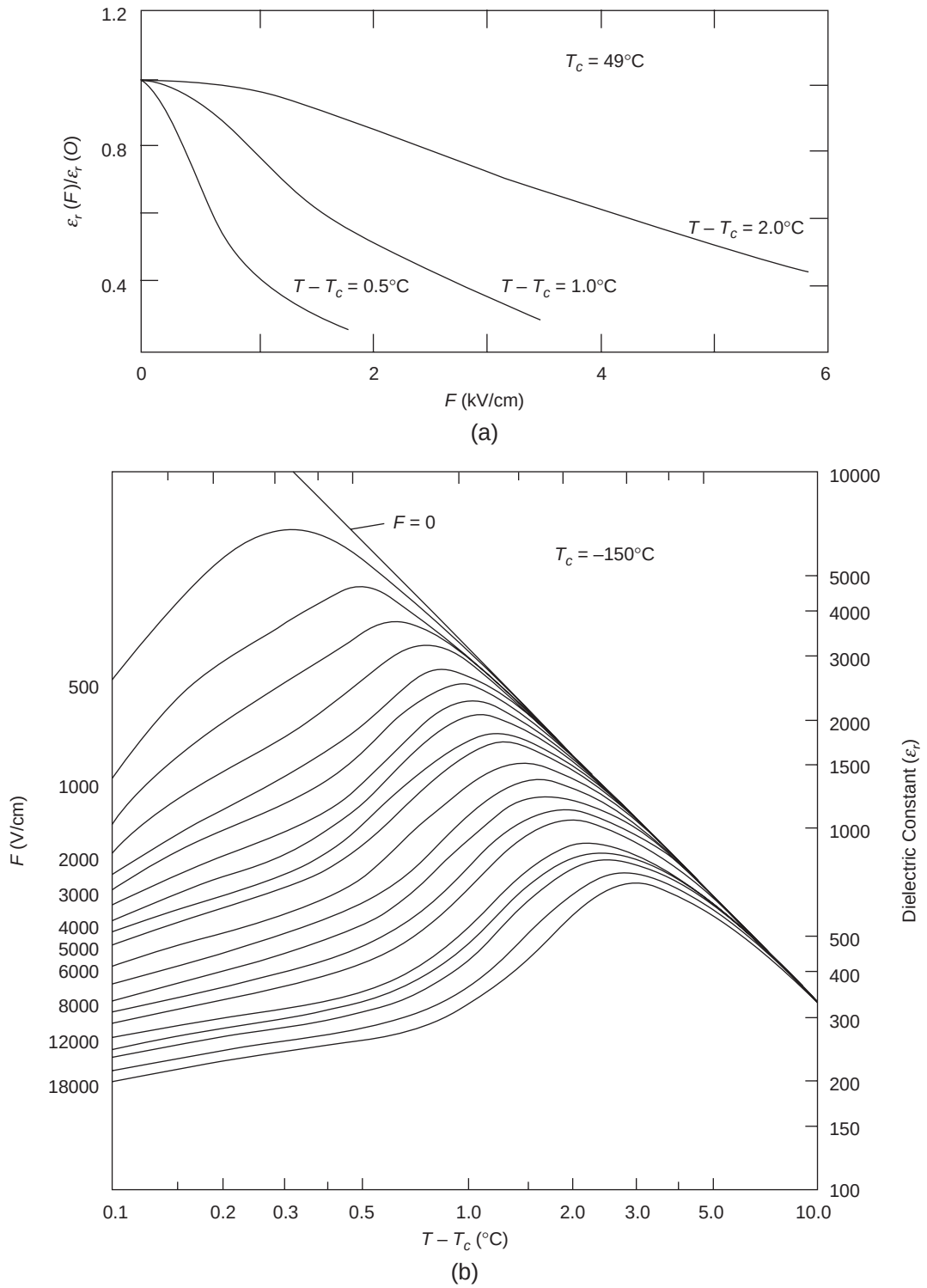


Figure 2-38 The variation of small-signal dielectric constant with DC bias field, with the field applied along the ferroelectric axis at temperatures above the Curie point temperature T_c for (a) triglycine sulphate (TGS) $[(\text{NH}_2\text{CH}_2\text{COOH})_3\text{H}_2\text{SO}_4]$, for which $\epsilon_r(F)$ is normalized to $\epsilon_r(F = 0)$, and (b) potassium dihydrogen phosphate (KDP) $[\text{KH}_2\text{PO}_4]$ plotted as a function of temperature at various DC bias fields.

molecules when the field is further increased. Under this situation, the polarization becomes saturated; this phenomenon is known as dielectric saturation. Several investigators⁶⁸⁻⁷² have observed experimentally that the static dielectric constant of dipolar solution decreases as the applied electric field is increased, and that the amount of the decrease in dielectric constant is approximately proportional to the square of the field strength. It should be noted that in order to observe the dielectric saturation phenomenon, the material used must be in fluid state so that molecules have freedom to orient; in a material in solid state, the molecules are frozen, losing their ability to move. Several investigators have studied this phenomenon theoretically and proposed equations to elucidate quantitatively the behavior of the dielectric constant at high fields. The first was Debye.³¹ Later, van Vleet,^{18,72} Böttcher,⁶ and several others⁷⁴⁻⁷⁹ also put forward their approaches based on their assumptions. All of the theoretical equations are derived on the basis of some assumptions, but none of them is very satisfactory. It is obvious that the tremendous local fields existing around ions and dipoles in the solution must have large effects on the dielectric behavior. An accurate method to calculate such local fields is still not available. We shall not include the discussion of various analyses for this dielectric saturation phenomenon here. The reader interested in this topic is referred to the references cited above. In the following, we shall present a simple analysis with the aim of showing the physical concept of this phenomenon.

In a dielectric solution under a high electric field, the average dipole moment in the direction of the field is given by [see Equations (2.123), (2.129) and (2.130)]

$$\langle u_F \rangle = u_o L(z) \quad (2-305)$$

where $L(z)$ is the Langevin function, which can be expressed in the form of a series as

$$\begin{aligned} L(z) &= \coth z - \frac{1}{z} \\ &= \left[\frac{1}{z} + \frac{z}{3} - \frac{z^3}{45} + \frac{2z^5}{945} - \dots \right] - \frac{1}{z} \\ &= \frac{z}{3} - \frac{z^3}{45} + \frac{2z^5}{945} - \dots \end{aligned} \quad (2-306)$$

in which z is

$$z = \frac{u_o F_{loc}}{kT} \quad (2-307)$$

Here, we must use the local field F_{loc} instead of the applied field F . The local field is always larger than F by a local field correction factor B , so Equation 2-302 can be rewritten as

$$z = \frac{u_o BF}{kT} \quad (2-308)$$

At high fields (HF) for which $z > 1$, we must include higher power terms of $L(z)$. For simplicity, we take only the terms up to z^3 . Thus, the polarization can be written as

$$\begin{aligned} P_{(HF)} &= [(\epsilon_{ro} - 1) + \epsilon_{ro(HF)}] \epsilon_o F \\ &= N \left[\alpha_e BF + \alpha_i BF + \frac{Bu_o^2 F}{3kT} - \frac{u_o}{45} \left(\frac{u_o BF}{kT} \right)^3 \right] \end{aligned} \quad (2-309)$$

Assuming that for the same high field, the dielectric constant remains the same as that obtained at low fields, based on the linear part of the Langevin function shown in Figure 2-17, the polarization can be written as

$$\begin{aligned} P &= [(\epsilon_{ro} - 1) + \epsilon_{ro}] \epsilon_o F \\ &= N \left[\alpha_e BF + \alpha_i BF + \frac{Bu_o^2 F}{3kT} \right] \end{aligned} \quad (2-310)$$

The difference between $P_{(HF)} - P$ in Equation 2-309 and Equation 2-310 gives the difference between $\epsilon_{ro(HF)}$ and ϵ_{ro} due to the effects of high fields. Thus, we have

$$\begin{aligned} \Delta P &= P_{(HF)} - P = [\epsilon_{ro(HF)} - \epsilon_{ro}] \epsilon_o F \\ &= -\frac{Nu_o}{45} \left(\frac{u_o BF}{kT} \right)^3 \end{aligned} \quad (2-311)$$

and the decrease in dielectric constant as

$$\Delta \epsilon_r = -\left(\frac{Nu_o^4 F^2}{45 \epsilon_o k^3 T^3} \right) B^3 \quad (2-312)$$

Depending on the local field correction factor B , $\Delta \epsilon_r$ is proportional to the square of the applied field and inversely proportional to T^3 . If the local field, based on Lorentz's approach, is used, B is given by

$$B = \frac{\epsilon_{rs} + 2}{3} \quad (2-313)$$

If the reaction field, based on Onsager's approach, is used, B is given by

$$B = \frac{3(\epsilon_{rs} - \epsilon_{r\infty})(2\epsilon_{rs} + \epsilon_{r\infty})}{\epsilon_{rs}(\epsilon_{r\infty} + 2)^2} \quad (2-314)$$

The basic difference among the different expressions put forward by other investigators lies mainly in the different expressions for B . However, this simple analysis, though not accurate, does provide a way to understand this phenomenon.

2.7 Dielectric Relaxation Phenomena

Using the time-domain approach, measurements of the growth or the decay of the polarization when a step-function electric field is suddenly applied or removed from the specimen will yield information about the relaxation behavior equivalent to that from the measurements of ϵ_r and ϵ_r' as functions of frequency based on the frequency-domain approach. In the previous section, we discussed the relation of the dielectric constant and absorption, with frequency based on the frequency-domain approach. In the present section, we shall use the time-domain approach to deal with relaxation phenomena. In general, the time-domain response provides conspicuous information about the nonlinearity of the dielectric behavior simply by varying the amplitude of the applied step-function field.

The basic experimental arrangement for the measurements of the time-domain response, (i.e., the transient charging or discharging current, resulting from the application or the removal of a step DC voltage) is shown schematically in Figure 2-39(a).

This arrangement was originally used by Williams.⁸⁰ The circuit is self-explanatory. The switch S_1 has two positions: one for turning on the step DC voltage to start the flow of the charging current, the other for short-circuiting the specimen to allow the discharging current to flow after the specimen has been fully charged to a steady-state level. The switch S_2 is used to short-circuit R_1 to provide a path for surge currents for a very short period of time to

protect the circuit; it also gives a chance to adjust the amplifier to a null position before recording the transient current. It is important to make the time constant of the amplifier, which depends on the stray capacitance in shunt with R_1 , much smaller than the time during which the transient current is flowing. The specimen has the guard and the guarded electrodes, the outer guard electrode being connected to ground to eliminate surface leakage currents from the specimen. The charging or discharging current is measured as a voltage appearing across R_1 by means of a DC amplifier. The voltage drop from point A to ground is made zero by a negative feedback in the amplifier circuit, which produces a voltage across R_2 equal and opposite to that across R_1 , thus making the applied step voltage across the specimen only. The step voltage, and the charging and discharging current as function of time, are schematically shown in Figure 2-39(b), in which I_o is the steady DC component of the charging current and the width of the step voltage is 63 seconds.

2.7.1 The Hamon Approximation

In principle, the equivalent frequency-domain response of a dielectric material can be obtained from a time-domain response by a Fourier transformation which extracts all the separate harmonic components from the charging or the discharging current. For a linear dielectric material in a capacitor of unit area in cross-section and d in thickness, as shown in Figure 2-39(a), the complex dielectric constant can be expressed as

$$\epsilon_r^*(\omega) = \epsilon_{r\infty} + \int_0^{\infty} \phi(t) e^{-j\omega t} dt - j \frac{\sigma_o}{\omega \epsilon_o} \quad (2-315)$$

where σ_o is the steady DC electrical conductivity and $\phi(t)$ is the decay function of the transient current $J(t)$.⁸¹ Thus, $\phi(t)$ can be written as

$$\phi(t) = \frac{J(t)}{\epsilon_o F} \quad (2-316)$$

where F is equal to V/d . Note that $\phi(t)$ does not include the contribution to the observed tran-

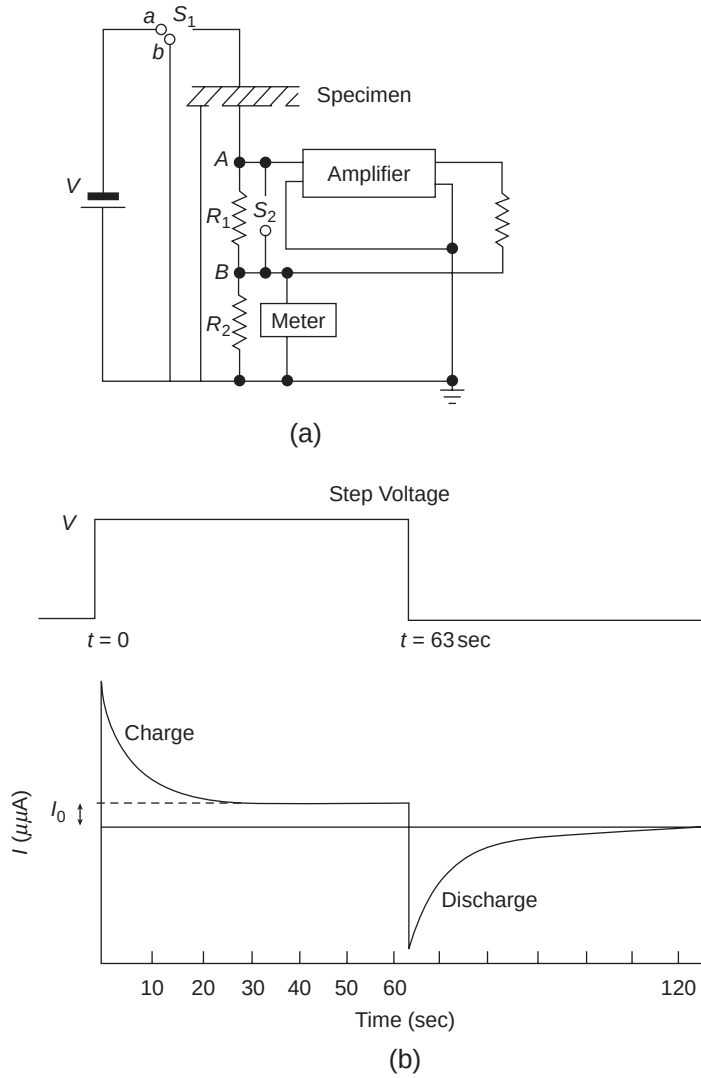


Figure 2-39 Schematic diagrams illustrating (a) the basic experimental arrangement for the measurements of the charging and the discharging current resulting from the application and the removal of a step voltage, and (b) the step voltage and the charging and the discharging current as functions of time. I_0 is the steady DC component of the charging current.

sient current $J(t)$ from the steady DC conductivity of the specimen. By considering only the relaxation component of the transient current, the complex dielectric constant should not include the term $\sigma_o/\omega\epsilon_o$, although it is part of the total loss factor. Thus, we have

$$\epsilon_r^*(\omega) = \epsilon_r(\omega) - j\epsilon_r'(\omega) = \epsilon_{r\infty} + \int_u^\infty \phi(t) e^{-j\omega t} dt \quad (2-317)$$

For a simple relaxation process involving only one single relaxation time based on the Debye theory, the decay function is

$$\phi(t) = A \exp(-t/\tau_o) \quad (2-318)$$

where

$$A = (\epsilon_{rs} - \epsilon_{r\infty})/\tau_o \quad (2-319)$$

However, for relaxation processes involving a distribution of relaxation times, as in poly-

mers, $\phi(t)$ would be a complicated function of time, and an explicit expression for the spectra of $\varepsilon_r(\omega)$ and $\varepsilon'_r(\omega)$ cannot be deduced by analytical integration of Equation 2-317. Of course, if $\phi(t)$ can be determined experimentally over the entire range of time, numerical solution of Equation 2-317, with the aid of a computer, would provide information about $\varepsilon_r(\omega)$ and $\varepsilon'_r(\omega)$.

However, the frequency-domain response for the case with a distribution of relaxation times can be expressed by empirical equations (see Equations 2-282 and 2-286 through 2-288), hinting that the time-domain response can also be expressed by an empirical equation such as

$$J(t) = At^{-r} \quad (2-320)$$

From Equation 2-316, the decay function can be written as

$$\phi(t) = \frac{J(t)}{\varepsilon_0 F} = Bt^{-r} \quad (2-321)$$

in which $B = A/\varepsilon_0 F$. At a fixed constant temperature, the constant A (and hence the constants B and r) can be determined for a specific dielectric material by a direct fitting of Equation 2-320 to the experimental data of the charging or the discharging current. Substitution of Equation 2-321 into Equation 2-317, followed by integration, yields

$$\varepsilon_r(\omega) = \varepsilon_\infty + B\omega^{r-1}\Gamma(1-r)\sin r\pi/2, \quad 0 < r < 1 \quad (2-322)$$

$$\varepsilon'_r(\omega) = B\omega^{r-1}\Gamma(1-r)\cos r\pi/2, \quad 0 < r < 2 \quad (2-323)$$

where $\Gamma(1-r)$ is the gamma function.^{16,39} If B and r are known, $\varepsilon_r(\omega)$ and $\varepsilon'_r(\omega)$ at any desired value of ω can be obtained from Equations 2-322 and 2-323.

To simplify the calculation, Hamon⁸² has proposed an approximation method, generally referred to as the Hamon approximation. Equation 2-323 for $\varepsilon'_r(\omega)$ may be rearranged to the form

$$\begin{aligned} \varepsilon'_r(\omega) &= \frac{Bt^{-r}}{\omega} \left[\frac{\Gamma(1-r)\cos r\pi/2}{(\omega t)^{-r}} \right] \\ &= \frac{J(t)}{\omega \varepsilon_0 F} \left[\frac{\Gamma(1-r)\cos r\pi/2}{(\omega t)^{-r}} \right] \end{aligned} \quad (2-324)$$

The basic criterion for the Hamon approximation is that at a certain value of t in Equation 2-324, $\varepsilon'_r(\omega)$ at any desired value of ω can be obtained simply by the following expression

$$\varepsilon'_r(\omega) = \frac{J(t)}{\omega \varepsilon_0 F} \quad (2-325)$$

This criterion implies that

$$\frac{\Gamma(1-r)\cos r\pi/2}{(\omega t)^{-r}} = 1$$

or

$$\omega t = [\Gamma(1-r)\cos r\pi/2]^{1/r} \quad (2-326)$$

Hamon found that for $0.1 < r < 1.2$, the right-hand side of Equation 2-326 is approximately constant at a value of $\pi/5 = 0.63$ at an accuracy of about 3%. Based on the Hamon approximation, the loss factor can be expressed simply as

$$\varepsilon'_r(\omega) = \frac{J(t = 0.63/\omega)}{\omega \varepsilon_0 F} \quad (2-327)$$

This equation indicates that the observed transient current $J(t)$ at $t = 1$ sec and 10 sec corresponds to the frequency $f = \omega/2\pi = 0.63/2\pi t = 0.1/t = 0.1$ Hz for $t = 1$ sec, and 10^{-2} Hz for $t = 10$ sec. That means that we can obtain $\varepsilon'_r(\omega)$ at $f = 0.1$ Hz and 10^{-2} Hz using the value of $J(t)$ at $t = 1$ sec and at 10 sec, respectively. The dielectric constant $\varepsilon_r(\omega)$, based on the Hamon approximation, can be deduced from Equations 2-322 through 2-324. It is

$$\varepsilon_r(\omega) = \varepsilon_\infty + \varepsilon'_r(\omega)\tan r\pi/2, \quad 0.1 < r < 1 \quad (2-328)$$

The Hamon approximation is not rigorous, since Equation 2-321 does not hold in practice over the entire range of times. Williams⁸³ has shown that the Hamon approximation can be derived more rigorously for low-frequency long relaxation time or high-frequency short relaxation time for dielectric materials following the Cole-Cole empirical relation (see Equation 2-282). Later, Hyde⁸⁴ treated this time-frequency transformation problem based on the data of the time-integral of the transient current, i.e., the charge $Q(t)$. The measurement of the charge on the capacitor plates rather than the

current through the specimen has some advantages in principle. The decay function for $Q(t)$ can be obtained directly, and $Q(t)$ is usually bound at short times, while $J(t)$ involves a much longer time to reach a steady level though the current is much easier to measure. Hyde's approach may be more accurate, but it involves more computations. Using in-line computation, Hyde's method can provide a rapid means of determining $\epsilon_r(\omega)$ and $\epsilon'_r(\omega)$ over the frequency range of 10^{-4} to 10^6 Hz.⁸⁴ The Fourier transformation technique has been extended to obtain very high frequency (10^8 to 10^{10} Hz) data. The experimental method is usually referred to as time-domain spectroscopy. It involves the use of a waveguide and very short-duration pulses with fast rise times, the reflected or the transmitted pulses being observed with a sampling oscilloscope.⁸⁵

However, the Hamon approximation is useful for a rapid appraisal of the dielectric losses and the dielectric constant from the polarization or the depolarization current in practical dielectric materials, such as polymers.

2.7.2 Distribution of Relaxation Times

Dipoles (permanent or induced) are generally present in inorganic, organic, and biological materials. Orientational polarization plays a decisive role in dielectric phenomena in materials. The behavior of a polar molecule in a material is similar to a body of ellipsoidal shape in a viscous fluid. If these bodies in the fluid are not all identical in size, then their orientations will involve more than one relaxation time. Since molecules are generally ellipsoidal in shape, the friction coefficients of the three axes are different; hence, three different relaxation times may exist. Many reasons exist for the relaxation times to be distributed in solids, the most obvious being the presence of inhomogeneities. It is likely that not all dipoles in a solid are situated in the same environment, so some are more free to rotate than others. Even in a single crystal, dipoles may find certain orientations more favorable than others and certain transitions between orientations easier than others. The variation of such local transition probabilities reflects the variation of the activation energy

H for dipole orientation and hence the relaxation time τ , as well as the preexponential factor τ_h based on Equation 2-289. In polymers, the molecules are complex; the orientation of the polar group related to the link segment along a polymer chain may involve many relaxation times. For example, in polyvinyl chloride, the degree of rotation of the polar groups about the C–C axis depends on the position and the angle of the section of the C–C links.

If a system has more than one type of dipole, the relaxation times will be distributed. Supposing that N is the number of dipoles of one type per unit volume, then we can write

$$N = N_o f(\tau) \quad (2-329)$$

where N_o is the total number of dipoles of all different types per unit volume, $f(\tau)$ is the distribution function of the relaxation time τ or the probability density of τ . Thus, $f(\tau)d\tau$ represents the probability density of $f(\tau)$ in the interval between τ and $\tau + d\tau$. So, we can write

$$\int_0^{\infty} f(\tau) d\tau = 1 \quad (2-330)$$

Taking into account the distribution of relaxation times, the Debye equations, (Equations 2-255 through 2-258) become

$$\epsilon_r = \epsilon_{r\infty} + (\epsilon_{rs} - \epsilon_{r\infty}) \int_0^{\infty} \frac{f(\tau) d\tau}{1 + \omega^2 \tau^2} \quad (2-331)$$

$$\epsilon'_r = (\epsilon_{rs} - \epsilon_{r\infty}) \int_0^{\infty} \frac{\omega \tau f(\tau) d\tau}{1 + \omega^2 \tau^2} \quad (2-332)$$

$$\tan \delta = \frac{\epsilon'_r}{\epsilon_r} = \frac{(\epsilon_{rs} - \epsilon_{r\infty}) \int_0^{\infty} \frac{\omega \tau f(\tau) d\tau}{1 + \omega^2 \tau^2}}{\epsilon_{r\infty} + (\epsilon_{rs} - \epsilon_{r\infty}) \int_0^{\infty} \frac{f(\tau) d\tau}{1 + \omega^2 \tau^2}} \quad (2-333)$$

If τ is distributed, then the peak value of ϵ'_r decreases and the spread width of ϵ'_r increases, as shown in Figure 2-40.

The area beneath the curve in Figure 2-40 can be expressed as

$$\begin{aligned} S &= \int_{-\infty}^{\infty} \epsilon'_r d(\ln \omega) = \int_0^{\infty} \epsilon'_r \frac{d\omega}{\omega} \\ &= (\epsilon_{rs} - \epsilon_{r\infty}) \int_0^{\infty} \frac{d\omega}{\omega} \int_0^{\infty} \frac{\omega \tau f(\tau) d\tau}{1 + \omega^2 \tau^2} \end{aligned} \quad (2-334)$$

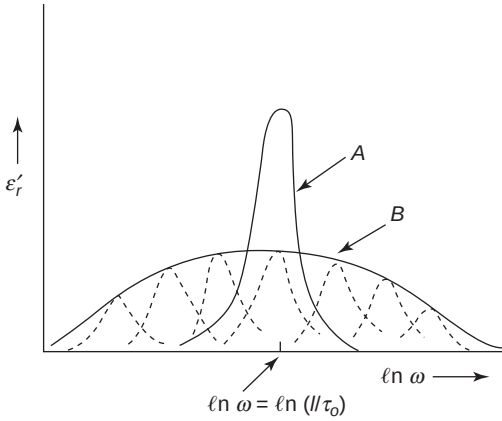


Figure 2-40 The $\epsilon'_r - \ln \omega$ curves (A) involving only one relaxation time, and (B) involving a set of relaxation times, each having its Debye-type loss peak.

By changing the integration order, we have

$$S = (\epsilon_{rs} - \epsilon_{r\infty}) \int_0^\infty \tau f(\tau) d\tau \int_0^\infty \frac{d\omega}{1 + \omega^2 \tau^2} \quad (2-335)$$

Since

$$\int_0^\infty \frac{d\omega}{1 + \omega^2 \tau^2} = \frac{\pi}{2\tau}$$

$$\int_0^\infty f(\tau) d\tau = 1$$

Equation 2-330 reduces to

$$S = \frac{\pi}{2} (\epsilon_{rs} - \epsilon_{r\infty}) \quad (2-336)$$

This equation implies that the area is not related to the distribution function $f(\tau)$. This means that the area beneath the $\epsilon'_r - \ln \omega$ curve for any number of relaxation times is identical irrespective of the dispersion mechanisms of τ and irrespective of $f(\tau)$. The quantity

$$\Delta \epsilon_r = \epsilon_{rs} - \epsilon_{r\infty} = \frac{2}{\pi} \int_{-\infty}^\infty \epsilon'_r d(\ln \omega) = \frac{2S}{\pi} \quad (2-337)$$

is called the relaxation strength, which is an important physical parameter in relaxation processes.

Since $f(\tau)$ is unknown, we cannot determine $\epsilon_r(\omega)$ and $\epsilon'_r(\omega)$ from Equations 2-331 and 2-332. Because the distribution of relaxation

times makes $\epsilon'_{r(\max)}$ decrease, Fuoss and Kirkwood⁵⁶ have proposed a parameter λ with $0 < \lambda < 1$ to modify Equation 2-257 to take into account the effect of the distribution of relaxation times empirically. They have changed Equation 2-257 to the following form:

$$\epsilon'_r = \frac{\lambda(\epsilon_{rs} - \epsilon_{r\infty})}{(\omega\tau)^\lambda + (\omega\tau)^{-\lambda}} \quad (2-338)$$

$$= \frac{1}{2}(\epsilon_{rs} - \epsilon_{r\infty}) \text{Sech}[\lambda \ln(\omega\tau)]$$

When $\omega\tau = 1$, the maximum value of the loss factor, termed $\epsilon'_{r(\max)}$, occurs. It is

$$\epsilon'_{r(\max)} = \lambda(\epsilon_{rs} - \epsilon_{r\infty}) \quad (2-339)$$

Substitution of Equation 2-339 into Equation 2-338 gives the expression exactly the same as Equation 2-288—the Fuoss–Kirkwood equation. When $\lambda = 1$, Equation 2-338 returns to the Debye equation in the case of only one relaxation time. So $\lambda < 1$ implies that the relaxation processes involve more than one relaxation time. Thus, the smaller the value of λ , the higher the number of different relaxation times, the wider the spread of $\epsilon'_r(\omega)$, and hence the smaller the value of $\epsilon'_{r(\max)}$. This implies that the distribution function $f(\tau)$ covers a wide range of frequencies.

To find $f(\tau)$, we can rewrite Equation 2-332 to the form⁸⁶

$$\epsilon'_r = (\epsilon_{rs} - \epsilon_{r\infty}) \sum_{j=1}^n \frac{\omega_i \tau_j f(\tau_j) \Delta \tau_j}{1 + \omega_i^2 \tau_j^2} \quad (2-340)$$

If the summation is represented by a matrix M , then $\epsilon'_r = Mf(\tau)$, where the elements of the matrix are

$$A_{ij} = (\epsilon_{rs} - \epsilon_{r\infty}) \left[\frac{\omega_i \tau_j \Delta \tau_j}{1 + \omega_i^2 \tau_j^2} \right] \quad (2-341)$$

So the distribution function $f(\tau)$ can be expressed in terms of the inverse matrix as

$$f(\tau) = M^{-1} \epsilon'_r \quad (2-342)$$

If ϵ'_r is known, $f(\tau)$ may be determined. However, this method is suitable for the continuous distribution of the relaxation times. Great errors may result if the relaxation process involves only a few, discrete relaxation times.

If all dipoles follow Debye's relaxation model, $f(\tau)$ may be expressed as

$$f(\tau) = \frac{1}{\tau} \exp(-t/\tau) \quad (2-343)$$

If the distribution of the relaxation times assumes a Gaussian type of distribution, then we have

$$f(\tau) = (b/\pi^2\tau) \exp\{-[b\ln(\tau/\tau_o)]^2\} \quad (2-344)$$

where b is a constant and τ_o is the central and the most probable relaxation time. The empirical relation developed by Cole and Cole⁵² corresponds to the distribution function^{52,56}

$$f(\tau) = \frac{\sin \alpha \pi}{2[\cosh(\ln \tau / \tau_o) + \cos \lambda \pi]} \quad (2-345)$$

Based on Equation 2-289, we can write

$$\frac{\tau}{\tau_o} = \exp\left[\frac{H - H_o}{kT}\right] = \exp\left[\frac{\Delta H}{kT}\right] \quad (2-346)$$

The distribution function for activation energies for dipole orientation $G(H)$ can be written as

$$G(H)d(\Delta H) = f(\tau)d(\ln \tau / \tau_o) \quad (2-347)$$

This expression represents the fraction of dipoles having activation energy for orientation between $H - d(\Delta H)/2$ and $H + d(\Delta H)/2$.

A great deal of work has been carried out in studying the distribution functions for the relaxation times and their associated activation energies. However, the molecular processes leading to such various distributions are still not well understood. We still rely on some empirical approaches. It is possible that a collective and cooperative mechanism may be involved.

2.7.3 The Relation between Dielectric Relaxation and Chemical Structure

This is a huge topic, and we shall not include a detailed review of it in this chapter. There are several excellent reviews dealing with this topic, and the reader who is interested in more details is referred to these reviews.^{26,39,87-94}

Here, we shall briefly discuss the relation between dielectric relaxation and chemical structure, taking polymers as an example.

In polymers, there are mainly two types of dielectric relaxation: dipolar segmental relaxation and dipolar group relaxation. The dielectric relaxation processes of these two types exhibit loss peaks on the curves of the temperature or frequency dependence of ϵ_r' . In amorphous polymers, the most intensive peaks of ϵ_r' or $\tan \delta$ occur in the region of the transition from the glassy to the rubbery (high elastic) state, i.e., near the glass transition temperature T_g . Dielectric losses caused by dipolar segmental relaxation are associated with the micro-Brownian motion of segments in polymeric chains, while dielectric losses caused by dipolar group relaxation are associated with the localized movement of molecules. Several distinct dielectric relaxation processes are usually present in solid polymeric materials. They can be observed by means of the thermally stimulated relaxation technique.⁹⁴ As the temperature is raised, the mobility of various types of molecules becomes successively enhanced, making it easier for dipolar orientation to occur. Usually, dielectric relaxation processes are labeled by α , β , γ , δ , and so on, beginning at the high-temperature end.

Figure 2-41 illustrates schematically the α , β , and γ peaks on the curve of the loss factor ϵ_r' as a function of temperature at a fixed frequency. In this convention, the dipolar segmental relaxation corresponds to the α relaxation, and the dipolar group relaxation processes correspond to the β , γ , δ , relaxations. The high-temperature α relaxation is mainly

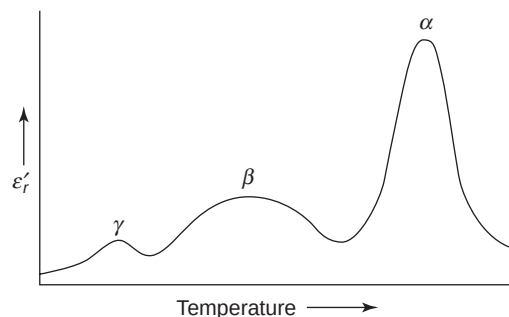


Figure 2-41 Schematic illustration of the loss factor ϵ_r' as a function of temperature showing the α , β , and γ loss peaks.

associated with the micro-Brownian motion of whole chain (segmental movement). The dielectric loss peak in the region of β relaxation is mainly due to the movement of side groups or small units of main chains. The relative strength of the α and β dielectric relaxation depends on the degree of the orientation of dipolar groups through the limited mobility allowed by the β process before the more difficult, but more extensive, mobility for the α process comes into play. There is a partitioning of the total dipolar alignment among the molecular rearrangement processes.

The mechanisms for the β process may depend on the nature of the dipolar group concerned and its position on the polymeric chain. Among the most important mechanisms are the following^{91,92}:

- Relaxation of a side group about a C–C chain,
- Conformational flip of a cyclic unit involving the transition from one chain form to another, altering the orientation of a polar substituent,
- Local motion of a segment of the main chain, since the small segment of a $(\text{CH}_2)_n$ chain can have such motion without involving the rest of the chain.

At lower temperatures, there is a γ relaxation peak due mainly to the movement of small kinetic units of the main or the side chains, which consist of a sequence of several carbon atoms, such as the movement of several CH_2 or CF_2 groups. Another possible mechanism for the γ relaxation is due to the crankshaft rotation⁹⁵ below the glass transition temperature. The γ relaxation process has been observed in polyethylene, aliphatic polyamides, polyester, and some polymethacrylates containing linear methylene chains in side branches.⁹⁵ The presence in the polymeric chain of runs of three or more CH_2 units, each of which is linked to immobile groups, would lead to the crankshaft rotation, thus resulting in γ relaxation in various polymers.^{93,96} However, the available experimental evidence indicates that γ relaxation due to crankshaft rotation is possible only in amor-

phous polymers, or in amorphous regions of crystalline polymers, because crankshaft rotation can only occur about two collinear bonds. This condition is not fulfilled in crystalline regions, where the sequences of methylene groups form mainly transconformations. Relaxation processes at very low temperatures have a quantum nature.⁹³

To study the dielectric relaxation processes, apart from the measurement of the thermally stimulated relaxation as a function of temperature, we need also the data from isothermal scans of the dielectric constant ϵ_r and the loss factor ϵ'' as functions of frequency, so that the effective dipole moments and activation energies may be obtained. Typical results from such measurements for polyvinyl chloride in the α relaxation region are shown in Figure 2-42. The results are from Ishida.⁹⁷ Such plots are sometimes referred to as dielectric spectra. From a series of such plots, the relaxation times can be determined for the individual relaxation processes as functions of temperature.

The α relaxation peak at $T = T_g$ of an amorphous polymer is much narrower than the β relaxation peak. Also, the temperature dependence of the α process is much steeper than that of the β process, indicating that a greater thermal activation energy is required for the motion. The α relaxation process near T_g is largely dependent on free volume. Molecular structure greatly affects the glass transition temperature and hence the dielectric relaxation times. Thus, a bulky side group would cause the T_g to decrease, thus preventing the chains from packing together tightly. T_g can also be deliberately reduced by doping a plasticizer into the polymer, such as diphenyl doped into polyvinyl chloride, which affects the α relaxation process because of the decrease of T_g .⁹⁸

The effect of the degree of crystallinity on the dielectric properties of polymers has been extensively studied. The so-called crystalline polymers always consist of many fully crystalline and fully amorphous regions. So, the overall dielectric relaxation processes may be regarded as a superposition of the relaxation processes in fully crystalline and in fully amorphous regions in the polymer. Usually, the

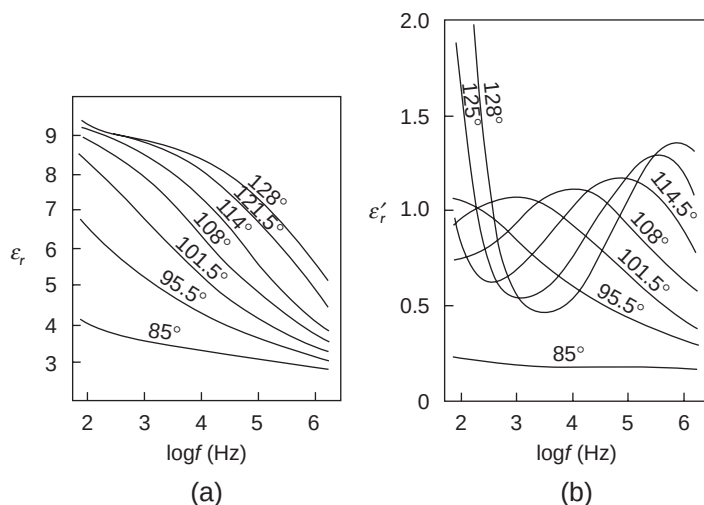


Figure 2-42 (a) The dielectric constant ϵ_r and (b) the loss factor ϵ'_r as functions of frequency at temperatures in the α relaxation region for polyvinyl chloride.

amorphous regions make the major contribution to both ϵ_r and ϵ'_r . As a result, ϵ_r and ϵ'_r will decrease with an increasing degree of crystallinity. Such a dependence of the dielectric properties on the degree of crystallinity is usually observed at temperatures higher than T_g of the amorphous regions. When the temperature is lowered further to the level at which τ becomes so large that $\omega\tau \gg 1$, then the dielectric constant becomes dominated by $\epsilon_{r\infty}$. Since the value of $\epsilon_{r\infty}$ for a number of nonpolar (e.g., polyethylene and polytetrafluoroethylene) and polar (e.g., polyamids) polymers differs little, at very low temperatures, the dielectric constants of these polymers are almost identical. Thus, when the relaxation spectra are frozen at low temperatures, the difference between nonpolar and polar polymers will disappear—at least, as far as the value of ϵ_r is concerned. This phenomenon has been experimentally observed.⁹³

Polymers may exist with linear, branched, and cross-linked chain structures. Linear polymers usually have long chains composed of a large number of repeating units, such as polyethylene and polytetrafluoroethylene. Branched polymers have side chains attached to the main chain, such as polymethyl methacrylate. Cross-linked polymers have a three-dimensional

space network formed by chemical bonds. They are sometimes called the network polymers. There are many ways to cause the formation of a cross-linked network. Polyethylene undergoes cross-linking under the action of ionizing radiation. Cross-linking can greatly restrict certain kinds of molecular movement and hence can affect the dielectric relaxation processes. In general, dielectric losses will diminish with an increasing degree of cross-linking. This tendency appears more prominently in the region of the transition from the glassy to the rubbery state. Thus, as the degree of cross-linking increases, the magnitude of the principal α relaxation peak decreases.

Obviously, the incorporation of any impurities into the polymer in order to modify the dielectric properties will also affect the relaxation processes and hence the values of ϵ_r and ϵ'_r . Examples include plasticizer to lower the glass transition temperature and polar impurities to increase the dielectric relaxation for nonpolar polymers.

References

1. L. Brillouin, *Wave Propagation in Periodic Structures*, (Dover, New York 1953).
2. J.A. Stratton, *Electromagnetic Theory*, (McGraw-Hill, New York, 1941), p. 113.

3. K.C. Kao, *Brit. J. Appl. Phys.*, **12**, 629 (1961).
4. W.R. Smythe, *Static and Dynamic Electricity*, (McGraw-Hill, New York, 1950).
5. J.H. Jeans, *The Mathematical Theory of Electricity and Magnetism*, (Cambridge University Press, Cambridge, 1925).
6. C.J.F. Böttcher, *Theory of Electric Polarization*, (Elsevier, Amsterdam, 1978).
7. R. Cooper and A.A. Wallace, *proc. Phys. Soc.*, **B66**, 1113 (1953).
8. K.H. Stark and C.G. Garton, *Nature*, **176**, 1225 (1955).
9. K.C. Kao, "The Bubble Theory for Electrical Breakdown of Liquid Dielectrics," A.I.E.E. Conference Paper No. CP 60-84 (American Institute of Electrical Engineers, New York, January 1960).
10. K.C. Kao, *Nature*, **208**, 279 (1965).
11. C.G. Garton and Z. Krasucki, *Proc. Royal Soc. (London)*, **208A**, 211 (1964).
12. K.C. Kao and J.B. Higham, *J. Electrochem. Soc.*, **108**, 522 (1961).
13. H.A. Pohl, *Dielectrophoresis*, (Cambridge University Press, Cambridge, (1978), and *J. Appl. Phys.*, **29**, 1182 (1958).
14. R.H. Cole, "Theories of Dielectric Polarization and Relaxation", in *Progress in Dielectrics*, Vol. 3, (Heywood, London, 1961) pp. 47-100.
15. R.E. Teachout and R.T. Pack, *Atomic Data*, **3**, 195 (1971).
16. A.K. Jonscher, *Dielectric Relaxation in Solids*, (Chelsea Dielectric Press, London, 1983).
17. R. Coelho, *Physics of Dielectrics*, (Elsevier, Amsterdam, 1979).
18. J.H. van Vleck, *Theory of Electric and Magnetic Polarizabilities*, (Oxford University Press, Oxford, 1932), p. 203.
19. N.F. Mott and I.N. Sneddon, *Wave Mechanics and Its Applications*, (Clarendon, Oxford, 1948), p. 166.
20. A.J. Dekker, *Electrical Engineering Materials*, (Prentice Hall, Englewood Cliffs, NJ, 1961).
21. A.J. Dekker, *Solid State Physics*, (MacMillan, London, 1963).
22. I. Pirenne and E. Kartheuser, *Physica*, **20**, 2005 (1964).
23. M. Born and K. Huang, *Dynamical Theory of Crystal Lattices*, (Clarendon, Oxford, 1954).
24. B.K.P. Scaife, *Principles of Dielectrics*, (Clarendon, Oxford, 1984).
25. R.J.W. LeFevre, *Dipole Moments*, (Methuen, London, 1953).
26. C.P. Smyth, *Dielectric Behavior and Structure*, (McGraw-Hill, New York, 1955).
27. A.R. Von Hippel, *Dielectrics and Waves*, (John Wiley and Sons, New York, 1954).
28. H. Frohlich, *Theory of Dielectrics*, (Clarendon, Oxford, 1958).
29. P. Langevin, *J. Physique*, **4**, 678 (1905); *Ann. Chim. Phys.*, **5**, 70 (1905).
30. P. Debye, *Physik Zeit.*, **13**, 97 (1912).
31. P. Debye, *Polar Molecules*, (Dover, New York, 1945).
32. C.P. Smyth and C.S. Hitchcock, *J. Amer. Chem. Soc.*, **55**, 1290 (1933).
33. C.P. Smyth and C.S. Hitchcock, *J. Amer. Chem. Soc.*, **56**, 1084 (1934).
34. N.E. Hill, "Theoretical Treatment of Permittivity and Loss" in *Dielectric Properties and Molecular Behaviour*, Ed. N.E. Hill, W.E. Vaughan, A.H. Price, and M. Davies (Van Norstrand Reinhold, London, 1969).
35. J.C. Slater, *Insulators, Semiconductors and Metals*, (McGraw-Hill, New York, 1955).
36. F.C. Nix and W. Shockley, *Rev. Mod. Phys.*, **10**, 1 (1938).
37. N.F. Mott and E.A. Davis, *Electronic Processes in Non-Crystalline Materials*, (Clarendon, Oxford, 1974).
38. L. Onsager, *J. Am. Chem. Soc.*, **58**, 1486 (1936).
39. N.G. McCrum, B.E. Read, and G. Williams, *Anelastic and Dielectric Effects in Polymeric Solids*, (John Wiley and Sons, New York, 1967).
40. J.G. Kirkwood, *J. Chem. Phys.*, **7**, 911 (1939).
41. F.E. Harris and B.J. Adler, *J. Chem. Phys.*, **21**, 1031 (1952).
42. H. Frohlich, *Physica*, **22**, 898 (1956).
43. J.J. O'Dwyer, *J. Chem. Phys.*, **26**, 878 (1957).
44. B.K.P. Scaife, *Proc. Phys. Soc. (London)*, **B70**, 314 (1957).
45. A.D. Buckingham, *Proc. Roy. Soc. (London)*, **A238**, 235 (1956); also, *Trans. Faraday Soc.*, **52**, 1035 and 1551 (1956).
46. R.H. Cole, *J. Chem. Phys.*, **27**, 33 (1957).
47. G.P. Mikhailov, in *Physics of Non-Crystalline Solids*, (North Holland, Amsterdam, 1965), p. 270.
48. H.A. Kramers, *Atti del Congresso Internazionali dei Fisici*, Como **2**, 545 (1927), *Collected Sci. Papers*, (North Holland, Amsterdam, 1956), p. 333.
49. R. de L. Kronig, *J. Optical Society of America*, **12**, 547 (1926).
50. P. Debye and W. Ramm, *Annalen der Physik*, **28**, 28 (1937).
51. P. Debye, *The Collected Papers of Peter J.W. Debye*, (Wiley Interscience, New York, 1954).

52. R.H. Cole and K.S. Cole, J. Chem. Phys., 9, 341 (1941); *ibid.*, 10, 98 (1942).
53. D.W. Davidson and R.H. Cole, J. Chem. Phys., 18, 1417 (1950); *ibid.*, 19, 1484 (1951).
54. S. Havriliak and S. Negami, J. Polymer Sci., Pt. C. 14, 99 (1966); also, Polymers, 8, 161 (1967).
55. A.M. Rad and P. Bordewijk, Rec. Trav. Chim., 90, 1055 (1971).
56. R.M. Fuoss and J.G. Kirkwood, J. Amer. Chem. Soc., 63, 385 (1941).
57. R.H. Cole and P.M. Cross, Rev. Sci. Instr., 20, 252 (1949).
58. R.H. Cole, J. Chem. Phys., 23, 493 (1955).
59. B.E. Read and G. Williams, Polymers, 2, 239 (1961); also, Trans. Faraday Soc., 57, 1979 (1961).
60. A. Schallamach, Trans. Faraday Soc., 42A, 495 (1946).
61. C. Herring, Bell System Tech. J., 34, 237 (1955).
62. E.M. Conwell, *High Field Transport in Semiconductors*, (Academic Press, New York, 1967).
63. E.M. Conwell, V.J. Fowler, and J. Zucker, *Report on Contract AF19 (604)-5714* (Air Force Cambridge Res. Center, Cambridge, Massachusetts, 1960).
64. A.F. Gibson, J.W. Granville, and E.G.S. Paige, J. Phys. Chem. Solids, 19, 198 (1961).
65. S. Triebwasser, IBM J. Res. Develop., 2, 212 (1958); also, Phys. Rev., 118, 100 (1960).
66. H. Baumgartner, Helv. Phys. Acta, 23, 651 (1950); *ibid.*, 24, 326 (1951).
67. E. Fatusso and W.J. Merz, *Ferroelectricity*, (North Holland, Amsterdam, 1967).
68. S. Ratnowsky, *Vorhandl deut Physik. Ges.*, 15, 497 (1913).
69. J. Herweg, Z. Physik, 3, 36 (1920).
70. J. Herweg and W. Potzsch, Z. Physik, 8, 1 (1922).
71. J. Malsch, Physik Z., 29, 710 (1928).
72. G.P. Jones and M. Gregson, *Total Dielectric Saturation Observed in a Dipolar Solution*, Chem. Phys. Lett., 4, 13 (1969).
73. J.H. van Vleck, J. Chem. Phys., 5, 556 (1937).
74. A.D. Buckingham, J. Chem. Phys., 25, 428 (1956).
75. J.J. O'Dwyer, Proc. Phys. Soc. (London), A64, 1125 (1951).
76. A. Anselm, Acta Physicochim, U.S.S.R., 19, 400 (1944).
77. J.B. Hasted, D.M. Ritson, and C.H. Collie, J. Chem. Phys., 16, 11 (1948).
78. S. Lielich, in *Dielectric and Related Molecular Processes*, Vol. 1, (The Chemical Society, Burlington House, London, 1972), (1972).
79. G.P. Jones, *ibid.*, Vol. 2 (1975).
80. G. Williams, Polymers, 4, 27 (1963).
81. M.F. Manning and M.E. Bell, Rev. Mod. Phys., 12, 215 (1940).
82. B.V. Hamon, Proc. IEE, (London), 99, 151 (1952), and also Austral. J. Phys., 6, 304 (1953).
83. G. Williams, Trans. Faraday Soc., 58, 1041 (1962); *ibid.*, 59, 1397 (1963).
84. P.J. Hyde, Proc. IEE. (London), 117, 1891 (1970).
85. A. Suggett, in *Dielectric and Related Molecular Processes*, Vol. 1 (Chemical Society, London, 1972), p. 100.
86. D.R. Uhlmann and R.M. Hakim, J. Phys. Chem. Solids, 32, 2652 (1971).
87. N.E. Hill, W.E. Vaughan, A.H. Price, and M. Davies (Eds.), *Dielectric Properties and Molecular Behaviour*, (Van Nostrand Reinhold, London, 1969).
88. A.M. North, Chem. Soc. Rev., 1, 49 (1972).
89. P. Hedvig, *Dielectric Spectroscopy of Polymers*, (Adam Hilger, Bristol, 1977).
90. P.E. Karasz (Ed.), *Dielectric Properties of Polymers*, (Plenum, New York, 1972).
91. A.R. Blythe, *Electrical Properties of Polymers*, (Cambridge University Press, Cambridge, 1979).
92. I.I. Perepechko, *An Introduction to Polymer Physics*, (Mir Publisher, Moscow, 1981), English Translation by A. Beknazarov).
93. I.I. Perepechko, *Low Temperature Properties of Polymers*, (Mir Publisher, Moscow, and Pergamon Press, Oxford, 1980).
94. P. Bräunlich, *Thermally Stimulated Relaxation in Solids*, (Springer-Verlag, Berlin, 1979).
95. T.F. Schatzki, J. Poly. Sci., 57, 496 (1962); also, Meeting of Am. Chem. Soc. Div. Polymer Chem., Atlantic City, September 1965, Polymer Preprints, 6, 646 (1965).
96. B.J. Wunderlich, J. Chem. Phys., 7, 10 and 2429 (1962).
97. Y. Ishida, Kolloid. Z., 168, 29 (1960).
98. R.M. Fuoss, J. Am. Chem. Soc., 63, 378 (1941).

3 Optical and Electro-Optic Processes

We can conceive darkness without thought of light, but we cannot conceive light without thought of darkness.

Christopher Morley

Optical and electro-optic processes are mainly associated with the interaction of light with matter. Light is simply energy and energy is intangible, being apprehended only on its interaction with matter, which acts as an origin or as a detector of energy. In this chapter, we shall discuss such interaction processes.

3.1 Nature of Light

Optical radiation is electromagnetic radiation covering a wide range of wavelengths, as was shown in Figure 1-21. As mentioned in *Electromagnetic Waves and Fields* in Chapter 1, light can be described by corpuscular theory or by wave theory. The two theories are not in conflict but complementary.

According to quantum mechanics, radiation is absorbed or emitted as particles, called photons, having discrete values of energy in quanta. Thus, photon energy can be written as

$$E = h\nu = \frac{hc}{\lambda} \quad (3-1)$$

where h is the Planck constant, c is the speed of light, and ν and λ are, respectively, the frequency and the wavelength. Energy is itself intangible; it can be apprehended only through its interaction with matter. The interaction of light with solids may produce diffraction or photoelectricity, leading to the concept of radiant energy, which can be described in terms of photons or quanta. Light, like matter, is composed of infinitesimal particles whose probable behaviors are describable by wave equations similar to those used for describing other waves. So, light is energy and apparently has a dual nature.

3.1.1 Corpuscular Theory

The Outstanding Differences between Photons and Electrons or Protons

There are several significant differences between photons and electrons or protons.¹

- Photons may be created and annihilated, whereas electrons or protons are conserved.
- Photons do not interact with each other, and they obey Bose–Einstein statistics; electrons or protons do interact with each other and obey Fermi–Dirac statistics.
- Photons do not have electrostatic charges, spin moments, or rest mass; these are possessed by electrons or protons.
- All photons have a common constant velocity c in free space (constant velocity $v = c/n$ in materials), whereas the velocities of electrons or protons are variable, depending on the accelerating voltage.
- Photons have diffraction wavelength λ_d equal to their radiation (conversion) wavelength λ_e , whereas electrons or protons have $\lambda_d \propto V^{-1/2}$ but $\lambda_e \propto V^{-1}$, where V is the accelerating voltage.
- Photons have momentum p and kinetic energy E_k , depending on their frequency, whereas electrons or protons have momentum and kinetic energy, depending on their velocity.

Frequency, Energy, and Momentum of Photons

On the basis of Einstein’s relativistic mechanics, the relation between mass m and energy E , and momentum p of a particle can be expressed as

$$E = mc^2 = \frac{m_0 c^2}{(1 - v^2/c^2)^{1/2}} \quad (3-2)$$

$$p = mv = \frac{m_0 v}{(1 - v^2/c^2)^{1/2}} \quad (3-3)$$

in which m_0 is the rest mass of the particle, and m is the mass of the particle when it is moving at a velocity v . If $v \ll c$, Equation 3-2 can be simplified to

$$E = m_0 c^2 + \frac{1}{2} m_0 v^2 = E_0 + \frac{1}{2} m_0 v^2 \quad (3-4)$$

of which $E_0 = mc^2$ is the rest mass energy and $\frac{1}{2} m_0 v^2$ is the kinetic energy of the particle.

For photons, we have the following relation based on Equations 3-2 through 3-4:

$$E^2 = m_0^2 c^2 + c^2 p^2 \quad (3-5)$$

Since the photon has no mass, so $m_0 = 0$, the momentum of a photon is

$$E = cp$$

or

$$p = \frac{E}{c} = \frac{h}{\lambda} \quad (3-6)$$

It can be imagined that a particle like a photon, whose rest mass is extremely small ($m \rightarrow 0$) and whose velocity is extremely large ($v \rightarrow c$), will have a finite momentum and energy.

3.1.2 Wave Theory

Light as an electromagnetic wave can be characterized by Maxwell's wave equation

$$\nabla^2 F = \frac{1}{c^2} \frac{\partial^2 F}{\partial t^2} \quad (3-7)$$

$$\nabla^2 H = \frac{1}{c^2} \frac{\partial^2 H}{\partial t^2} \quad (3-8)$$

where F and H are, respectively, the electric and magnetic fields. F and H are transverse waves, as shown in Figure 1-20. In free space, the waves propagate at a speed of light c . In other material media, the propagation speed is given by

$$v = \frac{c}{n} \quad (3-9)$$

where n is the refractive index of the medium, which is given by

$$n = (u_r \epsilon_r)^{1/2} \quad (3-10)$$

and

$$u_r = u/u_0 \\ \epsilon_r = \epsilon/\epsilon_0$$

where ϵ_0 and ϵ are, respectively, the permittivities of free space and the medium, and u_0 and u are, respectively, the permeabilities of free space and the medium.

In describing optical phenomena, we usually deal with the electric field vector, because the magnetic field vector behaves in the same way as the electric field vector. In a one-dimensional case, Equation 3-7 can be written as

$$\frac{\partial^2 F}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 F}{\partial t^2} \quad (3-11)$$

The solution of the above equation is

$$F(x, t) = F_0 \cos(\omega t - kx + \phi) \quad (3-12)$$

where

$k = \text{wave vector} = \frac{2\pi}{\lambda}$ (the wave number usually used is λ^{-1})

$\phi = \text{phase}$

$T = \text{one period of the time for one cycle} = \frac{1}{\nu} = \frac{2\pi}{\omega}$

Figure 3-1 illustrates the variation of F with the traveling distance x and the time t . Wavefronts or wavesurfaces are the surfaces with a constant phase, implying that on the wavefronts, F is a constant magnitude. This means

$$\omega t - kx + \phi = \text{constant} \quad (3-13)$$

This leads to

$$\frac{dx}{dt} = \frac{\omega}{k} = v\lambda = c \quad (3-14)$$

which is called the phase velocity.

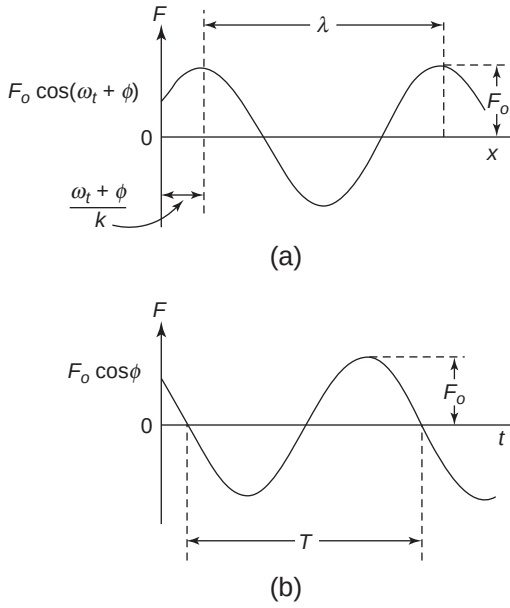


Figure 3-1 The variation of the electric field vector F with distance in x -direction and time t . (a) F as a function of x at a fixed value of t (say $t = 0$) and (b) F as a function of t at a fixed value of x (say $x = 0$).

Principle of Superposition

According to the principle of superposition, two or more waves can be added to form a resultant wave. For example, by adding two waves of the same frequency, but with different amplitudes and phases, such as

$$F_1 = F_{10} \sin(\omega t - kx + \phi_1) \quad (3-15)$$

$$F_2 = F_{20} \sin(\omega t - kx + \phi_2) \quad (3-16)$$

the resultant wave becomes

$$\begin{aligned} F &= F_1 + F_2 \\ &= (F_{10} \cos \phi_1 + F_{20} \cos \phi_2) \sin(\omega t - kx) \\ &\quad + (F_{10} \sin \phi_1 + F_{20} \sin \phi_2) \cos(\omega t - kx) \end{aligned} \quad (3-17)$$

which can be written in the form

$$F = F_o \sin(\omega t - kx + \phi) \quad (3-18)$$

with

$$F_o^2 = F_{10}^2 + F_{20}^2 + 2F_{10}F_{20} \cos(\phi_1 - \phi_2) \quad (3-19)$$

$$\tan \phi = \frac{F_{10} \sin \phi_1 + F_{20} \sin \phi_2}{F_{10} \cos \phi_1 + F_{20} \cos \phi_2} \quad (3-20)$$

Equation 3-18 implies that the resultant of two sinusoidal waves of the same frequency is itself a sinusoidal wave with the same frequency as the original waves.

In general, it is not possible in practice to produce perfectly monochromatic light. Usually, light consists of a group of waves of closely similar wavelength, moving in the form of a packet. To illustrate the concept of the wave packet, we use two waves, for simplicity, to demonstrate how the two waves form a packet. Let the two waves be

$$F_1 = F_o \cos(\omega t - kx) \quad (3-21)$$

$$F_2 = F_o \cos[(\omega + \delta\omega)t - (k + \delta k)x] \quad (3-22)$$

Based on the principle of superposition, these two waves can be added to form a resultant wave. This resultant wave is

$$\begin{aligned} F &= F_1 + F_2 \\ &= 2F_o \cos[(\omega + \delta\omega/2)t - (k + \delta k/2)x] \cos(\delta\omega t/2 - \delta kx/2) \end{aligned} \quad (3-23)$$

As we assume $\delta\omega \ll \omega$ and $\delta k \ll k$, Equation 3-23 can be simplified to

$$\begin{aligned} F &= 2F_o \cos(\delta\omega t/2 - \delta kx/2) \cos(\omega t - kx) \\ &= F'_o \cos(\omega t - kx) \end{aligned} \quad (3-24)$$

This is the wave of the same form as the original waves, but it has the amplitude

$$F'_o = 2F_o \cos(\delta\omega t/2 - \delta kx/2) \quad (3-25)$$

It is also a traveling wave. However, it can be seen from Equation 3-25 that there are maxima (antinodes) occurring at $\delta\omega t/2 - \delta kx/2 = m\pi$ with $m = 0, 1, 2, 3, \dots$ and minima (nodes) at

$$\delta\omega t/2 - \delta kx/2 = m' \frac{\pi}{2} \text{ with } m' = 1, 3, 5, 7, \dots$$

These maxima and minima are generally referred to as the beats. The propagation velocity of a specific beat is the signal velocity or group velocity v_g of the modulated wave. Thus, by setting $\delta\omega t/2 - \delta kx/2 = \text{constant}$ (the plane of constant amplitude), we obtain

$$v_g = \frac{dx}{dt} = \frac{d\omega}{dk} \quad (3-26)$$

Figure 3-2 illustrates this phenomenon of beats. From Equations 3-9 and 3-14, using v instead of c for phase velocity for general cases because in solids $v = c/n$, we can write

$$\omega = kv\lambda = kv \quad (3-27)$$

and obtain the group velocity in the form

$$v_g = \frac{d(kv)}{dk} = v - \lambda \frac{dv}{d\lambda} \quad (3-28)$$

Group velocity and phase velocity are identical only in media in which the propagation velocity is independent of frequency, that is, the wave propagation in nondispersive dielectric media.

Note that in most cases, the light source is a point source of light radiating in all directions. The wavefronts are then a series of concentric spherical shells. So, the electric field vector is given by

$$F = \frac{A}{r} \cos(\omega t - kr) \quad (3-29)$$

where A is the strength of the light source and r is the distance from the light source. Equation 3-29 implies that the amplitude of the electric field vector F is proportional to r^{-1} . Human eyes

and most light detectors are not sensitive to electric field or magnetic field vectors; they are sensitive to the power density of light, which is the flow of energy per unit time per unit area and is generally referred to as the irradiance I . I is, in fact, the Poynting vector. For the present case, the irradiance I_r can be written as

$$I_r = FH \propto F^2 \propto \frac{1}{r^2} \quad (3-30)$$

Interference and Interferometry

When two or more waves with different amplitudes and phases are mixed, they will interfere with each other. For example, when two waves, such as those given by Equations 3-15 and 3-16, are mixed, the resultant wave is given by Equation 3-18, with F_o given by Equation 3-19. In this case, the resultant irradiance can be written as

$$I = F_o^2 = F_{10}^2 + F_{20}^2 + 2F_{10}F_{20} \cos(\phi_1 - \phi_2) \quad (3-31)$$

If $E_{10} = E_{20}$, then I becomes

$$\begin{aligned} I &= 2F_{10}^2 [1 + \cos(\phi_1 - \phi_2)] \\ &= 4F_{10}^2 \cos^2[(\phi_1 - \phi_2)/2] \end{aligned} \quad (3-32)$$

It can be seen that as $(\phi_1 - \phi_2)/2$ varies from 0 to π , I will vary from $4F_{10}^2$ to zero and then from

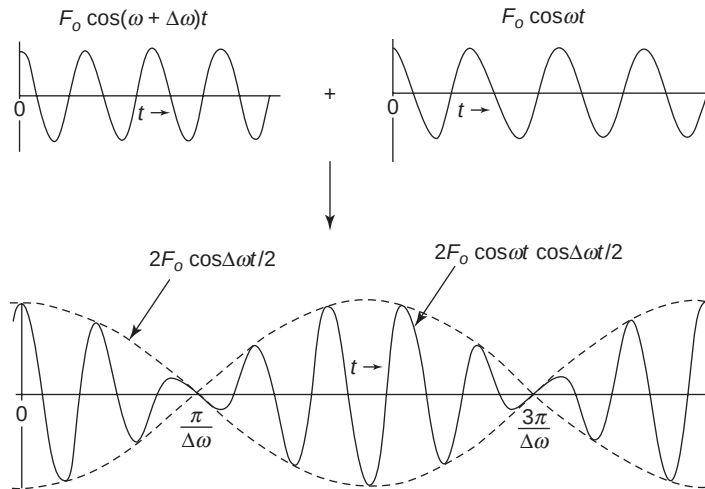


Figure 3-2 Schematic diagrams showing the beats produced by two waves of slightly different frequencies for $x = 0$.

zero to $4F_{10}^2$. If $F_{10} \neq F_{20}$, then I varies between $(F_{10} + F_{20})^2$ and $(F_{10} - F_{20})^2$, as $(\phi_1 - \phi_2)$ varies between 0 and π .

In general, interference produces the so-called interference fringes, which are the variation of the resultant amplitude and hence the relative energy and illumination (or brightness) from point to point in the field of view. They are mainly caused by the different paths between the interfering beams of light. The conditions for the occurrence of interference are as follows:

- The sources of light must be coherent. This means that the waves emanating from the sources, while they need not be exactly in phase, must keep their phase relationship unaltered during the period of interference.
- The amplitudes of two wavefronts must be similar.
- The wavelengths of the light from the sources must be the same.
- The ways for obtaining interference in accordance with the above conditions can be divided into two groups: division of wavefront and division of amplitude.^{2,3}

Division of Wavefront

The earliest experiment about interference caused by the division of wavefront was performed by Thomas Young in 1802.⁴ The experiment basically involves two slits S_1 and S_2 with a separation d between them, and light from a monochromatic source A illuminating a narrow slit S , which then acts as a light source to illuminate simultaneously S_1 and S_2 , as shown in Figure 3-3. Then the two narrow beams of light of different cylindrical wavefronts from S_1 and S_2 will fall upon the screen B . Generally, the narrower the slit as compared with d , the greater the illuminated area, because of diffraction at the slits. The distances D_1 and D_2 determine the phase difference $\phi_1 - \phi_2$ between the two waves. Thus, the phase difference between the two waves on the screen can be written as

$$\phi_1 - \phi_2 = \frac{2\pi}{\lambda}(D_1 - D_2) \quad (3-33)$$

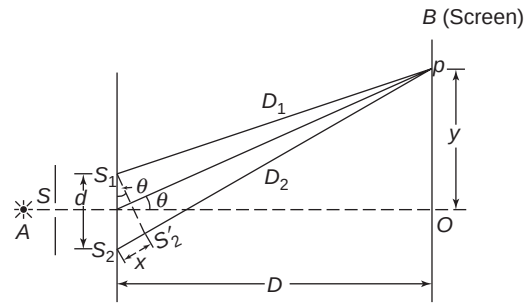


Figure 3-3 Schematic diagram showing the basic arrangement of Young's double-slit experiment on interference.

Obviously, at point O on the screen, $D_1 = D_2$, both light waves from S_1 and S_2 , are exactly in phase (i.e., $\phi_1 - \phi_2 = 0$); bright fringe will appear at point O on the screen. When $x = \lambda/2$, as shown in Figure 3-3, the two waves will be exactly out of phase and tend to cancel each other, so there will be no light (i.e., there will be darkness) at point p on the screen. Bright fringes occur when $\phi_1 - \phi_2 = \pm 2m\pi$, that is, when $D_1 - D_2 = m\lambda$, and the dark fringes occur when $\phi_1 - \phi_2 = \pm(2m + 1)\pi$, or when $D_1 - D_2 = \pm \frac{(2m+1)\lambda}{2}$,

where m is any integer, 0, 1, 2, 3, ... To find a point on the screen at a distance y from point O , we assume that both d and y are much smaller than D . For this case, bright fringes occur at the points p located at a distance y from point O when $y = \pm 2m\lambda D/d$.⁵ Similarly, dark fringes occur for the points at $y = \pm(2m + 1)\lambda D/d$.

Division of Amplitude

The various colors appearing in a soap bubble, or those reflected from the oil film on the top surface of a pool of water, are examples of interference encountered daily. This phenomenon is due to the interference caused by the division of amplitude.

The simplest example of interference caused by the division of amplitude is the interference from a thin film of refractive index n_2 deposited on a substrate of refractive index n_3 . This system is placed in a medium of refractive index n_1 (usually in air with $n_1 = 1$), as shown in Figure

3-4. The beams emerging from A and B at the interface between the medium and the film are parallel; a lens is therefore required to bring them together to produce interference. The beam from A is due directly to the reflection of the incident beam, while the beam from B is due to the reflection of the refracted beam on the rear surface of the film. For simplicity, we choose $n_1 = 1$. The difference in optical path between the beam from B and that from A is given by³

$$\begin{aligned} L_{\text{diff}} &= n_2(AD + DB) - AC \\ &= 2n_2AD - AC \end{aligned} \quad (3-34)$$

Using Snell's law, $n_1 \sin \theta_1 = n_2 \sin \theta_2$, we obtain

$$\begin{aligned} AC &= 2AD \sin \theta_2 \sin \theta_1 \\ &= 2n_2 AD \sin^2 \theta_2 \end{aligned} \quad (3-35)$$

Substituting Equation 3-35 into Equation 3-34, we obtain

$$L_{\text{diff}} = 2n_2 d \cos^2 \theta_2 \quad (3-36)$$

So, bright fringes occur when

$$\frac{4\pi n_2 d \cos \theta_2}{\lambda} = 2m\pi \quad (3-37)$$

or

$$2n_2 d \cos \theta_2 = m\lambda$$

and dark fringes occur when

$$\frac{4\pi n_2 d \cos \theta_2}{\lambda} = \frac{(2m+1)\pi}{2}$$

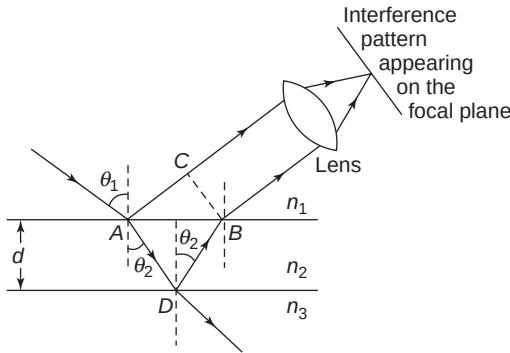


Figure 3-4 Schematic diagram showing the interference caused by the division of amplitude due to beams of light emerging from a thin film of refractive index n_2 with $n_2 > n_1$.

or

$$2n_2 d \cos \theta_2 = \frac{(2m+1)\lambda}{2} \quad (3-38)$$

where m is any integer, 0, 1, 2, 3, . . .

Applying the same principle of analysis for two beams, it is quite straightforward to extend the case for two beams to the cases for multiple beams. Multiple beams will produce interference caused either by the division of wavefront or by the division of amplitude. For more details about this subject, the reader is referred to two very good references.^{3,4} Many applications are derived from the interference phenomenon. Here, we just mention a few examples.

Nonreflecting Surface—It is possible to deposit or to insert a thin dielectric layer on the surface of any object, when reflection of the incident wave is not desired. Such a thin layer or thin film, as shown in Figure 3-5, is sometimes called an antireflection coating. The incident wave reflected from the interface between the film and the medium (usually the air with $n_1 = 1$), called wave A , and the wave reflected from the interface between the film and the object of refractive index n_3 , called wave B , will interfere with each other. In order to make them to interfere destructively, the optical thickness of the thin film must be equal to $\lambda/4$, so that the optical path difference between wave A and wave B is equal to $\lambda/2$, or their phase difference equal to π . To satisfy this condition, the following relation, based on Fresnel's reflection equation, must be satisfied³:

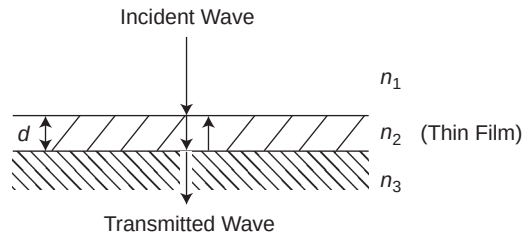


Figure 3-5 Nonreflecting surface produced by insertion of a dielectric film with thickness $d = \lambda/4$ and refractive index $n_2 = (n_1 n_3)^{1/2}$.

$$\frac{n_2 - n_1}{n_2 + n_1} = \frac{n_3 - n_2}{n_3 + n_2} \quad (3-39)$$

This leads to

$$n_2 = (n_1 n_3)^{1/2} \quad (3-40)$$

Appreciable reduction in the reflected irradiance has been found with films of magnesium fluoride or cryolite on the glass surface.

Interferometers—The various types of interferometers may be divided into two main groups based on their operational principles:

- On the basis of the division of wavefront by means of a number of apertures similar to that used in Young's experiment, or by means of prisms and mirrors.
- On the basis of the division of amplitude by means of semireflecting surfaces, part of the light being reflected and part being transmitted. The classical Michelson interferometer belongs to this group.

The basic principle of interferometers is to bring the reflected or transmitted beams together to produce bright and dark interference fringes. Based on the difference in the optical paths and the associated bright and dark interference fringes, it is possible to determine the distance between two points or the thickness of a thin film. The classical method for measuring the distance is the Michelson interferometer, and in fact, nearly all other interferometers are variations of this basic instrument.⁶⁻⁸

Because our aim is to show the basic principle of the interferometer, we have chosen a simple one, called the Fizeau interferometer, because the essential design of this instrument is originally due to Fizeau, but the arrangement shown in Figure 3-6 is based on a more recent development.^{9,10} In this instrument, light from a quasi-monochromatic source L_p is collimated by the lens L and falls on the film T at nearly normal incidence. The light reflected from the top and bottom surfaces of the film returns through L and converges to an apparatus A , which produces the bright and dark interference fringes on the focal plane F . Usually, a black cloth is placed at the base S of the instrument

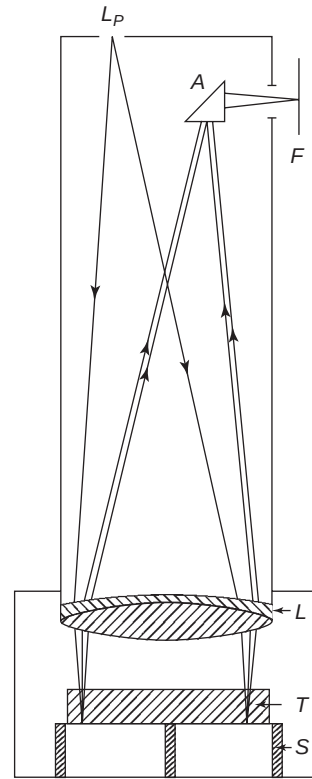


Figure 3-6 The basic arrangement of the essential parts for the Fizeau interferometer.

to absorb the transmitted light. The bright and dark fringes appearing on the focal plane F provide the information about the optical thickness of the film.

Diffraction

If an opaque object is placed between a beam of light and a screen, the shadow cast on the screen is not perfectly sharp. Some light is also present in the dark zone. Similarly, a beam of light that emerges from a small aperture or a narrow slit is observed to spread out. In daily experience, we may have noticed that the beams of a car's headlights are spread out by the mist on a foggy night; the mist is composed of water particles, which scatter the light beams. This phenomenon is called diffraction. If the particles are extremely small, such as atoms arranged in the lattice of a crystal, an

incident beam striking the crystal will undergo Bragg diffraction. The arrangement of the atoms and the structure of the crystal can be determined by examining the angle and the intensity of the diffracted beams. Since the separation between atoms (lattice constant) is extremely small, the wavelength of the incident beam must be small. This is why x-rays must be used. This is well known as x-ray diffraction spectroscopy.

The basic theory of diffraction is that of Huygens and Fresnel, which is so far the most powerful and adequate theory for treating most problems encountered in optical instruments.¹¹ Based on this theory, every point of a wavefront may be considered a center of a secondary disturbance which gives rise to many spherical wavelets spreading out in all directions. Secondary wavelets also mutually interfere with each other. The waves of such spherical wavelets from the center of the disturbance will be superposed to form a new wavefront.

Because of the diffraction effect, images of any object through a lens or a microscope always have some defects, implying that the images always have some distortion in shape from the actual object. This may be caused by diffraction, aberration, or optical noise. Here, we will consider only the effect of diffraction. Let us take the resolution of a microscope as an example. Airy was the first to compute the diffracted image,¹² showing in 1834 that for a diffraction at a circular aperture of diameter D , the angle α between the first maximum (measured from the center) and the first dark ring is given by

$$\sin \alpha = \frac{1.22\lambda}{D} \quad (3-41)$$

where λ is the wavelength of the light. The maxima are separated by the minima, which occur at the angle of diffraction. For a circular aperture, the pattern would consist of a central bright area surrounded by concentric dark and bright rings. According to Airy, about 84% of the light is concentrated within the central spot, which is called the Airy or diffraction disk, as shown in Figure 3-7(a). For instance, when

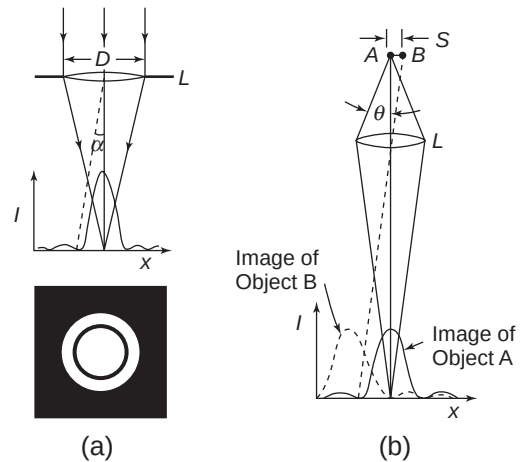


Figure 3-7 (a) Diffraction at the aperture of a lens showing an Airy disk, and (b) the resolution limit of an optical microscope.

light from a lamp, through a small pinhole, falls on a cardboard sheet located at a distance of several meters, we can see a bright spot and also dark rings. The same kind of pattern can be observed when a pointed object is imaged by a microscope.

Generally, a microscope is commonly used to determine the separation between two small objects. This is exactly what photolithography does in today's microelectronic technology. In this case, we may have three possible resolutions:

- The two objects produce an overlapping image, so we can see only one object through the microscope.
- The two objects can be observed distinguishably, but their relative positions are not distinguished.
- The two objects are clearly resolved. However, the demarcation between the "resolved" and the "not resolved" is not perfectly sharp.

The Rayleigh criterion imposes the condition that the two objects can be clearly distinguished when the central maximum, created by object A, coincides with the first minimum created by the other object B, as shown in Figure 3-7(b). Thus, for a distinguishable image, the separa-

tion S between the two objects should follow the criterion

$$S = \frac{1.22}{2n \sin \theta} \quad (3-42)$$

where n is the refractive index of the medium. This means that for the two objects in the image are just resolved, S must be the value not smaller than that limited by Equation 3-42. The numerical aperture (NA) is given by

$$NA = n \sin \theta$$

So

$$S = \frac{0.6\lambda}{NA} \quad (3-43)$$

Thus, the higher the value of NA of a lens, the higher the quality of the lens. Based on this criterion, if the system is placed in dry air, $n = 1$, values of NA up to 0.95 have been used. However, if the system is immersed in oil, the value of NA can be higher than 1.45. Thus, for $\lambda = 5500\text{\AA}$, the minimum value of S having good resolution is about $3.5 \times 10^{-5}\text{cm}$ in dry air and about $2.5 \times 10^{-5}\text{cm}$ in oil. So there is some advantage to placing the system in a medium having a high n and hence high NA .

Polarization, Reflection, and Refraction

Any beam of ordinary light comprises many individual waves, so the planes of the vibration of the electric field vectors are randomly oriented. Such a beam of light is unpolarized, and the resultant electric field vector changes its orientation randomly in time. A beam of light with its electric field vector vibrating only in a specific plane is referred to as plane-polarized, as shown in Figure 3-8. The wave propagates in the z direction, while the electric field vector F vibrates in the x - z plane (i.e., $y = 0$) and the magnetic field vector H vibrates in the yz plane (i.e., $x = 0$). We have mentioned that in describing optical phenomena, we usually consider only the electric field vector F because the magnetic field vector H behaves in the same way as F . The simplest form of polarization is plane polarization. There are other forms of polarization. Suppose that we have two plane waves,

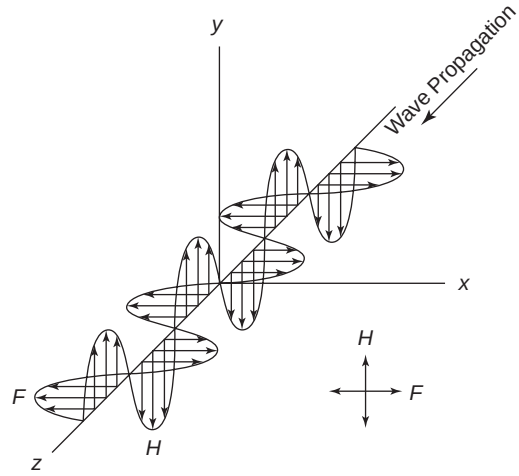


Figure 3-8 Plane-polarized light wave with the electric field vector F vibrating in the x - z plane ($y = 0$) and the magnetic field vector H vibrating in the y - z plane ($x = 0$).

one with F_x vibrating in the xz plane and the other with F_y vibrating in the yz plane (see Figure 3-8), then we can write

$$F_x = F_{x0} \cos \omega t \quad (3-44)$$

$$F_y = F_{y0} \cos(\omega t + \theta) \quad (3-45)$$

By eliminating ωt and combining Equations 3-44 and 3-45, we obtain

$$\left(\frac{F_x}{F_{x0}}\right)^2 - \frac{2F_x F_y}{F_{x0} F_{y0}} \cos \theta + \left(\frac{F_y}{F_{y0}}\right)^2 = \sin^2 \theta \quad (3-46)$$

There are three special cases.

First, if $F_{x0} \neq F_{y0}$ and $\theta = \pm \frac{\pi}{2}$, then Equation 4.46 can be simplified to

$$\left(\frac{F_x}{F_{x0}}\right)^2 + \left(\frac{F_y}{F_{y0}}\right)^2 = 1 \quad (3-47)$$

This is an elliptically polarized wave.

Second, if $F_{x0} = F_{y0}$ and $\theta = \pm \frac{\pi}{2}$, then Equation 4-46 becomes

$$F_x^2 + F_y^2 = F_{x0}^2 \quad (3-48)$$

This is a circularly polarized wave.

Third, if $F_{x0} \neq F_{y0}$ and $\theta = \pm \pi$, then Equation 3-46 becomes

$$\frac{F_x}{F_{x0}} + \frac{F_y}{F_{y0}} = 0 \quad (3-49)$$

This is a linearly polarized wave.

Of course, if θ were different from 0, $\pi/2$, and π , the shapes of the polarized waves would be different from the three cases just mentioned.

Polarizers

Certain sources provide polarized light without the use of polarizers, such as light from lasers and light from the blue sky, which is partially polarized. However, plane-polarized light may be produced from unpolarized light in a number of ways. The earliest experiment to produce plane-polarized light from ordinary, unpolarized light is due to Malus.^{3,14} He used two parallel, unsilvered mirrors, such as glass plates, as shown in Figure 3-9. The light is strongly reflected from G_1 . If mirror G_2 is turned in such a direction to reflect most of the light out of the x - y plane, then mirror G_1 makes the light strongly reflected in the x - y plane and not in the direction perpendicular to the x - y plane. In other

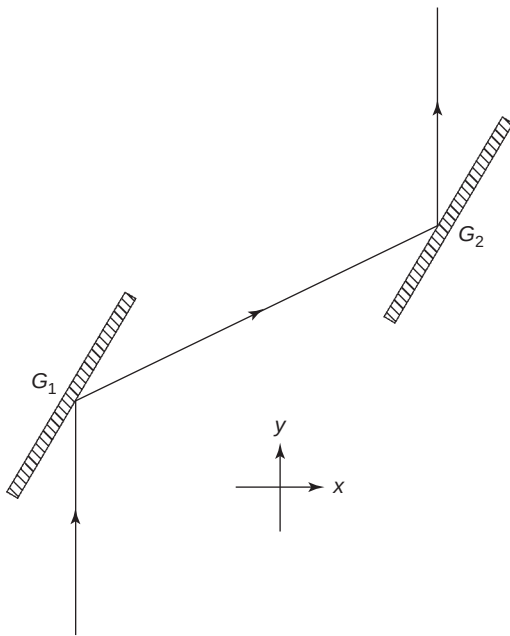


Figure 3-9 Malus's successive reflection experiment employing two parallel transparent mirrors G_1 and G_2 .

words, the first mirror G_1 polarizes the light, and the second mirror G_2 reveals that it is polarized. An apparatus that can produce a beam of plane-polarized light from a beam of unpolarized light is called a polarizer. This is similar to a filter, which allows only a narrow band of frequencies to pass through and filters out all other frequencies. An apparatus that can detect plane-polarized light is called an analyzer. Generally, an apparatus capable of being a polarizer is also capable of being an analyzer, and vice versa. We shall consider two commonly used polarizers: one is by reflection and the other by transmission; both are based on the condition of no reflection at the Brewster angle. The former suffers a great loss of light, since only a fraction of the incident beam is reflected, while the latter does not give complete polarization, though a pile of glass plates is used.

Light Polarized by Reflection—When a beam of unpolarized light strikes the boundary between two media, the incident beam on the plane of incidence perpendicular to the boundary will be partly reflected and partly refracted, thus changing the direction of propagation, amplitude, and phase, as shown in Figure 3-10(a). The electric field vector F of the incident beam may be vibrating in any direction, such as along the r -direction but in the rr_1 plane. However, F can be resolved into two components: one vibrating in the plane of incidence (i.e., parallel to the plane of incidence), termed F_p , and the other vibrating perpendicular to the plane of incidence (i.e., normal to the plane of incidence), termed F_n , as shown in Figure 3-10(b). The symbol used for F_p is $\equiv$ and that for F_n is $\cdots$. Thus, the symbol for an unpolarized beam, with F mixed with F_p and F_n , is $\equiv \cdots$. The normal reflection and refraction for an unpolarized light wave is shown in Figure 3-10(c). However, when the angle of incidence is equal to Brewster angle θ_b , (i.e., $\theta_i = \theta_b$) if the incident beam is a normally plane-polarized beam, the reflected beam and the refracted beam are both normally plane-polarized, as shown in Figure 3-11(a). When the incident beam is parallel plane-polarized, the incident

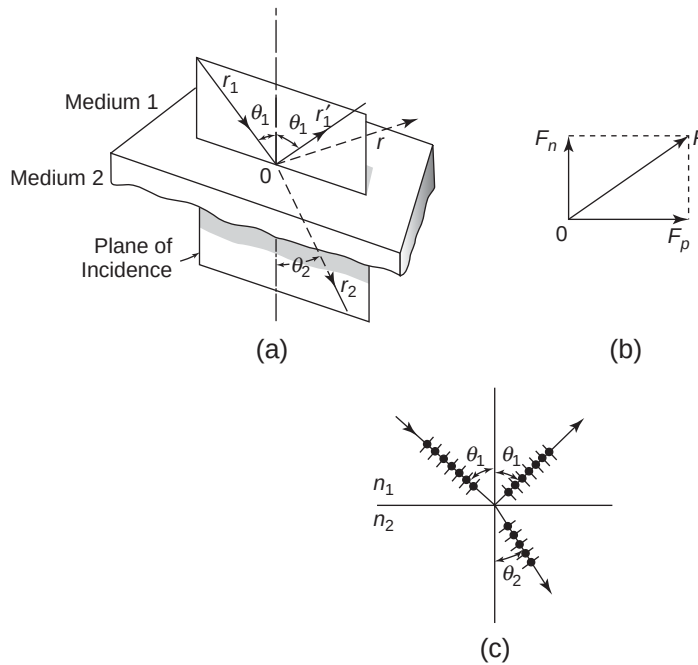


Figure 3-10 Schematic diagrams showing (a) the plane of incidence and the angles of incidence θ_i , of reflection θ_r , and of refraction θ_2 , (b) the normal component F_n and the parallel component F_p of the electric field vector F , and (c) the reflection and refraction of an unpolarized beam at any angle of incidence θ_i .

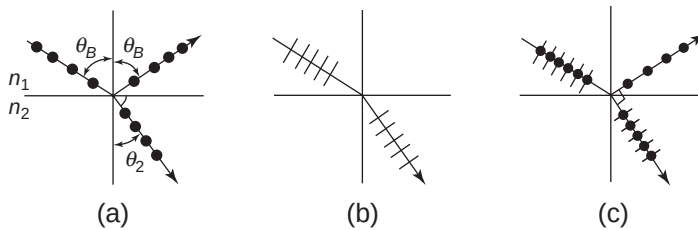


Figure 3-11 The reflection and refraction for the incident beam striking the boundary between two media at the angle of incidence $\theta_i = \theta_B$ (Brewster angle): (a) for the normal plane-polarized beam, (b) for the parallel plane-polarized beam, and (c) for the unpolarized beam.

beam and the refracted beam are parallel plane-polarized, but there is no reflected beam, as shown in Figure 3-11(b). For the incident unpolarized beam, the refracted beam is still partially unpolarized, but the reflected beam is highly plane-polarized. In this case, the plane-polarized reflected beam is polarized in the direction normal to the plane of incidence. Therefore, this method can be used to produce a normally plane-polarized beam from an unpolarized beam.

From Figure 3-11, it can be seen that only for the incident beam parallel plane-polarized at $\theta_i = \theta_B$, there is no reflection. When $\theta_i = \theta_B$, then

$$\theta_2 = \pi/2 - \theta_B \quad (3-50)$$

Based on Snell's law,

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \quad (3-51)$$

so from Equations 3-50 and 3-51, θ_B can be expressed as

$$\theta_B = \tan^{-1}(n_2/n_1) \quad (3-52)$$

If medium 1 is air, $n_1 = 1$, and medium 2 is glass, $n_2 = 1.5$, then $\theta_B = 56.3^\circ$. For the reverse case—medium 1 is glass and medium 2 is air—then $\theta_B = 33.7^\circ$. At this critical angle θ_B , the reflectance from the boundary is equal to zero for parallel plane-polarized light.^{15,16} No reflection occurs, because the electric field vector F_p in medium 2 sets the atomic electrons into oscillation. This oscillation of atomic electrons acts like that of Hertzian oscillators, which radiate energy only in the direction perpendicular to the direction of the oscillation but radiate no energy in the direction of the oscillation. That is why there is no reflection in medium 1 in the direction parallel to the direction of the oscillation of the atomic electrons in medium 2.

The reflectance, which is defined as the ratio of the reflected to the incident energy, is given by

$$R_p = \frac{(F_{ip}^r)^2}{(F_{ip}^i)^2} \quad (3-53)$$

for a parallel plane-polarized wave, and

$$R_n = \frac{(F_{in}^r)^2}{(F_{in}^i)^2} \quad (3-54)$$

for a normally plane-polarized wave. Both R_p and R_n depend on the angle of incidence θ_1 and on the optical constants of the medium. In general, the optical constants should be expressed in complex form because of the dispersion of the material. The complex refractive index is given by

$$n^* = n - jk \quad (3-55)$$

where n is the refractive index and k is the absorption index or the extinction coefficient, which represents the attenuation of the wave per unit wavelength. The calculation for the reflectance in general is quite mathematically involved. For the simple case, assuming that medium 1 is a transparent material with $n_1^* = n_1$ and $k_1 = 0$, and medium 2 is an absorbing material with $n_2^* = n_2 - jk_2$, then the reflectance can be expressed as¹⁷

$$R_p = \frac{(n_2^2 + k_2^2)\cos^2 \theta_1 - 2n_1n_2 \cos \theta_1 + n_1^2}{(n_2^2 + k_2^2)\cos^2 \theta_1 + 2n_1n_2 \cos \theta_1 + n_1^2} \quad (3-56)$$

and

$$R_n = \frac{(n_2^2 + k_2^2) - 2n_1n_2 \cos \theta_1 + n_1^2 \cos^2 \theta_1}{(n_2^2 + k_2^2) + 2n_1n_2 \cos \theta_1 + n_1^2 \cos^2 \theta_1} \quad (3-57)$$

Only for normal incidence (i.e., $\theta_1 = 0$), $R_p = R_n$. Otherwise, R_p and R_n are different. The reflectance as a function of the angle of incidence is shown in Figure 3-12. It can be seen that only for R_p there is a Brewster angle θ_B , at which $R_p = 0$.

Light Polarized by Transmission—Figure 3-11(c) shows that some of the light polarized normal to the plane of incidence is reflected; all of the light parallel plane-polarized is transmitted to medium 2. If a pile of glass plates is used for this polarizing process, it is possible to filter out the light normally plane-polarized, leaving the portion parallel plane-polarized to pass through, as shown in Figure 3-13. After passing through a pile of glass plates, the transmitted light becomes a highly plane-polarized light, but in this case, the light is polarized parallel to the plane of incidence.

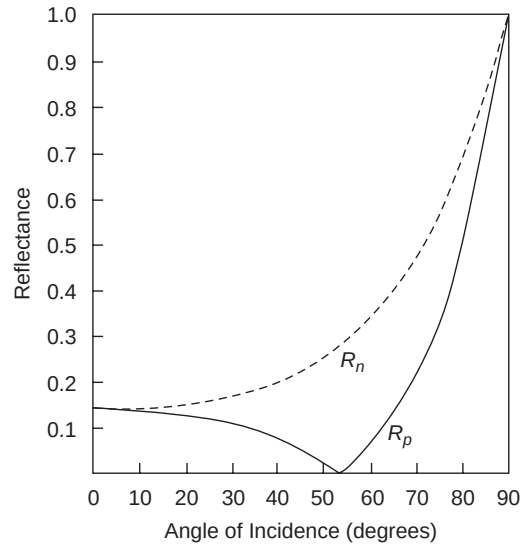


Figure 3-12 Reflectance as a function of angle of incidence for the light-polarized parallel (R_p) and normal (R_n) to the plane of incidence for the light reflected from water surface.

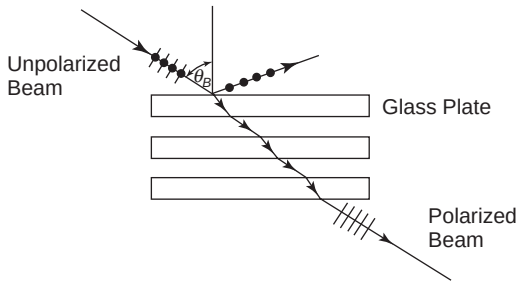


Figure 3-13 Polarization by transmission through a pile of glass plates with the incident unpolarized beam striking the surface of the first glass plate at the Brewster angle.

Total Internal Reflection

Based on Snell's law, it is possible to make the angle of refraction $\theta_2 = \pi/2$ by changing the angle of incidence to a critical angle termed θ_c . For this to occur, n_1 must be larger than n_2 , (i.e., $n_1 > n_2$). So, when a light beam in medium 1 strikes the surface of medium 2, which is a less dense medium, at the critical angle θ_c , which is

$$\theta_c = \sin^{-1}(n_2/n_1) \quad (3-58)$$

then the incident beam will be totally reflected. Usually, the incident beam is not reflected at the striking point on the boundary surface, but penetrates some distance into medium 2 before it is reflected. The reflecting surface of the reflected beam is displaced slightly behind the real reflecting surface. This is known as the Goos-Hanchen effect.¹⁸ The total internal reflection is important in optical fiber waveguides.

Dispersion

A pulse or signal of any electromagnetic waves may be constructed to form wave packets by superposition of all harmonic waves covering a frequency interval with prescribed amplitudes. The velocities, with which the constant phase surface of these component waves are propagated, depend on the parameters ϵ (permittivity), μ (permeability), and σ (conductivity) of the medium. If the medium is non-conducting and ϵ and μ are independent of the frequency of the waves, then the phase velocity is constant and the signal is propagated

without distortion. Such a medium is not dispersive. On the other hand, if the medium is conducting, then the propagating harmonic waves suffer attenuation, relative displacements in phase in the direction of propagation, and the signal arrives at a distant point in a distorted form. Such a medium, in which the phase velocity is frequency-dependent, is dispersive.

At optical frequencies, a material may exhibit this dispersive phenomenon even when the conductivity due to free charges is totally negligible. In nonmagnetic dielectric materials, the phase velocity is related to the refractive index $n = (\epsilon_r)^{1/2}$. Strictly speaking, however, there is no material free of dielectric losses; therefore, there is no material completely free of dispersion. Dispersion is an intrinsic property of all dielectric materials, and all other properties must coexist with it. So, the refractive index must be in complex form, as given by Equation 3-55.

A dielectric solid may be considered an assembly of oscillators (each atom or molecule is an oscillator) set into forced vibration by the excitation force or by the radiation of optical frequencies. Under an applied electric field or the electric-field vector of the incident electromagnetic wave of radiation, the bound electrons (or ions) will be displaced from their equilibrium positions, giving rise to polarization. These bound electrons (or ions) will also experience a restoring force proportional to the displacement Δx and a damping force proportional to the velocity $d(\Delta x)/dt$. Thus, the motion of an electron (or an ion) is governed by the following equation

$$m \frac{d^2 \Delta x}{dt^2} + mG \frac{d\Delta x}{dt} + m\omega_0^2 \Delta x = qF_x \exp(j\omega t) \quad (3-59)$$

where m is the mass of the electron (or the ion), ω_0 is the natural frequency of the oscillator, G is the force constant, and $F_x \exp(j\omega t)$ is the excitation field in the x direction. The solution of Equation 3-59 is

$$\Delta x = \frac{qF_x}{m[(\omega^2 - \omega_0^2) + jG\omega]} \quad (3-60)$$

Thus, we have the complex relative permittivity (or dielectric constant)

$$\begin{aligned}\epsilon_r^* &= 1 + \frac{N(q\Delta x)}{\epsilon_o F_x} \\ &= 1 + \frac{Nq}{\epsilon_o F_x} \left\{ \frac{qF_x}{m[(\omega^2 - \omega_o^2) + jG\omega]} \right\}\end{aligned}\quad (3-61)$$

Thus, we can write

$$\epsilon_r^* = \epsilon_r - j\epsilon_r' = (n - jk)^2 = (n^2 - k^2) - j2nk \quad (3-62)$$

This is the dispersion equation. It implies that any light beam passing through a dielectric medium will suffer attenuation and hence dispersion. More details about this topic will be given in Section 3.3.4.

3.2 Modulation of Light

If optical beams are used as information carriers, then transmission of the information requires the optical beams to be modulated. This process of placing information onto an information carrier is referred to as modulation, and the converse process is referred to as demodulation (the recovery of the information). Optical beams can be modulated by varying one or more of the wave parameters: frequency, amplitude, phase, polarization, and propagation direction. Frequency, amplitude, and phase modulations are well known in radio and microwave communications. Polarization and propagation direction are related to the variation of the refractive index. In this section, we shall discuss several important effects that lead to the modulation of light.

3.2.1 Double Refraction and Birefringence

In Section 3.1.2, we mentioned that the superposition of two waves, one with the electric field vector vibrating in the x direction and the other vibrating in the y direction, will produce elliptical, circular, or linear polarizations, depending on the relative values in amplitude and phase of the two waves. There are some materials in which the refractive index for light linearly polarized in one direction (e.g., in the x direction) is different from that for the light linearly polarized in another (e.g., in the y

direction). This implies that such materials are anisotropic and that the light with the electric vector in one direction travels faster than that in another. Some materials consist of non-spherical, long molecules, and these molecules may be arranged with their long axes parallel to each other. Because of the structure of the molecules, it is possible that the electrons respond more easily to the electric field oscillation in the direction parallel to the long axes of the molecules than they would respond to the electric field perpendicular to these axes. Such materials, which have double refractive indices, are said to be birefringent. If a light beam is incident normally on such birefringent materials, the beam will be split in two upon refraction, except in certain preferred directions. Since the x -directed and the y -directed electric field vectors travel with different velocities, their phases change at a different rate as the light propagates through the material. Although on the material surface the x -directed and the y -directed electric field vectors are in phase, the phase difference between these two electric field vectors inside the material becomes proportional to the depth the light has traveled into the material. Anisotropic materials, such as Calcite (CaCO_3), quartz (SiO_2), and KDP (potassium dihydrogen phosphate, KH_2PO_4), belong to this kind of material. If a narrow beam of unpolarized light is incident normally on the surface of a birefringent plate, the beam will be split into two refracted beams: the ordinary ray (O ray) and the extraordinary ray (E ray), as shown in Figure 3-14(a). The O ray passes through the plate, but the E ray diverges as it passes through the plate and then emerges parallel to the original beam direction. If the emergent beam from Plate I is allowed to fall on Plate II, whose optic axis is parallel to that of Plate I, the result is shown in Figure 3-14(b). The combination of these two plates behaves like a single plate of the thickness equal to the sum of the thicknesses of the two plates (i.e., $1d + d/2 = 1.5d$). If Plate II is rotated through an angle of π about the direction of the incident beam, the deflections in the two plates are of opposite signs. Thus, the combination of these two plates behaves like a single plate of the

thickness equal to the difference in thickness between the two plates, (i.e., $1d - d/2 = d/2$), as shown in Figure 3-14(c).

The direction of polarization of the O ray is perpendicular to that of the E ray. According to Huygens,³ since there are two rays, the wave should consist of two equiphase wavefront surfaces, one for the O ray and the other for the E ray. The O ray obeys Snell's law in all directions, indicating that the surface to represent its equiphase wavefront surface is a sphere, while for the E ray, the surface to represent its equiphase wavefront surface should be an ellipsoid of revolution. Huygens also assumed that all birefringent crystals can be considered uniaxial and that the spheroid touches the sphere either externally, as shown in Figure 3-15(a), or internally, as shown in Figure 3-15(b), depending on whether the uniaxial crystal is positive

or negative. In either case, the whole surface is formed by revolving the curves about the line joining the points of contacts. If the wave propagation is along this line, both the O ray and the E ray have the same propagation velocity. This line is called the optic axis. The two rays have mutual perpendicular polarization. The plane containing the optic axis and the O ray is called the principal plane for the O ray, and that containing the optic axis and the E ray is called the principal plane for the E ray. The O ray is polarized perpendicular to the principal plane of the O ray, while the E ray is polarized in the principal plane of the E ray. For positive uniaxial crystals, the refractive index for the O ray n_o is smaller than that for the E ray n_E (i.e., $n_o < n_E$), as in quartz (SiO_2), while for negative uniaxial crystals, the refractive index for the O ray is larger than that for the E ray (i.e., $n_o > n_E$).

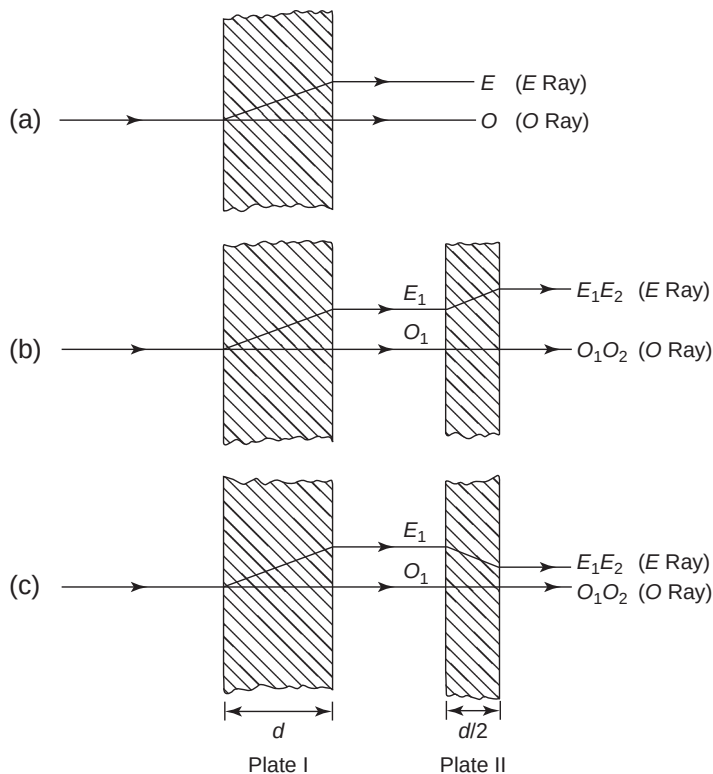


Figure 3-14 (a) Double refraction of light (birefringence) in an anisotropic crystal plate, (b) double refraction in two parallel plates with the same orientation of the crystal axes, and (c) double refraction in two parallel plates, but in this case, the second has been rotated through an angle of π about the direction of the incident light beam.

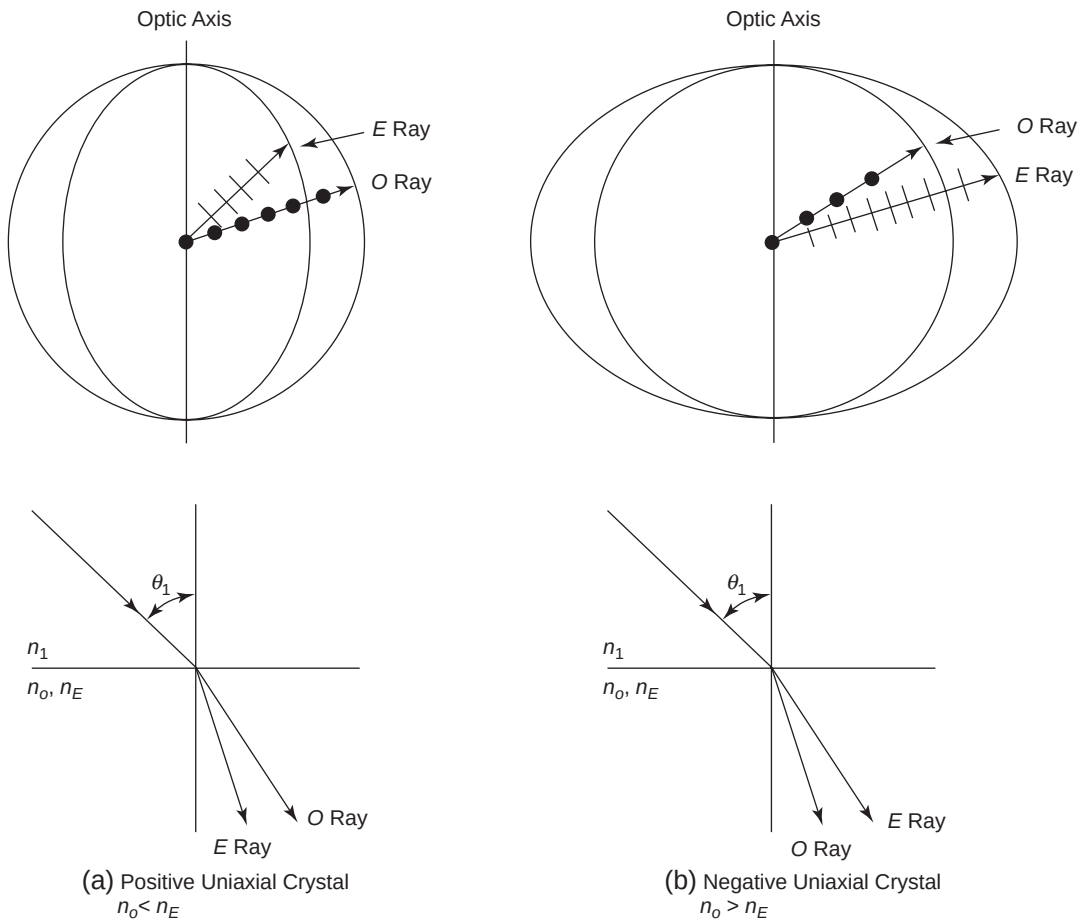


Figure 3-15 Equiphase wavefront surfaces for ordinary and extraordinary rays for (a) positive uniaxial crystal and (b) negative uniaxial crystal.

Suppose that a point source of light is radiating uniformly inside the crystal; the O ray component is polarized with its electric field vector perpendicular to its principal plane and has a spherical equiphase wavefront surface. This is why the O ray propagates at the same velocity in all directions and the crystal for the O ray behaves like an isotropic crystal. The E ray is polarized in its principal plane and orthogonally to the O ray. Its equiphase wavefront surface is a spheroid, and this is why the velocity of the E ray propagation varies with direction. Only along the optic axis, its propagation velocity is the same as that of the O ray. At right angles to the optic axis, the propagation velocity of the E ray is either a maximum

for the negative uniaxial crystal, in which $n_o > n_E$, or a minimum for the positive uniaxial crystal, in which $n_o < n_E$, as shown in Figure 3-15. If an unpolarized light beam is incident on the surface of a birefringent material, there are two refracted beams inside the material: the O ray and the E ray. Their refraction angles depend on the value of n_o and n_E , as shown in Figure 3-15.

Quarter-Wave Plates (QWP)

It can be seen from Figure 3-15 that when a light beam propagates along the direction perpendicular to the optic axis in a uniaxial crystal, the difference in velocity between the O ray and

the E ray is the greatest. Therefore, there will be a phase difference between the two beams. After these two beams have traveled a distance d inside the crystal, the phase difference is

$$\Delta\phi = \frac{2\pi}{\lambda}(n_o d - n_E d) \quad (3-63)$$

Suppose we have a uniaxial crystal plate of thickness d . If the difference in optical paths $(n_o - n_E) d = \lambda/4$ (quarter wavelength), then the phase difference between the O ray and the E ray becomes

$$\Delta\phi = \pi/2 \quad (3-64)$$

Thus, the combination of these two rays will have a resultant electric field vector that will trace an ellipse on the exit facet of the crystal. Based on the same principle, we can also produce half-wave or whole-wave plates. In addition, we can adjust the direction of the incident beam relative to the optic axis. If the angle between the direction of the incident beam and the optic axis is 45° , the emergent beam will be circularly polarized. However, a variety of the forms of polarization can be obtained using uniaxial crystal plates. Such plates are often used in light modulation systems.

Optical Activity

When a beam of plane-polarized light passes through certain types of crystals, the light remains plane-polarized but rotates about the direction of polarization, either in clockwise rotation (called right-hand or positive rotation) or in counterclockwise rotation (called left-hand or negative rotation). Such crystals are generally referred to as optically active materials, and such an intrinsic phenomenon is referred to as natural optical activity. Obviously, only anisotropic crystals may have this property. Some birefringent crystals exhibit a gyration phenomenon due to this natural rotation of the polarization plane. However, optical activity can be induced by an external excitation force. Some materials, which are normally isotropic in their optical properties, may become anisotropic and optically active by an external force, such as an applied electric field,

a magnetic field, optical radiation, or mechanical stress. This leads to many ways to produce light modulation. In the following sections, we shall discuss some commonly used methods for making material optically active.

3.2.2 Electro-Optic Effects

This section deals with the ways of making some crystals optically active by means of an applied electric field. An applied electric field across a crystal will alter the normal distribution of electrons in it and hence cause changes in polarization P and refractive index n . Thus, the field dependence of the refractive index can be written as

$$n_F = n + aF + bF^2 + \dots$$

or

$$\Delta n = n_F - n = aF + bF^2 + \dots \quad (3-65)$$

where n is the refractive index in the absence of an applied field (i.e., $F = 0$), Δn is the change of the refractive index, and a and b are electro-optic constants. The effect of the electric field on the change of the refractive index is referred to as the electro-optic effect. It is evident that if a crystal has a center of symmetry or a material has a random structure, there is no electro-optic effect. Only crystals that are noncentrosymmetric exhibit the electro-optic effect. In the following sections, we shall discuss two well known electro-optic effects commonly used in light-modulation systems.

The Pockels Effect or Linear Electro-Optic Effect

The Pockels effect is the first order electro-optic effect, which was first discovered by F. Pockels in 1893¹⁹ and is now probably the most useful element in modulation applications.²⁰ Let us consider a three-dimensional dielectric ellipsoid, as shown in Figure 3-16, whose shape varies with the electric field. In the absence of an applied electric field ($F = 0$), the refractive index is isotropic, and the ellipsoid should be a sphere. Suppose a beam of plane-polarized light is propagating in the z direction

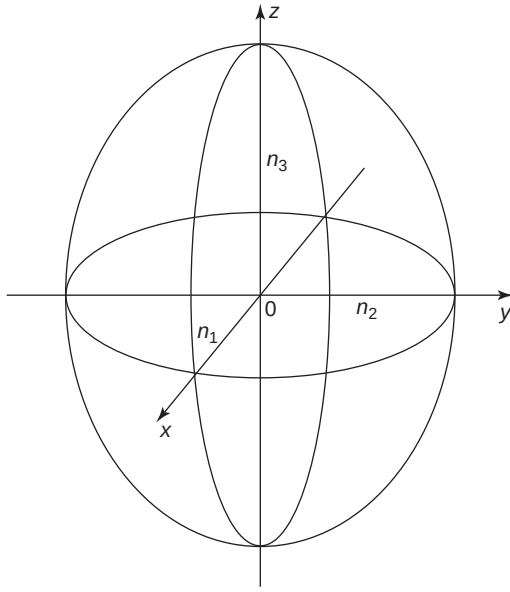


Figure 3-16 An ellipsoid for a biaxial crystal used for explaining the effect of an applied electric field.

through an electro-optic crystal such as KDP, and also an electric field F is applied across the crystal along the z direction. The refractive index does not change in the z direction but changes in the transverse direction. For KDP, if F is along the z direction, then the original principal axes of the ellipsoid are rotated through 45° into new principal axes x' and y' , due to the induced optical activity. The refractive index changes in such a way that the refractive index for the polarization in the x' direction changes by $+\Delta n$ and that for the polarization in the y' direction by $-\Delta n$, so we can write

$$\begin{aligned} n_{x'} &= n + \Delta n \\ n_{y'} &= n - \Delta n \end{aligned} \quad (3-66)$$

where

$$\Delta n = \frac{n^3}{2} r F \quad (3-67)$$

and r is the electro-optic coefficient for the Pockels effect and n is the refractive index in the absence of an applied electric field. The derivation of the Pockels coefficient r is based on the dependence of the refractive index on the electric polarization P , and the analysis

involves tensors. A linear electro-optic effect in quartz was first analyzed by F. Pockels in 1893.¹⁹ For the details about the derivation, see references 5 and 19–23.

The basic arrangement of a Pockels light modulator is shown in Figure 3-17. Since the two beams travel at different velocities in the crystal, the emergent beams will have a phase difference given by

$$\begin{aligned} \Delta\phi &= \frac{2\pi}{\lambda} (n_{x'} - n_{y'}) d \\ &= \frac{2\pi}{\lambda} (2d) \frac{n^3}{2} r F \\ &= \frac{2\pi}{\lambda} n^3 r V \end{aligned} \quad (3-68)$$

where d is the thickness of the crystal and V is the applied voltage ($V = Fd$). With this phase difference, the emergent beam is generally elliptically polarized. This elliptically polarized beam is then made horizontally plane-polarized by an analyzer before it is transmitted to the other optical transmission system.

Since the electric field vectors $F_{x'}$ and $F_{y'}$ of the beam emerging from the electro-optic crystal depend on the phase difference $\Delta\phi$, we can write

$$F_{x'} = \frac{F_{x'o}}{\sqrt{2}} \cos(\omega t + \Delta\phi/2) \quad (3-69)$$

$$F_{y'} = \frac{F_{y'o}}{\sqrt{2}} \cos(\omega t - \Delta\phi/2) \quad (3-70)$$

It can easily be shown that the total electric field vector $F_T = F_{x'} + F_{y'}$, after passing through the horizontal polarizer (analyzer), is in a form similar to Equations 3-18 and 3-19. The transmitted light intensity (i.e., irradiance) is $I \propto F_T^2$. It can be shown that the transmitted light intensity is given by

$$\begin{aligned} I &= I_o \sin^2(\Delta\phi/2) \\ &= I_o \sin^2\left(\frac{\pi}{\lambda} n^3 r V\right) \end{aligned} \quad (3-71)$$

where I_o is the irradiance (i.e., intensity) of the light incident on the electro-optic crystal.⁵ The irradiance of the output light varies with applied voltage, and there is a maximum output, which occurs at $V = V_m$ when

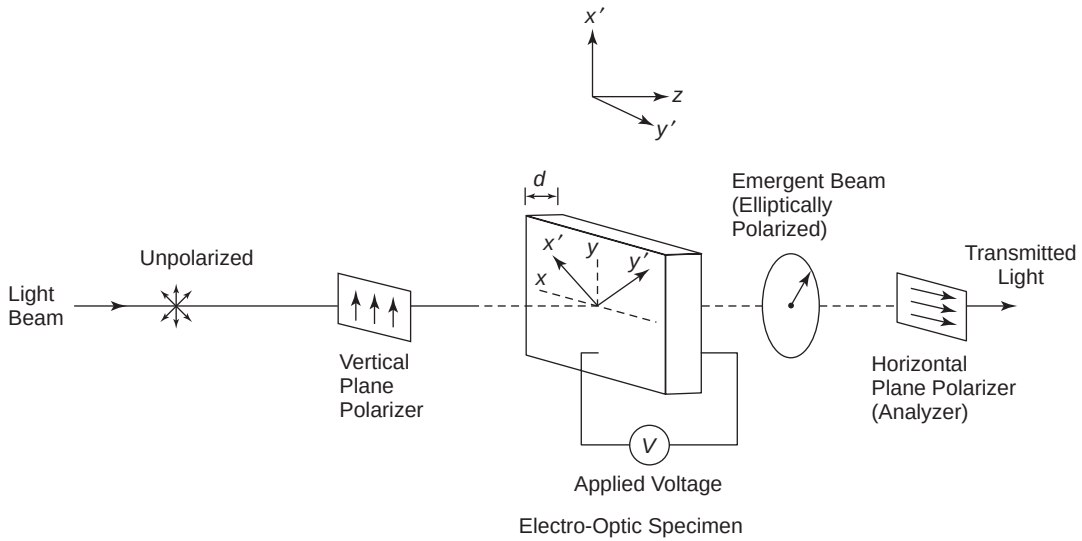


Figure 3-17 The basic arrangement of the components of a Pockels light modulator.

$$\frac{\pi}{\lambda} n^3 r V = \frac{\pi}{2} \quad (3-72)$$

or when

$$V = V_m = \frac{\lambda}{2n^3 r}$$

The variation of I/I_0 with V is shown in Figure 3-18(a).

If the applied voltage V is superimposed on a small signal, the transmitted light will be modulated according to the small signal, as shown in Figure 3-18(b). Obviously, the Pockels electro-optic cell can be used as a light switch. We can adjust the irradiance of the light output from its maximum value by increasing the applied voltage to the value of V_m , and we can turn off the light by reducing the applied voltage to zero, as indicated in Figure 3-18(a).

The Kerr Effect or Quadratic Electro-Optic Effect

The Kerr effect is the second order electro-optic effect. In 1875, John Kerr was the first to show that many isotropic materials behave like a uniaxial crystal with the optic axis in the direction of the electric field.²⁴ When an electric field

is applied across the material in the direction perpendicular to the incident light beam, however, birefringence will be induced. Kerr used glass at that time to show this effect.^{3,20,24} If the applied field F or the applied voltage ($V = FL$) is in the direction perpendicular to the incident light beam, the change in refractive index is given by

$$\Delta n = n_p - n_n = K\lambda F^2 \quad (3-73)$$

where K is the Kerr constant (or the birefringence coefficient) and λ is the wavelength of the incident light in free space. The dimension of the Kerr constant is mV^2 .

If an optically inactive, transparent, dielectric liquid is composed of long asymmetric molecules and these molecules are polar, possessing permanent dipoles, then at zero applied field, the dipoles are randomly oriented; however, at high applied fields, the molecules will be polarized and the dipoles will tend to line up, making the liquid become anisotropic and birefringent, and exhibit optical activity. This means that under a high applied field, the propagation velocity of the incident light polarized in the direction parallel to the optical axis is different from that normal to the optical axis. The applied electric field F produces two effects:

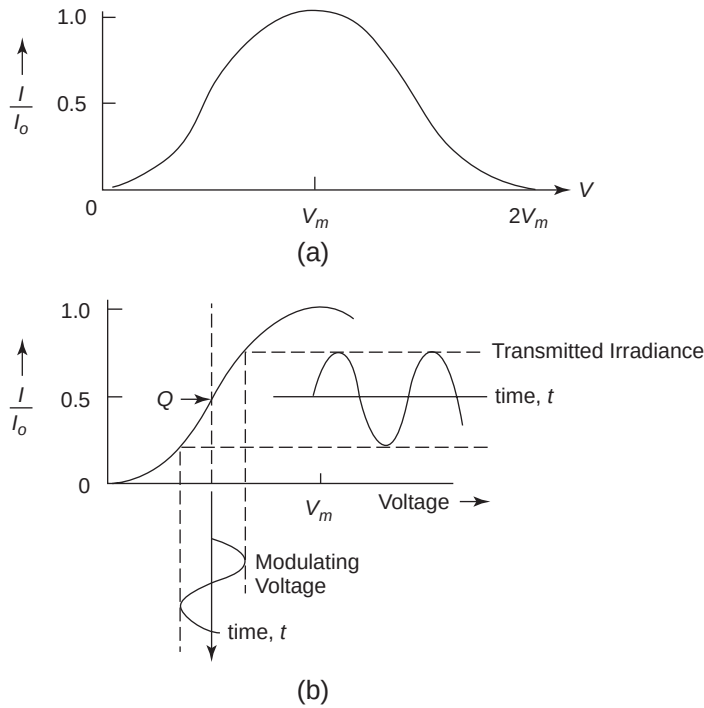


Figure 3-18 (a) The ratio of output light irradiance to input light irradiance as a function of applied voltage and (b) the variation of the irradiance of the modulated output light with time corresponding to the variation of the modulating signal voltage with time.

The first effect is the orientation of the dipolar molecules (dipoles) with their axes of greatest polarizability in the direction of the field, and the other effect is the orientation of these dipoles. In nonpolar materials, only the first effect is present, but in polar materials, both effects will occur, with the second effect usually dominant. The field dependence of the difference between the refractive index n_p when the electric field vector of the light wave F_p is parallel to the applied field F , and n_n when the field vector F_n is normal to F , is shown in Figure 3-19.

It can be imagined that molecular dipoles are very sluggish in responding to the periodic change of the electric field vector of the light beam, whose frequency is of the order of 10^{14} Hz. This implies that it is the applied electric field F that causes the orientation and polarization of the molecules, resulting in the change of the refractive index. In the Fundamental Concepts section of Chapter 2, we assumed the

electric susceptibility χ and polarizability α to be independent of applied electric field. However, in nonlinear optics, the polarization P is no longer just equal to $\chi\epsilon_0 F$, but should be written as

$$P = (\chi + aF + bF^2 + \dots)\epsilon_0 F \quad (3-74)$$

If we consider only the quadratic field dependence, P may be simplified to

$$P = \epsilon_0(\chi F + bF^3) \quad (3-75)$$

Based on this field dependence of the polarization, the Kerr constant K has been derived to find the permittivity of the material at the optical frequency.^{5,25,26} This constant is given by

$$K = \frac{N}{30\epsilon_0 kT} \left(\Delta\alpha + \frac{u_o^2}{kT} \right) \quad (3-76)$$

where N is the number of molecules per unit volume, u_o is the dipole moment of the molecular dipole, $\Delta\alpha = \alpha_p - \alpha_n$, and α_p and α_n are, respectively, the induced polarizability along

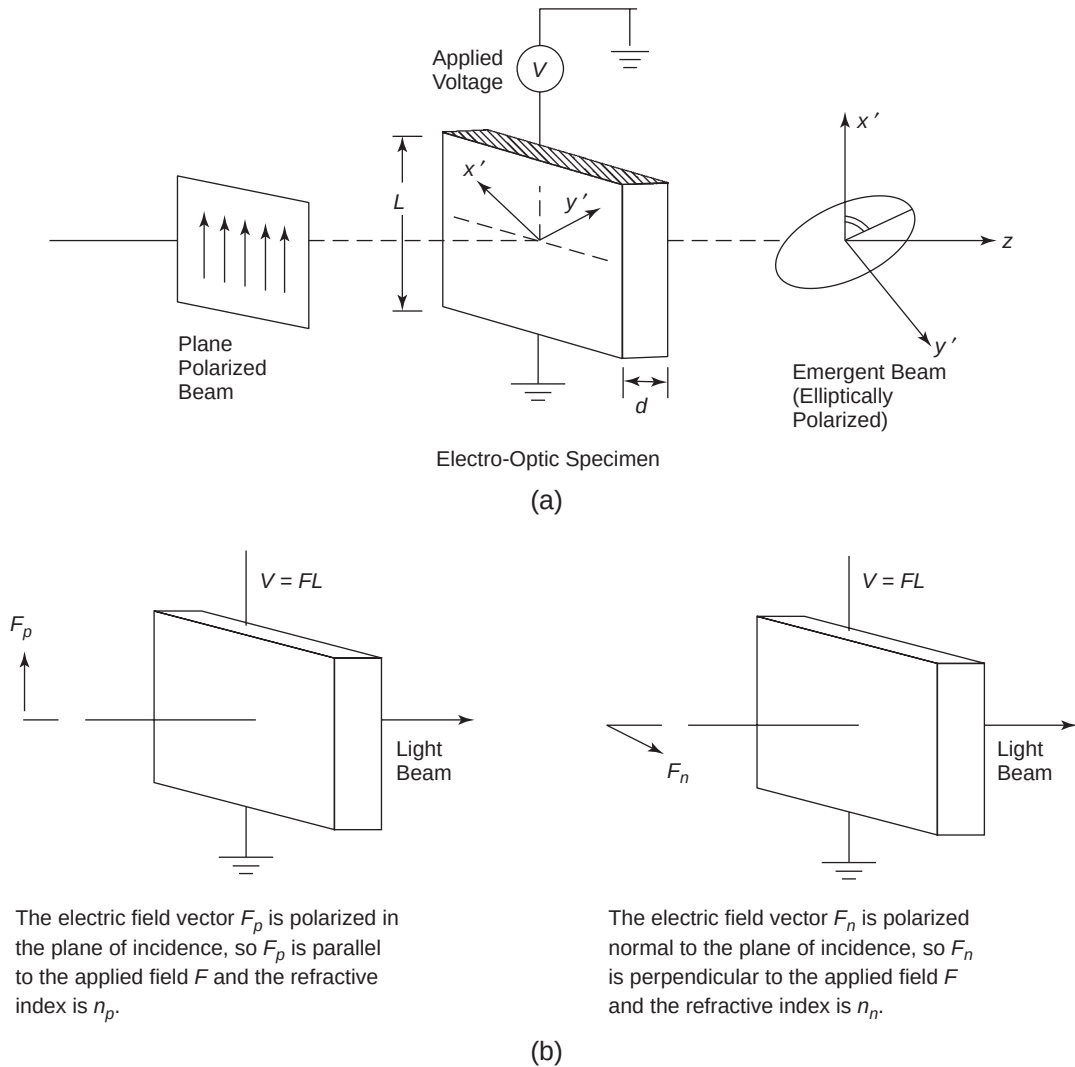


Figure 3-19 Schematic diagrams showing (a) the basic arrangement for a Kerr effect light modulator and (b) the definition of n_p and n_n .

the optic axis and that perpendicular to the optic axis. For more details about the Kerr effect, see references 25–28.

The electric field that induces optical anisotropy can be either a static or a time-varying field. In the former case, the Kerr effect can be used for light modulators or light switches; in the latter case, the Kerr effect can provide information about molecular reorientation. The Kerr effect can also be used as a method for measuring the spatial distribution

of space charges or electric field in fluids with high Kerr constants.²⁹ If the field for inducing the Kerr effect is in the optical frequency range, the effect is called optically induced birefringence. This effect has been theoretically studied and experimentally observed in several liquids by several investigators.^{30,31} This effect can be used for self-focusing of laser beams, because the intense field vector of the beams can enhance the refractive index.³² In polar materials, the optically induced birefringence is

smaller than the static field-induced birefringence for the same field intensity, since in the optical frequency fields, the orientation of the molecules is due only to the anisotropy of the polarizability, while in the static fields the orientation of the permanent dipoles becomes more important.

Liquid Kerr cells containing CS_2 , nitrobenzene, or other similar liquids have been used either as a light modulator or as a light switch. Practical Kerr cells usually employ one of the molecular liquids, since the reorientation of the molecules in those liquids under an applied electric field can produce a large change in refractive index. However, even though those liquids have a high value of the Kerr constant K , a high voltage of the order of 25 kV is sometimes required to produce sizable amplitude modulation. This is why most of electro-optic light modulators used with lasers are Pockels cells, which use one of the noncentrosymmetric crystals and produce birefringence through the second order nonlinearity, which induces a linear rather than a quadratic birefringence, as discussed earlier. Adequate modulation can be achieved in most applications with the modulation voltage of a few kV or less. The most important materials for Pockels cells are potassium dihydrogen phosphate (KDP) and ammonium dihydrogen phosphate (ADP), or Lithium niobate (LiNbO_3).

Electro-Optic Ceramics

Before closing this subsection, it is worth mentioning the electro-optic effect in ceramic materials. About 1960, highly transparent ceramics in the $\text{PbZrO}_3 - \text{PbTiO}_3 - \text{La}_2\text{O}_3$ (PLZT) system were developed.^{33–35} These materials have a remarkable facility for changing their polar state under an applied electric field. Since these materials are made of many compounds, a highly homogeneous mixture is necessary for electro-optic applications in order to avoid scattering due to local variation in composition and to ensure uniform electro-optic properties. This is why material fabrication processes are extremely important for producing PLZT ceramics for optical applications. For details

about ceramic fabrication processes, see references 36–38.

The optical transparency of the PLZT ceramics is highly dependent on the concentration of lanthanum oxide (La_2O_3) in the material. The specific concentrations of La that yield high optical transparency are dependent on the concentration ratio of Zr/Ti. For Zr/Ti = 65/35, the concentrations of La expressed in atomic percentage should be in the range of about 5 to about 16 atomic percentage. The type of hysteresis loop depends on the composition of the PLZT ceramics. For a Zr/Ti ratio of about 65/35 and a concentration of La of about 5 atomic percentage, the hysteresis loop is a square loop with a low coercive field, as shown in Figure 3-20(a), while for a Zr/Ti ratio of about 35/65 and a concentration of La of about 8 atomic percentage, the hysteresis loop is also a square loop, but with a high coercive field, as shown in Figure 3-20(b). By increasing the concentration of La to about 11 atomic percentage for the case of a Zr/Ti ratio of about 65/35, the hysteresis loop becomes a slim loop exhibiting field-enhanced ferroelectric behavior, as shown in Figure 3-20(c). The corresponding variations of the change in refractive index $\Delta n = (n_E - n_o)$, with the applied electric field for the three main types of characteristics, are also shown in Figure 3-20. The experimental data are from Haertling.^{36,39,40}

Ferroelectric materials are noncentrosymmetric and possess the property of birefringence. These materials have one optic axis. A light beam traveling in the direction of the optic axis with its electric field vector vibrating in the direction perpendicular to the optic axis has a phase velocity c/n_o , while a beam traveling in the direction perpendicular to the optic axis with its electric field vector vibrating in the direction parallel to the optic axis has a phase velocity c/n_E , in which n_o and n_E are the same as those defined for the ordinary and extraordinary rays. Let us go back to Figure 3-20. The hysteresis loop shown in (a) is suitable for memory applications (see Ferroelectric Materials and Application of Ferroelectrics in Chapter 4). However, the PLZT with the hysteresis loop

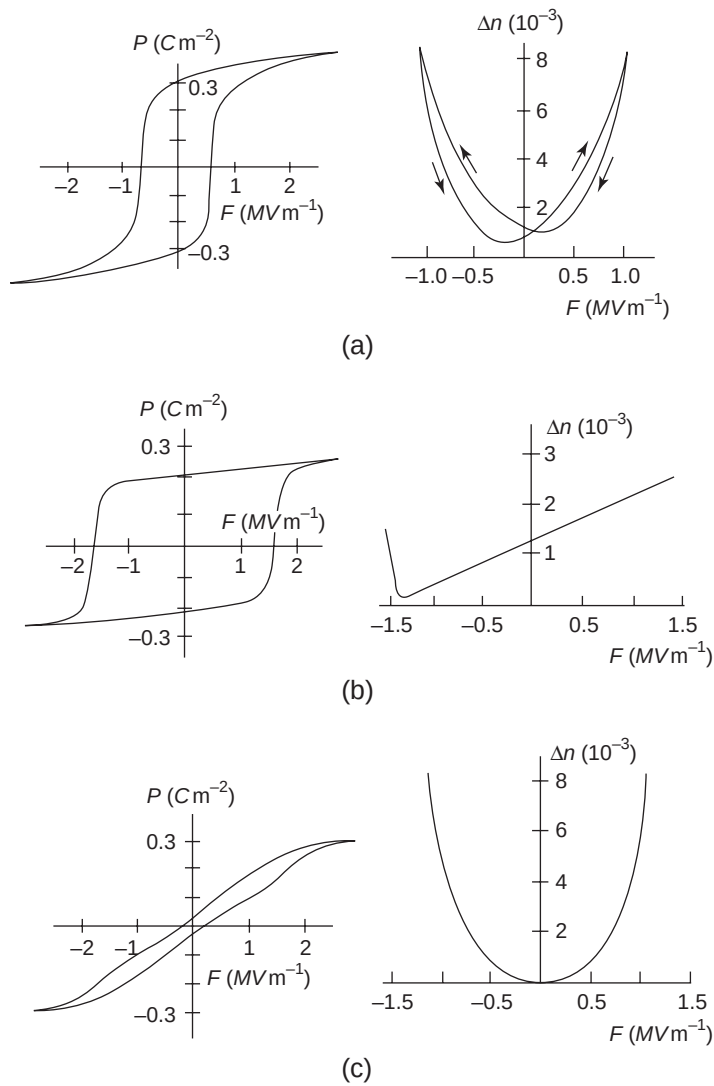


Figure 3-20 The hysteresis loops and the variation of the change in a refractive index with the applied electric field for (a) a square loop with a low coercive field, (b) a square loop with a high coercive field, and (c) a slim loop of PLZT ceramics.

shown in (b) exhibits a linear electro-optic effect similar to the Pockels effect. The composition of this PLZT system is rich in the concentration of PbTiO_3 . A high concentration of PbTiO_3 favors tetragonal distortion from the cubic structure, and a high tetragonality yields a high coercivity. The hysteresis loop shown in (c) is a slim loop; the PLZT exhibits a quadratic electro-optic effect similar to the Kerr effect.

The composition of this PLZT system lies close to the ferroelectric rhombohedral-tetragonal boundary (see the phase diagram for the PZT system, Figure 4-19, in Phenomenological Properties and Mechanisms in Chapter 4).

For more information regarding the properties and applications of electro-optic materials, the reader is referred to some excellent references, 36–43.

3.2.3 The Photorefractive Effect

Some materials' refractive index will change when they are subjected to light radiation. This phenomenon is referred to as the photorefractive effect. Several investigators discovered in 1966 that an intensive blue or green laser beam irradiating LiNbO_3 or LiTaO_3 crystals caused a change in refractive index of these materials.⁴⁴ They also found that such crystals, after their refractive index was changed by irradiation, could be returned to their original homogeneous and equilibrium state by being heated to about 200°C in a short period of time. Since then, many investigators have studied this effect in three directions: the mechanisms responsible for this effect, the materials exhibiting this effect, and the applications of this effect.⁴⁵⁻⁵⁰ The photorefractive effect simply refers to the optically induced changes of the refractive indices. This effect has been observed in a variety of electro-optic materials, including BaTiO_3 , KNbO_3 , CdS , etc. Depending on the band gap of the material, the refractive index changes may be induced not only by visible light but also by infrared or ultraviolet light.

The basic experimental arrangement for detecting the optically induced changes of refractive indices in crystals is shown in Figure 3-21(a). Both LiNbO_3 and LiTaO_3 have a large linear electro-optic effect and large spontaneous polarization of the order of $60\mu\text{C cm}^{-2}$.^{51,52} The laser beam is illuminating the specimen in the direction perpendicular to the c -axis (the polar axis of the crystal), and only a small region of the specimen is illuminated. For the specimen with a size of about 1.25 cm along the a -axis, 0.25 cm along the c -axis, and 0.25 cm along the b -axis, the diameter of the illuminating beam is about 10^{-2} cm. Usually, the incident beam is linearly polarized, with the plane of polarization at 45° from the c -axis of the crystal. The incident light beam in this case is found to split into three beams inside the crystal: one beam (the ordinary ray) does not change the direction, while the other two beams (extraordinary rays) are displaced from each other. The splitting of the light beam

indicates that the laser beam has caused the changes of refractive indices most for the extraordinary rays and least for the ordinary ray, resulting in $n_E < n_o$ for LiNbO_3 . The change of the refractive index $\Delta n = |n_E - n_o|$ is the highest at the center of the c -axis (i.e., the region in which the intensity of the laser beam is the highest) and Δn gradually decreases along both sides of the c -axis and the b -axis from the center, as shown schematically in Figure 3-21(b). The refractive index changes $\Delta n = |n_E - n_o|$ within the illumination region increase linearly with increasing laser beam intensity and also with increasing the light exposure time, as shown in Figure 3-22. The experimental results are from F.S. Chen.⁴⁶

When a crystal specimen is illuminated by a light beam, the illumination inside the specimen is not uniform. The amount of photo-generated charge carriers depends on the concentration of suitable donors present in the specimen, as well as the intensity of the light beam. Most of the crystals of interest are transparent to visible light, so electrons and holes excited from impurities present in the crystal are important. All charge carriers generated by light will diffuse from the region of high light intensity to the region of low light intensity, or to the region without illumination. These charge carriers may also move by drift due to the photovoltaic effect. The migration of the charge carriers will form space charges along the c -axis and the b -axis, and hence create an internal field (called a space charge field). This internal field causes the changes of refractive indices based on the electro-optic effect. The basic difference between the conventional electro-optic effect and the photorefractive effect lies in the fact that for the former, the change of the refractive index is caused by an externally applied electric field, while for the latter, it is caused by the internal field created by space charges and the photovoltaic effect due to the optical excitation.

However, for a large photorefractive effect, suitable donors, traps, and efficient charge carrier movement are essential. In crystals, traps are present due either to structural defects or to chemical defects (irregular composition or

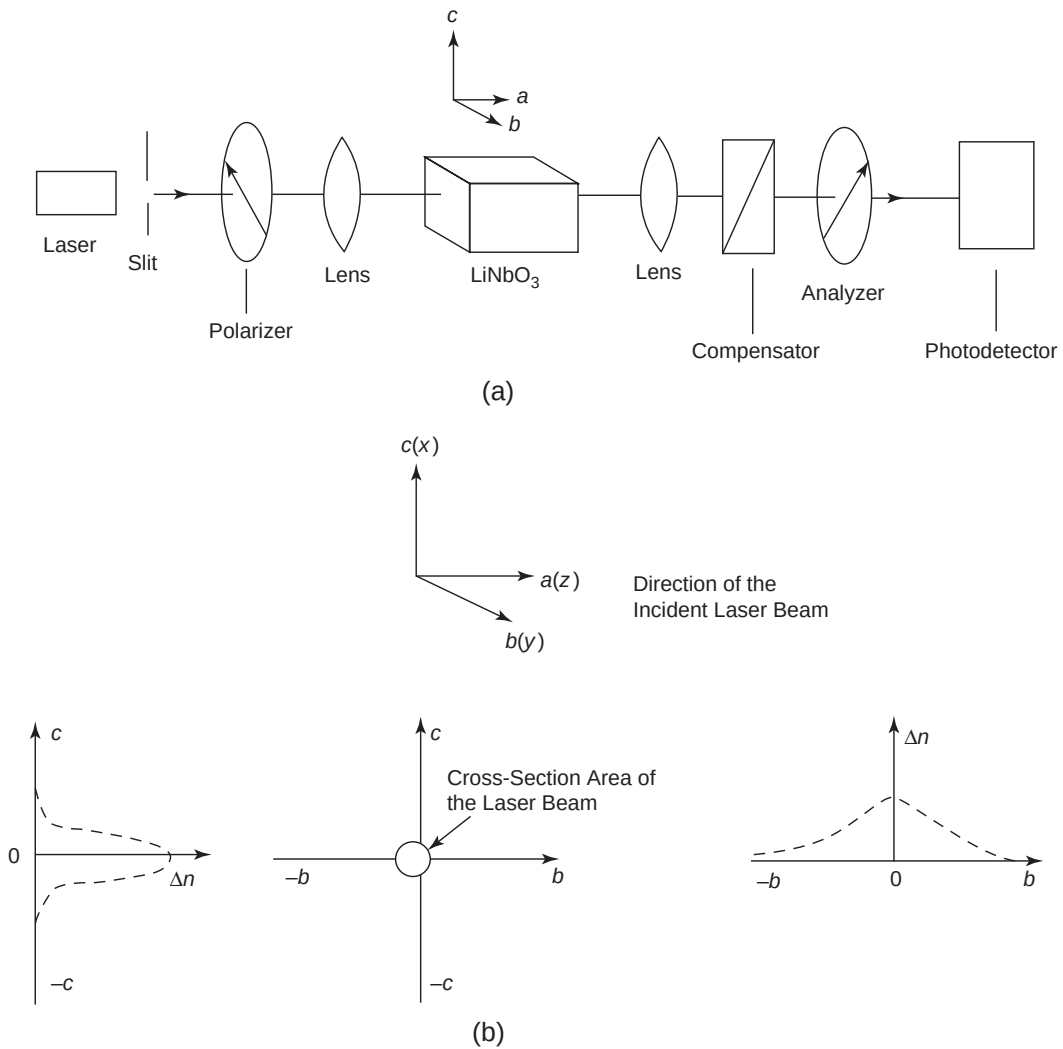


Figure 3-21 (a) The basic experimental arrangement for detecting the optically induced changes of refractive indices, and (b) the schematic diagrams showing the spatial variation of the changes of refractive indices along the c -axis and along the b -axis.

impurities). Materials doped with Fe impurities have a high photorefractive effect since Fe acts as a donor-acceptor trap via intervalence exchanges, such as $\text{Fe}^{2+} \rightleftharpoons \text{Fe}^{3+}$. The ion Fe^{2+} acts as a donor, and Fe^{3+} acts as an acceptor, which may act as an electron trap to capture an electron from the conduction band or as an acceptor to produce a hole in the valence band. The change of refractive index is more concentrated in the illuminating region along the c -axis, as shown in Figure 3-21(b). This

may indicate that a higher internal field may be created not only by space charges similar to the Dember photovoltaic effect, but also by anomalous photovoltaic effect, due to an asymmetric charge transfer process and Franck–Condon shifts of the excited ions along the polar axis (c -axis) of pyroelectric crystals, such as LiNbO_3 .⁴⁸ It has been reported that even in pure ferroelectric or electro-optic materials, electron-phonon coupling and collective Franck–Condon relaxation processes

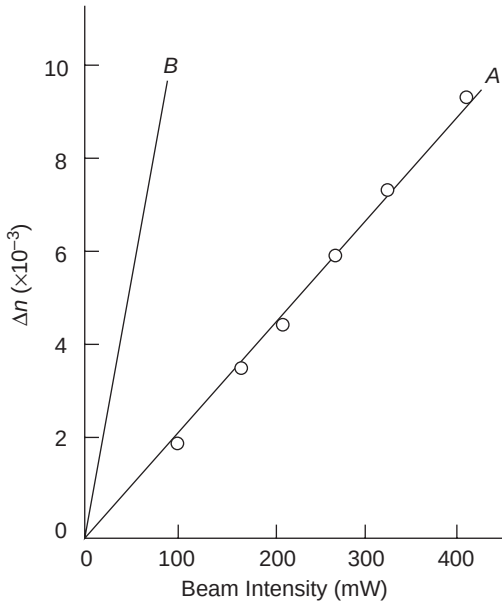


Figure 3-22 The change of the refractive index Δn of LiNbO_3 crystals as a function of the intensity of the illustrating laser beam. The beam diameter is 10^{-2} cm. The exposure time is 0.02 sec for A and 0.1 sec for B.

may produce an anomalous photovoltaic effect.^{53,54}

We will consider a simple case to illustrate the factors controlling the photorefractive effect, assuming that only electrons in the conduction band are mobile carriers, that the donors and other impurities or traps are immobile, and that the formulation is for the one-dimensional case (i.e., along the c -axis [x -axis]). The equations governing the behavior of the photorefractive effect are as follows.

The equation for the rate of photoionization is

$$\frac{\partial N_D^+}{\partial t} = SI(N_D - N_D^+) - RnN_D^+ \quad (3-77)$$

where N_D and N_D^+ are, respectively, the concentrations of the total donors and the ionized donors, n is the mobile electron concentration, R is the recombination coefficient, S is the cross-section of photo-ionization, and I is the incident light intensity expressed in the number of photons per unit area and per unit time (I expressed in mW cm^{-2} must be converted to the number of photons per cm^2 and per second by

multiplying it by $(h\nu)^{-1}$ or $(hc/\lambda)^{-1}$, where $h\nu$ is the energy of each photon).

The current flow equation is

$$\begin{aligned} j &= qun(F_{ph} + F_{sc}) + qD \frac{dn}{dx} \\ &= qun(F_{ph} + F_{sc}) + ukT \frac{dn}{dx} \end{aligned} \quad (3-78)$$

The continuity equation is

$$\frac{\partial \rho_{sc}(x, t)}{\partial t} = \frac{\partial j(x, t)}{\partial x} \quad (3-79)$$

and Poisson's equation is

$$\frac{dF_{sc}}{dx} = \frac{\rho_{sc}(x, t)}{\epsilon_r \epsilon_0} \quad (3-80)$$

where F_{ph} and F_{sc} are, respectively, the photovoltaic field and the space charge field; D is the diffusion coefficient, which is equal to ukT/q based on Einstein relation; $\rho_{sc}(x, t)$ is the space charge density, which is $q(n - N_D^+ + N_A^-)$; and N_A^- is the concentration of trapped electrons captured by traps other than the ionized donors, which also act as electron traps.

Even if we ignore the photovoltaic field F_{ph} , to obtain F_{sc} by solving Equations 3-77 through 3-80, we need to know the spatial distribution of N_D^+ , n , and u . It is much easier to solve this problem if we use a light source having a definite spatial distribution of its light intensity. In the following, we describe briefly a simple double-beam coupling system that can provide a stable spatial distribution of the incident light intensity. With two light beams of intensities I_{10} and I_{20} incident on the specimen surface at an angle θ from the a -axis, as shown in Figure 3-23, the spatial distribution of the resultant light intensity is given by

$$I(x, t) = I_o \exp(-\alpha x \cos \theta) [1 - m \cos Kx] \quad (3-81)$$

along the x -axis (c -axis) for the one-dimensional case, where α is the light absorption coefficient, $K = 2\pi/\Lambda$ is the spatial frequency, $\Lambda = \lambda/2 \sin \theta$ is the interference fringe spacing, $I_o = I_1 + I_2$, and I_1 and I_2 are, respectively, the light intensities of the two incident interfering beams (I_{10} and I_{20}) with a modulation index m .^{49,55} The value of m is given by

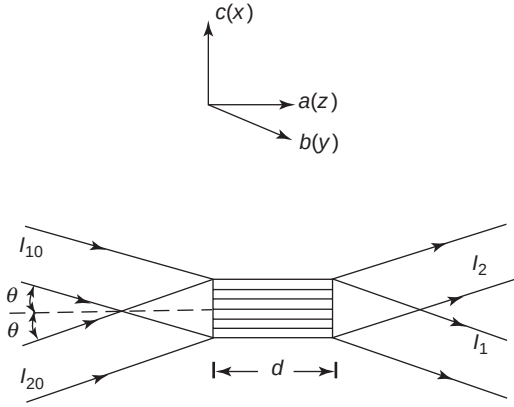


Figure 3-23 A double-beam coupling system for producing a stable spatial distribution of light intensity.

$$m = \left[\frac{2(I_1 I_2)^{1/2}}{I_1 + I_2} \right] \cos 2\theta \quad (3-82)$$

With this double-beam coupling system, the generation of photo-induced electrons can be considered to be uniform over the crystal length along the z direction. The generation rate of photo-induced electrons along the x direction (c -axis) can be written as

$$g(x) = g_o(1 + m \cos Kx) \quad (3-83)$$

where g_o is given by

$$g_o = \frac{\beta \alpha I_o}{h\nu} = \frac{\beta \alpha I_o \lambda}{hc} \quad (3-84)$$

where β is the quantum efficiency for exciting a charge carrier from the donor.

Using the double-beam coupling system, several investigators have solved Equations 3-77 through 3-80 for the space charge field F_{sc} . By ignoring the contribution from the anomalous photovoltaic effect, the space charge field along the c -axis (x -axis) is given by

$$F_{sc} = m \frac{kT}{q} \left[\frac{K}{(1 + K/K_o)^2} \right] \cos 2\theta \quad (3-85)$$

where

$$K_o^2 = \frac{qN_{D(av)}^+ [N_D - N_{D(av)}^+]}{qkTN_D} \quad (3-86)$$

and $N_{D(av)}^+$ is the mean value of N_D^+ distributed along the x -axis.^{49,56} Once F_{sc} is known, the

change of refractive index for the linear electro-optic crystals induced by the photorefractive effect can be estimated by

$$\begin{aligned} \Delta n_o &= -\frac{n_o^3 r_{13} F_{sc}}{2} \\ \Delta n_E &= -\frac{n_E^3 r_{33} F_{sc}}{2} \end{aligned} \quad (3-87)$$

These expressions are similar to that for the Pockels effect. The electro-optic coefficients r_{13} and r_{33} are related to the relative direction of the O ray and that of the E ray to the direction of F_{sc} . For most crystals, $r_{33} > r_{13}$. For example, $BaTiO_3$ has $r_{13} = 10 \text{ pmV}^{-1}$ and $r_{33} = 93 \text{ pmV}^{-1}$.^{57,58}

Since the discovery of the photorefractive effect, a great deal of effort has been devoted to insight into the microscopic mechanisms responsible for this photorefractive phenomenon and the factors controlling its behavior under various conditions. Details about developments in this area are beyond the scope of this book. However, before closing this subsection, we would like to mention some general information about the materials and applications. Up to the present, the materials exhibiting the photorefractive effect may be placed into three groups:

Group 1: This group includes mainly ferroelectric materials. The oxygen-octahedral materials, particularly those containing Fe, exhibit the largest electro-optic effect, such as LiNbO_3 , LiTaO_3 , KNbO_3 , and BaTiO_3 . Highly transparent ferroelectric ceramics of the PZT family, with the perovskite structure with part of lead ions substituted by lanthanum ions, form a $(\text{Pb}_{1-x}\text{La}_x)(\text{Zr}_y\text{Ti}_{2-y})\text{O}_3$ or PLZT system.

Group 2: This group includes mainly nonferroelectric or paraelectric materials, such as $\text{Bi}_{12}\text{SiO}_{20}$, and $\text{Bi}_{12}\text{GeO}_{20}$. They possess electro-optic and photoconductive properties.

Group 3: This group includes mainly semiconducting materials such as GaAs, InP, and CdTe. These materials are transparent up to the infrared frequency region, so they exhibit the photorefractive effect in the infrared frequency range.

To select material for a specific application, it is important to consider the following factors:

- Photorefractive sensitivity
- The range of the refractive index changes
- The phase shift between the refractive index and light intensity distributions
- Photorefractive signal recording time and erasure time
- Spatial frequency dependence
- Electric field dependence
- Light wavelength (i.e., photon energy) for inducing refractive index changes
- Resolution
- Signal-to-noise ratio
- Operation temperature

Similar to the electro-optic effect, there are a variety of applications of the photorefractive effect.^{49,59} Here we just mention a few because of the limited space of this subsection. Some important applications are volume hologram storage,^{47,49,60} real-time holography and interferometry,^{61,62} real-time information processing,^{63,64} and coherent light amplification.^{49,65}

3.2.4 The Magneto-Optic Effect

In general, any method that can change the refractive index of a crystal can be used to deflect a light beam and hence to modulate a light beam. Obviously, the electro-optic effect is a more efficient and convenient means for modulating light than the magneto-optic effect. The magnetic field may also change the optical properties of some materials, but the handling of an electric field is much easier than the handling of a magnetic field, so any optical devices based on the electro-optic effect are preferred to those based on the magneto-optic effect.

The Faraday Effect

Faraday discovered in 1848 that when a plane-polarized light beam passed through a glass plate in the direction of an applied magnetic field, the plane of polarization was rotated.^{66,67}

The amount of rotation in angle is proportional to the magnetic field H and is given by

$$\theta = VHL \quad (3-88)$$

where L is the path length in the material and V is the so-called Verdet constant, whose dimension is the rotation per unit path length and per unit magnetic field.

The Faraday effect is usually small and depends on the wavelength of the light beam. For example, the rotation in quartz at a magnetic field of 10^4 oersteds and the wavelength of the light beam of $5900\text{\AA}$ is only about $2^\circ 46'$ for a path length of 1 cm. A plane-polarized light beam will become circularly polarized when the crystal is subjected to a strong magnetic field along the direction of the light beam propagation. Depending on the electric field vector direction, the circularly polarized wave can be a right-hand circularly polarized (RHCP) wave or a left-hand circularly polarized (LHCP) wave. As a result, the two counter-rotating circularly polarized waves propagating along the direction of the magnetic field in the z direction have different phase velocities, c/n_R and c/n_L , where n_R and n_L are, respectively, the refractive indices for the RHCP and the LHCP waves. So θ can also be expressed in terms of n_R and n_L as

$$\theta = \frac{2\pi}{\lambda}(n_R - n_L)L \quad (3-89)$$

It can be seen from Equations 3-88 and 3-89 that $n_R - n_L$ depends on H . Thus, by varying the magnetic field H , we can modulate the light beam for various applications.

The Voigt Effect

Voigt found in 1905 that a magnetic field may also induce birefringence for a plane-polarized light beam passing through a crystal in the direction perpendicular to the magnetic field. The component of a linear polarized light beam with the electric field vector parallel to the magnetic field travels at a phase velocity different from that with its electric field vector perpendicular to the magnetic field, thus causing the emerging light beam from the crystal to become

elliptically polarized.^{68,69} We may consider the Faraday effect as the magnetic counterpart to the Pockels effect and the Voigt effect as the magnetic counterpart to the Kerr effect.

3.2.5 The Acousto-Optic Effect

The refractive index of a medium can be changed by the mechanical strains caused by the passage of an acoustic wave through the medium. The relation between the change of the refractive index and the change of the mechanical strain and stress is rather complicated.⁷⁰ However, we may explain this effect from the Lorentz–Lorenz equation (see Chapter 2, Equation 2-192), which is given by

$$\frac{n^2 - 1}{n^2 + 2} = \frac{N\alpha}{3\epsilon_0} \quad (3-90)$$

The basic mechanism responsible for the change of the refractive index is the change of the density of the molecules by ΔN due to the mechanical strain produced by the acoustic wave. So we can write

$$\frac{(n - \Delta n)^2 - 1}{(n - \Delta n)^2 + 2} = \frac{(N - \Delta N)\alpha}{3\epsilon_0} \quad (3-91)$$

From Equations 3-90 and 3-91, the change of the refractive index may be expressed as⁷¹

$$\frac{6n\Delta n}{(n^2 + 2)^2} = \frac{\Delta N\alpha}{3\epsilon_0} \quad (3-92)$$

or expressed in the form of

$$\Delta n = \frac{n^3}{2} \left[\frac{(n^2 - 1)(n^2 + 2)}{3n^4} \right] \frac{\Delta N}{N} \quad (3-93)$$

The acoustic wave causes only the locally elastic change in the density of the molecules in the region where the wave passes through. There is no change in the total mass of the medium, so the acoustic wave causes only a fractional change in volume of the medium ΔV . So, we can express the tensile strain h as

$$h = \frac{\Delta V}{V} = \frac{\Delta N}{N} \quad (3-94)$$

The photoelastic constant is given by

$$p = \frac{(n^2 - 1)(n^2 + 2)}{3n^4} \quad (3-95)$$

Substitution of Equations 3-94 and 3-95 into Equation 3-93 yields

$$\Delta n = \frac{n^3 p h}{2} \quad (3-96)$$

The change of the refractive index Δn is directly proportional to the mechanical strain h .

The acousto-optic effect is sometimes referred to as the elasto-optic effect or the photoelastic effect, referring mainly to the change induced in the relative dielectric impermeability tensor by mechanical strain.⁷¹ The relative dielectric impermeability tensor is defined as $B_{ij} = (\epsilon_r^{-1})_{ij}$, which is a second-rank tensor. The impermeability B_{ij} is symmetrical, so $B_{ij} = B_{ji}$. For a weak strain, such as that caused by an acoustic wave, the change of the tensor can be considered as a linear function of strain. However, since the relative dielectric impermeability tensor and the strain tensor each involves nine tensor components, the solution of this problem for the acousto-optic effect would be quite complicated. Because it is beyond the scope of this book, we shall not discuss the mathematical involvement further.

When a light beam enters an elastic medium in which an acoustic wave has already produced a sinusoidal variation of elastic pressure in the medium, the portion of the light wavefronts near the pressure maxima will encounter a higher refractive index and hence travel at a lower phase velocity than the portion of the wavefronts near the pressure minima. Thus, the light wavefronts in the medium exhibit a wavy variation according to the variation of the acoustic wave. The velocity of an acoustic wave (i.e., a sound wave) is much lower than that of a light wave. So, we may ignore the small wave variation of the wavefront and consider the variation of the refractive index to be stationary in the medium.

In general, acousto-optic modulation cells are produced by means of moving gratings in an elastic medium produced by the strain accompanying acoustic waves. The light beam passing through the gratings will suffer diffraction; the light beam will then be spatially deflected, and the deflected beams have phases different from each other. There are two types

of diffraction: Raman–Nath diffraction and Bragg diffraction. For Raman–Nath diffraction, the grating is thin so that the light beam will be diffracted producing deflected beams. The amplitude of the deflected beam depends on the deflection angle from the direction of the incident light beam. If only the deflected beam in the same direction as that of the incident beam is considered, then the modulated beam will follow the modulating signal directly. For Bragg diffraction, we consider a plane-polarized light beam incident on the grating planes at an angle of incidence θ in a manner very similar to the Bragg diffraction of x-rays from the planes (layers) of atoms in a crystal. The grating planes can be considered stationary acoustic wavefronts, so the incident beams are scattered from the successive planes. Constructive interference is formed by mixing the beams reflected from each of the planes (the grating planes imply a thick grating). So, the amplitude of the modulated beam is directly related to the modulating signal. Here, we have used a simple approach to describe the basic concept of the interaction between light waves and acoustic waves. For more details, the reader is referred to some excellent references, 5 and 71–74.

3.3 Interaction between Radiation and Matter

The interaction between radiation and matter deals primarily with the processes of energy exchange between radiation and atoms. There are discrete energy levels in an atomic (or molecular) system. The system can make a transition from a state at a lower level E_1 to a state at a higher level E_2 by the absorption of a photon of energy $\Delta E = E_2 - E_1 = h\nu$. The system can also reverse the process; that is, an atom in a state of energy level E_2 may change to a state of lower energy level E_1 by the emission of a photon of the same energy $\Delta E = E_2 - E_1$. The optical spectra are associated with the motion of the outermost valence electrons of atoms or molecules and with the motion of the atoms or molecules, which occur in the microwave, infrared, visible, or ultraviolet frequency regions. The optical spectra provide informa-

tion about the behavior of the outermost valence electrons of an atom,^{75–77} while X-ray spectra give information about the inner electrons in the atom.^{77,78} This is a huge subject. In the limited space of this chapter, we shall deal only briefly with the topics related to dielectric phenomena.

3.3.1 Generation of Radiation

Radiation generated from candlelight, the fire in a fireplace, or sunlight is common in our daily lives, and we often overlook its presence and importance. But life could not exist without radiation.

There are two outstanding processes by which a material can generate radiation after absorbing suitable extraneous primary energy. The energy thus given to the electrons in solids may be transformed into thermal energy by collisions with neighboring atoms or molecules, resulting in thermal radiation. The energy given to the electrons may also be used to raise them from a state at a relatively lower energy level to another state at a relatively higher energy level, resulting in luminescent radiation.

Thermal Radiation

Radiation can emerge from a small hole in a highly polished cavity, as shown in Figure 3-24(a). It should be noted that almost every heated metal, such as tungsten filament, produces a radiation spectrum. The spectral radiation pattern depends on the radiator temperature. The ability of a body to radiate is closely related to its ability to absorb radiation. This is to be expected, since a body at a constant temperature in thermal equilibrium with its surroundings must absorb energy from those surroundings at the same rate as it emits energy. Generally, we consider an ideal body that absorbs all radiations incident upon it, regardless of frequency. Such a body is called a blackbody. The cavity shown in Figure 3-24(a) is a good approximation of an ideal blackbody. A small fraction of the thermal radiation can escape through a small hole in the wall of the cavity. Both the density and the wavelength of the emerging radiation can be measured by

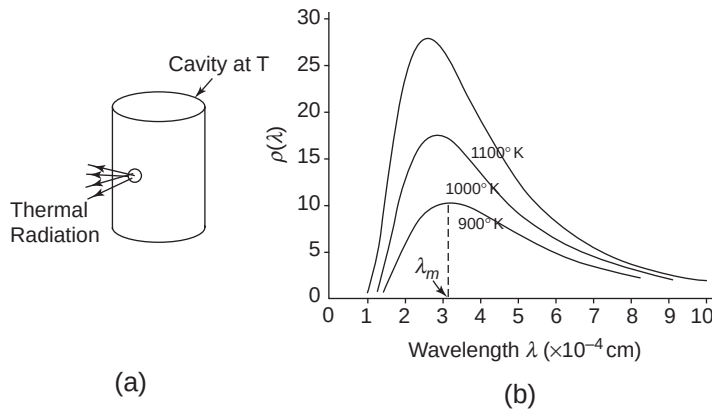


Figure 3-24 (a) The cavity approximates an ideal blackbody showing schematically the thermal radiation emitted from a small hole on the wall of the cavity, and (b) the distribution function of the radiation energy as a function of wavelength.

standard apparatus. In order to obtain emission of thermal radiation from a solid in the visible frequency region, the temperature of the solid must be raised to a temperature higher than 900 K to achieve an appreciable probability of getting 1.7 eV (or greater) for electronic excitation by the accumulative action of phonons.

The thermal radiation of a solid approximates closely the thermal radiation of an ideal black body. Thus, the total energy (for all wavelengths) emitted in unit time per unit area of the wall of the cavity, according to Stefan's law,^{3,79} is given by

$$E = \alpha T^4 \quad (3-97)$$

where α is Stefan's constant, which is $5.67 \times 10^{-8} \text{ Wm}^{-2}\text{K}^{-4}$ for an ideal blackbody. Most hot material surfaces only approach the ideal; therefore, α would be smaller than this value.

The distribution of energy with wavelength can be described by a function $\rho(\lambda)$. This means that the total energy, with a wavelength range between λ and $\lambda + d\lambda$, is $\rho(\lambda)d\lambda$. This function $\rho(\lambda)$, according to Planck's radiation law, is given by

$$\rho(\lambda) = \frac{8\pi hc}{\lambda^5} \left[\frac{1}{\exp(hc/\lambda kT) - 1} \right] \quad (3-98)$$

or in term of frequency ν instead of wavelength

$$\rho(\nu) = \frac{\lambda}{\nu} \rho(\lambda) = \frac{8\pi h \nu^3}{c^3} \left[\frac{1}{\exp(h\nu/kT) - 1} \right] \quad (3-99)$$

The radiation energy distribution $\rho(\lambda)$ as a function of wavelength for three different temperatures is shown in Figure 3-24(b). It can be seen that $\rho(\lambda)$ has a peak value at λ_m , and the wavelength for the occurrence of the peak is inversely proportional to temperature T , according to Wien's displacement law, given by

$$\lambda_m T = \text{constant} \quad (3-100)$$

Luminescent Radiation

Strictly speaking, luminescence is a process whereby matter generates nonthermal radiation. This implies that luminescence affords photon emission in excess of that produced entirely by thermal radiation. Luminescence is generally produced by the excitation of a solid by primary photons or charge particles in a material having individual energies ranging from about 1 eV to over 10^6 eV and affording emitted photons with energies in excess of 1 eV. The luminescence process as a whole includes excitation, temporary storage of the energy, and radiation (or emission). Various processes for producing luminescence will be discussed in Section 3.4.

Units of Light

In Gauss's law in Chapter 1, we discussed the fundamental and derived units of the SI system. Here we describe just briefly the units for light.

The method of measuring the energy of electromagnetic radiation, when all frequencies involved are treated equally, is called radiometry. The watt is a radiometric unit. Light is a form of energy, and radiant energy and power are measured in joules and watts, respectively. Radiant power is sometimes referred to as flux. The amount of flux per unit area delivered to a surface is the irradiance I . In fact, this unit is often referred to as radiation intensity. In SI units, however, it is called irradiance and expressed in watts. If the light is from a point light source, the irradiance I is not expressed in watts alone because the irradiance measured at point P depends on the distance r from point P to the point light source, so it must be expressed in watts per unit solid angle: radiant power per unit area within a unit solid angle (i.e., watts per steradian [W/sr⁻¹]).

Photometry was the method used to measure the radiation energy or power in the past. The fundamental difference between radiometry and photometry is that in radiometry, the measurements are made with electronic instruments, while in photometry, the measurements are based on the human eye. Therefore, different units are used in photometry for radiant power. The unit is lumens (lm) instead of watts. The unit for power density is lux (lumens per square meter) instead of watts per square meter. In this book, we use mainly the SI units based on radiometry.

3.3.2 The Franck–Condon Principle

The concept of the configuration diagram, based on the Franck–Condon principle, is useful in explaining the nonradiative transition. In the Born–Oppenheimer approximation, the total wave function of a vibronic state Ψ can be expressed as the product of electronic wave function ϕ_e and vibrational wave function ϕ_v . Thus, we may write

$$\Psi_{\ell i} = \phi_{\ell e}(q, r) \phi_{\ell i}(r) \quad (3-101)$$

as the total wave function of the ℓ_i vibronic state, where i indicates the i th vibrational state of the lower electronic state ℓ , and

$$\Psi_{mj} = \phi_{me}(q, r) \phi_{mj}(r) \quad (3-102)$$

as the total wave function of the mj vibronic state, where j indicates the j th vibrational state of the higher electronic state m , and q and r are, respectively, the electronic and nuclear coordinates. According to the Franck–Condon principle, there are no changes in the nuclear coordinates during an electronic transition, and therefore, the most probable vibronic transitions are vertical because the time (less than 3×10^{-14} sec) required for an electronic transition is negligibly small compared to the period of nuclear vibration. Electronic transitions are most probable when the kinetic energies of the nuclei are minimal (that is, at the end points of their vibrations; for example, at a_2c_2 and $a_2^*c_2^*$ in Figure 3-25). The lowest vibrational levels may be possible exceptions to this principle.⁸⁰ If the molecule is in its ground state E_ℓ with vibrational energy corresponding to lowest sublevel a_0c_0 , as shown in Figure 3-25, the electron may be excited from this level to state E_m at vibrational-level end point a_2^* by absorption of a photon with energy $E_A = h\nu_A$. The excited molecule then vibrates between a_2^* and c_2^* and may emit phonons and relax to sublevel $a_0^*c_0^*$. Then, it may proceed back to state E_ℓ at sublevel a_2c_2 by emitting a photon of energy $E_B = h\nu_B$. The molecule can relax to its lowest sublevel $a_0^*c_0^*$ by emitting phonons.

The energy emitted E_B is lower than the energy absorbed E_A . The difference between these two energies is called the Franck–Condon shift. In general, the degradation of optical energy is called the Stokes shift. Thus, the Franck–Condon shift is a Stokes shift due to the displacement of the molecule Δr . The absorption spectrum depends on the relative position of the minimum potential energy level and on whether the molecules are in vapor or in solution, as shown in Figure 3-25.

3.3.3 Radiative and Nonradiative Transition Processes

The emission of radiation is the inverse of the absorption process. An electron occupying a quantum state at a level relatively higher than it would under thermal equilibrium conditions, will tend to make a transition to an empty state

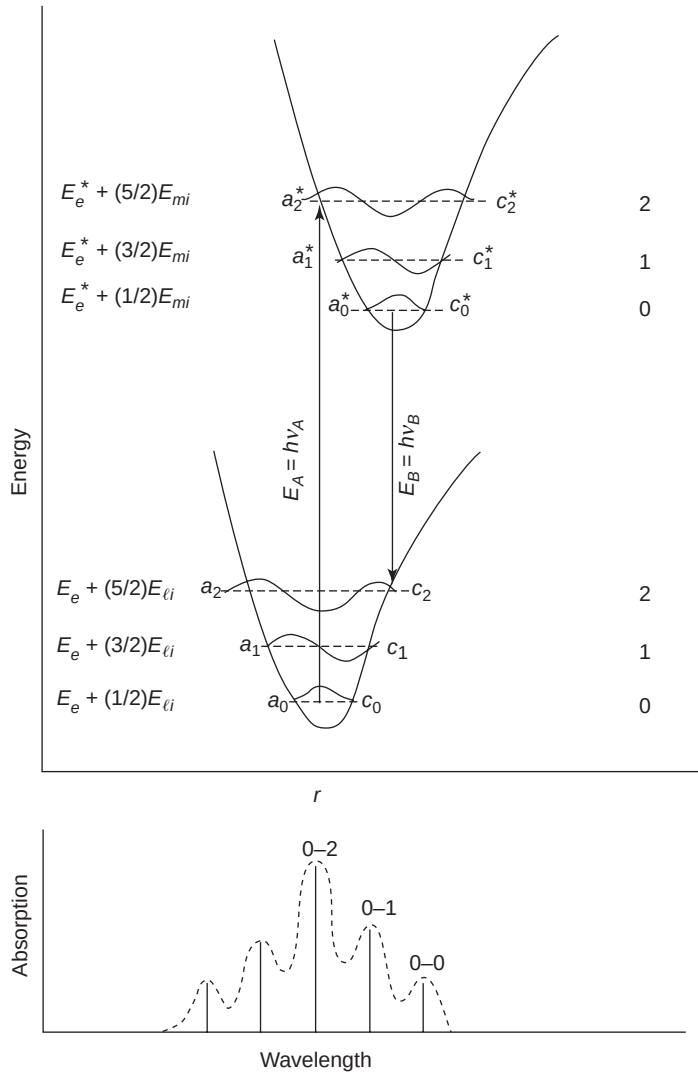


Figure 3-25 Franck-Condon configuration diagram and absorption spectrum for a diatomic molecule. In the absorption curves, the dashed curve is for molecules in solution and the solid lines are for molecules in vapor form. E_A is the absorption at the Franck-Condon maximum, and E_B is the emission at the Franck-Condon minimum; r is the nuclear separation.

at a lower energy level, and all or most of the energy released due to this transition will be emitted as electromagnetic radiation. The radiative transition rate is determined by the product of the density of the occupied upper states and the density of the empty lower states, and also the probability P per unit volume for one carrier in the upper state to make a radiative transition to one lower empty state. Thus, the radiative transition rate can be written as

$$R = P n_u n_\ell$$

where n_u and n_ℓ are, respectively, the densities of the occupied upper states and the empty lower states.

The Franck-Condon configuration diagram indicates that the phonon emission process may accompany radiative transition when the electron is excited from the ground state at the a_0 level to the excited state at the a_2^* level; the

system will then partly relax to the a_o^* before making the transition to return to its ground state, as shown in Figure 3-25. This involves a displacement Δr of the atom. The excited electron loses some energy in this process, and the energy thus lost is dissipated in the atomic displacement, that is, in the form of phonon emission. When the electron returns to the ground state at the c_2 level, a new atomic displacement will take place for $c_2 \rightarrow c_o$, again in the form of phonon emission.

However, whether the transition made by an electron from an excited state at level m to a state at a lower level ℓ is radiative or non-radiative depends on the transition moment (or matrix elements of the electric dipole moment) between two states $\psi_{\ell i}$ and $\psi_{m j}$ given by Equations 3-101 and 3-102, based on the Franck-Condon principle. The transition moment can be expressed as

$$\begin{aligned} M_{mj \rightarrow \ell i} &= \langle \psi_{mj} | M | \psi_{\ell i} \rangle \\ &= \langle \phi_{me} | M | \phi_{\ell e} \rangle \langle \phi_{mi} | \phi_{\ell i} \rangle \\ &= |\bar{M}_{m\ell}| S_{mj, \ell i} \end{aligned} \quad (3-103)$$

where M is the electric dipole operator, $\bar{M}_{m\ell}$ is the mean electronic transition moment, and $S_{mj, \ell i}$ is the vibrational overlap integral. The Franck-Condon maximum, which represents a vertical transition on the configuration diagram, corresponds to a maximum overlap integral ($|S_{mj, \ell i}|$). The probability of the radiative transition from a state corresponding to wave function ψ_{mj} to a state corresponding to wave function $\psi_{\ell i}$ is proportional to the square of the electronic transition moment

$$\begin{aligned} p_{m \rightarrow \ell}^r &\propto |M_{mj \rightarrow \ell i}|^2 \\ &= |\bar{M}_{m\ell}|^2 |S_{mj, \ell i}|^2 \\ &= A_{m \rightarrow \ell} F \end{aligned} \quad (3-104)$$

where F is the Franck-Condon factor, which is equal to $|S_{mj, \ell i}|^2$; $A_{m \rightarrow \ell}$ is Einstein's coefficient of spontaneous emission, which is equal to $|\bar{M}_{m\ell}|^2$ and is given by

$$A_{m \rightarrow \ell} = (8\pi h \nu_{mj \rightarrow \ell i}^3 \eta^3 c^{-3}) B_{m \rightarrow \ell} \quad (3-105)$$

in which h is the Plank constant, η is the refractive index of the medium where a molecule undergoes a transition, c is the light

velocity, $\nu_{mj \rightarrow \ell i}$ is the radiative frequency corresponding to the energy difference between the states $m j$, and ℓi , $B_{m \rightarrow \ell}$ is Einstein's coefficient for induced absorption or induced emission.^{70,81}

The probability of nonradiative transition from a state of wave function ψ_{mj} to a state of wave function $\psi_{\ell i}$ is proportional to the electronic factor

$$\begin{aligned} p_{m \rightarrow \ell}^{nr} &\propto |H_{mj \rightarrow \ell i}|^2 \\ &= |J_{m\ell}|^2 F \end{aligned} \quad (3-106)$$

where

$$|H_{mj \rightarrow \ell i}| = \langle \phi_{mj} | J_{m\ell} | \phi_{\ell i} \rangle \quad (3-107)$$

in which

$$|J_{m\ell}| = \langle \phi_{me} | J_N | \phi_{\ell e} \rangle \quad (3-108)$$

where $|J_{m\ell}|^2$ is the electronic factor and J_N is the nuclear kinetic energy operator.⁸¹

Nonradiative transition processes always compete with radiative transition processes, and are therefore very important in energy transfer processes and luminescent phenomena. Unfortunately, these processes are poorly understood. Both $|M_{m\ell}|^2$ and $|J_{m\ell}|^2$ involve the wave functions of the initial and the final electronic states, so nonradiative transitions are subject to the same multiplicity, symmetry, and parity selection rules as radiative transitions.⁸¹ So far, we have discussed only the transitions in a completely isolated molecule. But the analysis of a solid containing interreacting molecules or atoms involves a complete Hamiltonian and would therefore become formidable. As our aim is to understand the basic concept of radiative and nonradiative transitions, we will use a simple model to show the rules governing the transition processes. We consider one electron in a one-dimensional box, whose behavior is governed by the time-dependent Schrodinger equation. In the box, there are quantum states of various discrete energy levels. Let us look into the behavior of this electron between any two states: one state n at an energy level higher than the other state, called state m . When the electron in state n makes a transition to state m , starting at time $t = 0$, how does this electron move during the time inter-

val ($0 < t < \tau$), where τ is the time required for the electron to reach state m ? By solving the Schrodinger equation for this simple case, it can be shown that the wave function $\psi(x,t)$ of this electron during the time interval of transition is given by

$$\psi(x,t) = \psi_n \cos \frac{\pi t}{2\tau} + \psi_m \sin \frac{\pi t}{2\tau}, \quad 0 < t < \tau \quad (3-109)$$

where ψ_n and ψ_m are, respectively, the wave functions of the electron in state n and in state m . They are

$$\psi_n = \sqrt{\frac{2}{L}} \sin \frac{n\pi x}{L} \exp(jE_n t/\hbar) \quad (3-110)$$

$$\psi_m = \sqrt{\frac{2}{L}} \sin \frac{m\pi x}{L} \exp(jE_m t/\hbar) \quad (3-111)$$

The electron is in state n at $t = 0$, and in state m at $t = \tau$. The average value (or the expectation value) of the energy for the electron is

$$\begin{aligned} \langle E \rangle &= \int_0^L \bar{\psi} j \hbar \frac{\partial}{\partial t} \psi dx \\ &= E_m + (E_n - E_m) \cos \frac{\pi t}{2\tau} \end{aligned} \quad (3-112)$$

where $\bar{\psi}$ is the complex conjugate of ψ . Similarly, the average position (or the expectation position) of the electron in the box is

$$\begin{aligned} \langle x \rangle &= \int_0^L \bar{\psi} x \psi dx \\ &= \frac{L}{2} + \frac{4L}{\pi^2} \sin \frac{\pi t}{2\tau} \cos \frac{(E_n - E_m)t}{\hbar} \\ &\quad \times \frac{4nm}{(n^2 - m^2)^2} \end{aligned} \quad (3-113)$$

where m and n are integers signifying the energy levels of the quantum states.

By examining Equations 3-112 and 3-113, it can be concluded that when $n - m$ is an odd number, the position of the electron $\langle x \rangle$ oscillates about its average position at $\langle x \rangle = L/2$, and the oscillation frequency is $(E_n - E_m)/\hbar$ during transition, as shown in Figure 3-26. This implies that the transition generates photons and is thus radiative, producing luminescence. Based on the classical theory of radiation, an electron oscillating in space will radiate electromagnetic waves.

If $n - m$ is an even number, $\langle x \rangle = L/2$, implying that there is no time change of the electron position during transition. This means that the transition generates phonons whose energy is dissipated as energy loss to the lattice vibration. The above model, though very simple, shows the transitions typical to electric dipole transitions governed by the selection rules.

The performance of a laser depends on a particle (electron, atom, ion, or molecule) chang-

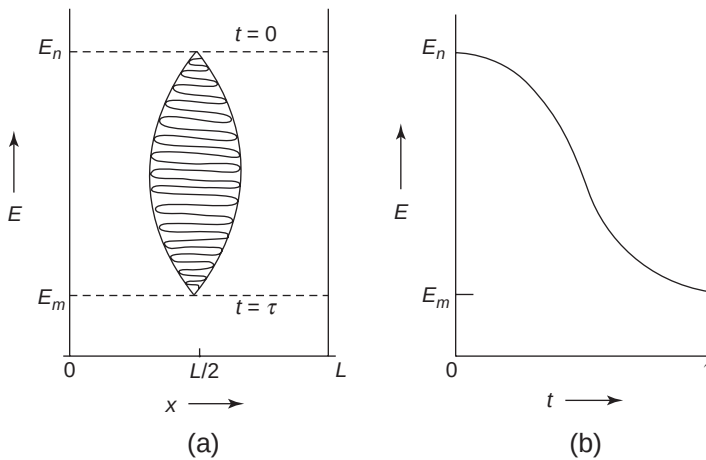


Figure 3-26 Schematic diagrams showing (a) the displacement (oscillation) of an electron in a one-dimensional box during the transition from state n to state m , and (b) the average energy of the electron in the box during transition from $t = 0$ at state n to $t = \tau$ at state m .

ing its energy state; in so doing, a photon can be generated. There are two basic types of solid-state lasers: one consists of rather well-isolated ions embedded in a host crystal, and the other is generally a semiconductor. The former can be considered a molecule with tightly bound electrons having discrete energy levels, somewhat similar to the simple model just discussed.

In general, there are five processes for radiative transitions in solids⁸²:

1. Band-to-band transitions
2. Transitions via shallow donor or acceptor levels
3. Donor-acceptor transitions
4. Transitions via deep recombination centers
5. Exciton transitions

In inorganic semiconductors, processes 1–4 may be dominant. However, in organic semiconductors, process 5 is the most important transition process. The quantum yield of luminescence depends on the relative probability of radiative and nonradiative transitions, and the relative efficiency of producing radiative excitons. There are many possible processes for nonradiative transitions. In inorganic semiconductors, multiphonon transition and Auger recombination processes are the dominant nonradiative transition processes.⁸³ In organic semiconductors, however, again the process of exciton transitions leading to nonradiative transitions is the most important nonradiative transition process. In the following sections, we shall discuss briefly some nonradiative transition processes. The radiative transition processes will be discussed in Section 3.4, dealing with luminescence.

Multiple-Phonon Transition

The excited electron in the Franck–Condon configuration diagram, shown in Figure 3-27, can be excited from point A at the lower level at r_a to point B at a higher level at r_b by thermal excitation at a higher temperature. At such a high temperature, the electron at point B may have enough energy to cross over to the ground

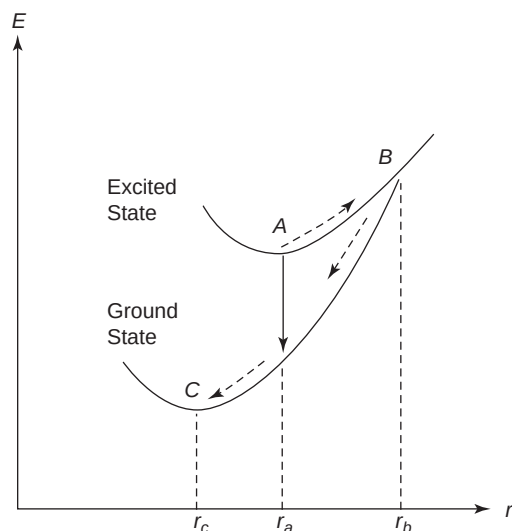


Figure 3-27 Schematic diagram showing a nonradiative multiple-phonon transition.

state, and then relax to a lower level in the ground state by the emission of phonons. This type of multiple-phonon nonradiative transition process may occur in narrow bandgap materials, such as Ge and Si.^{84,85} For more information about multiple-phonon transition, see the original work on this process.^{86,87}

Auger Recombination Processes

A number of Auger recombination processes are possible, depending on the nature of the transition and the concentration of carriers. This is a three-body collision event. Figure 3-28(a) shows the processes without involving localized states in the band gap; it is a band-to-band recombination but requires first the collision of two electrons in the conduction band or two holes in the valence band. Figure 3-28(b) shows the processes involving one level of localized states, which may be near either the conduction band or the valence band, and (c) shows the processes involving two levels of the localized states. The illustrations are self-explanatory. For example, in the case of band-to-band recombination, after electron–electron collision, one electron will recombine with a hole in the valence band, and the energy

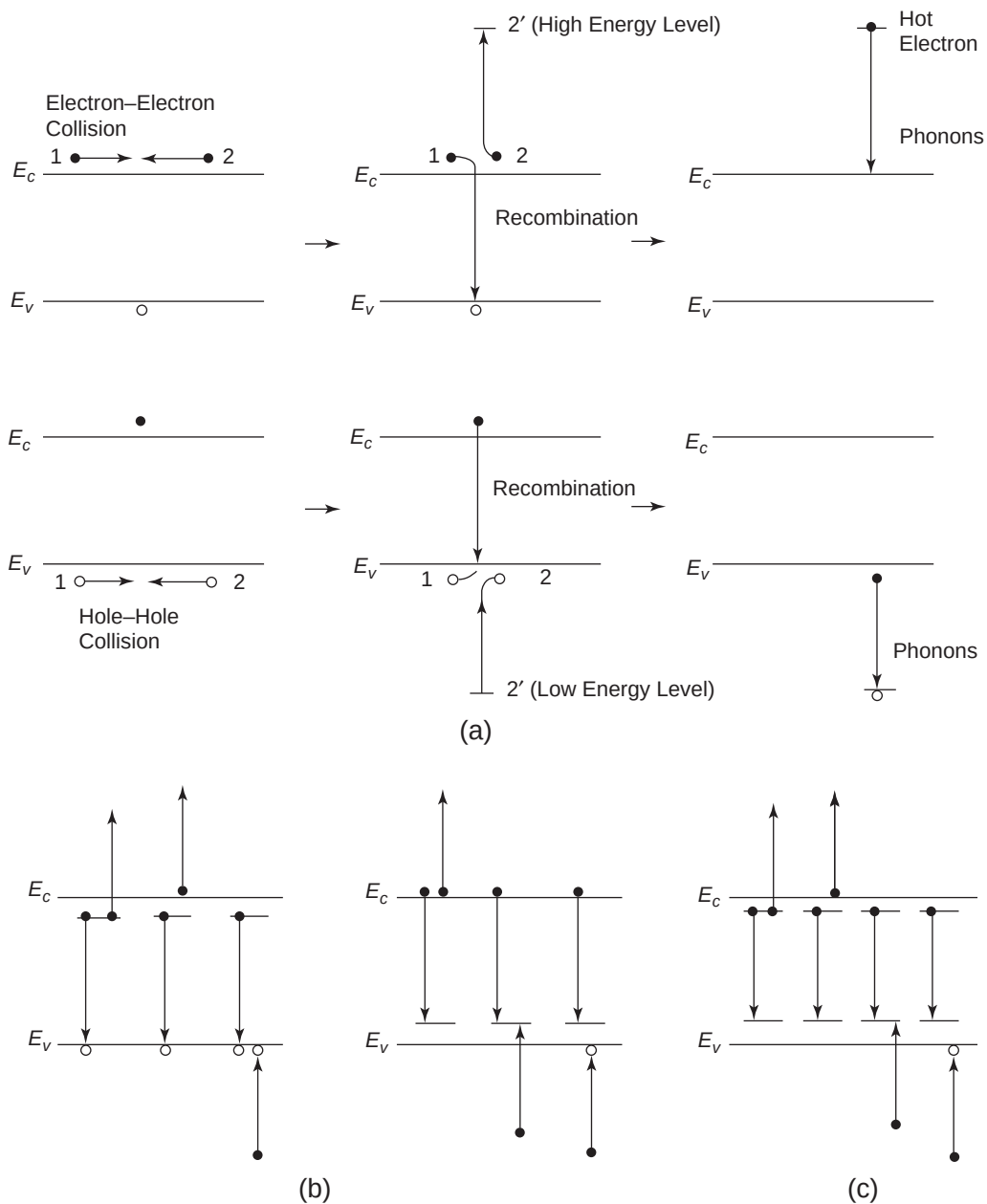


Figure 3-28 Schematic diagrams showing the Auger recombination processes: (a) band-to-band transitions without involving localized states in the bandgap, but involving either electron–electron collision in the conduction band or hole–hole collision in the valence band, (b) the transition involving one level of localized states either near the conduction band or near the valence band, and (c) the transitions involving two levels of localized states.

released due to this recombination will be transferred to another electron, making it a hot electron. This hot electron will then relax to a low energy level by emission of multiple phonons.

The Auger effect depends mainly on carrier–carrier interaction. Thus, this effect increases with increasing concentration of carriers.⁸⁸ Since the concentration of carriers increases

with increasing temperature, this effect also increases with increasing temperature. The Auger effect is an important nonradiative transition process. It plays an important role in the electrical discharge and breakdown processes in dielectric materials. It accounts for the non-radiative trapping and recombination of the carriers injected from the electrical contacts at high fields, thus causing gradual degradation of the material.⁸⁹

Nonradiative Transition due to Defects

There are always surface states at the surface of a crystal or the interface between two different materials, such as the metal–semiconductor interface, because of the presence of dangling bonds. These surface states create shallow and deep traps in the band gap. This makes band-to-band recombination unable to take place near the surface, implying that the recombination nearby must go via the surface states, thus rendering the recombination nonradiative. This effect occurs also in the internal interface between the inclusion (e.g., foreign particle) and the host material.

A crystal always consists of various defects, either structural or chemical, such as microvoids or pores, grain boundaries, dislocations, a cluster of vacancies, or precipitates of impurities. Such defects cause nonradiative transitions.

Nonradiative Transition in Indirect Bandgap Materials

Group IV elemental semiconductors and some III–V compounds belong to the category of indirect bandgap materials. In the energy band structure of direct bandgap materials, the wave vector k_c of the conduction band edge (the bottom of the conduction band) and the wave vector k_v of the valence band edge (the top of the valence band) are the same, so the law of the conservation of wave vectors (momenta) for optical transition is automatically satisfied. Since the wave vector of electrons is of the order of π/a , where a is the lattice constant, and the wave vector of photons is $2\pi/\lambda$, where λ is the wavelength of the light wave, the photon

wave vector is much smaller than the electron wave vector for visible light or even higher optical frequencies. So, an electron must make a transition between states having the same wave vector. This means that only vertical transition is allowed, so both the optical absorption and the emission are a direct transition process. However, in indirect bandgap materials, the wave vector k_c of the conduction band edge and k_v of the valence band edge are not the same, so the law of the conservation of momenta for optical transition is not satisfied. For this case, absorption or emission of phonons is required to provide momenta to satisfy the momentum conservation law. Such a transition involves either absorption or emission of phonons in an indirect transition process.

In general, an electron may be excited without change of wave vector from the top of the valence band to the conduction band by direct transition in an indirect band gap material, as shown in Figure 3-29(a). This transition creates a hole at the top of the valence band, but the electron in the conduction band is in a state with a higher energy than that at the conduction band edge (the bottom of the conduction band). This electron will quickly make a transition to the state at the bottom of the conduction band with emission of phonons.

If the photon energy $h\nu$ is less than E_g or higher than E_g , optical transition is still possible. For the case $h\nu < E_g$ the transition process is as follows:

$$E_g = h\nu + \Delta E_1$$

or

$$h\nu = E_g - \Delta E_1$$

involving absorption of a phonon.

If the photon energy is $E_g < h\nu < E_g + \Delta E_2$, then we have

$$E_g = h\nu - \Delta E_2,$$

or

$$h\nu = E_g + \Delta E_2$$

involving emission of a phonon.

These optical absorption processes are shown in Figure 3-29.

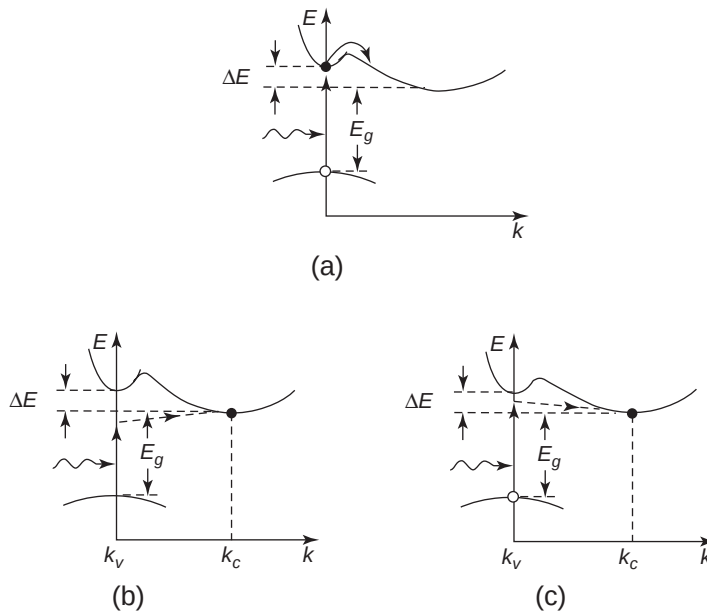


Figure 3-29 Optical absorption processes in indirect bandgap semiconductors (a) direct transition requiring a photon of energy $h\nu = E_g + \Delta E$, and the excited electron moving quickly to the bottom of the conduction band, (b) indirect transition involving absorption of a phonon of energy ΔE_1 , and (c) indirect transition involving emission of a phonon of energy ΔE_2 .

Similarly, we can easily write the processes for radiative emission. Suppose now that an electron in the bottom of the conduction band makes a transition to recombine with a hole in the valence band, but the hole is available only at the top of the valence band. In this case, absorption or emission of a phonon is again required to satisfy the momentum conservation law. If the emitted photon has energy $h\nu$ smaller than E_g , then

$$E_g = h\nu + \Delta E_1,$$

or

$$h\nu = E_g - \Delta E_1$$

involving emission of a phonon.

If the energy of the emitted photon is larger than E_g , then we have

$$E_g = h\nu - \Delta E_2$$

or

$$h\nu = E_g + \Delta E_2$$

involving absorption of a phonon.

These optical emission processes are shown in Figure 3-30.

Phonon energies are of the order of 0.01 eV, much smaller than E_g . The phonon energies are mainly from the lattice vibrations. The indirect transition involves three particles—electron, hole, and phonon—and the three particles must match each other perfectly before the transition would become radiative. Furthermore, any transition involving the absorption of more than one phonon is very improbable. Obviously, it is highly improbable that the three particles will match each other perfectly to produce radiative transition. This is why the probability of radiative band-to-band recombination p is very small for indirect bandgap materials. The probability of band-to-band radiative transition for indirect bandgap materials, such as Si and Ge, is some 10^6 times smaller than that for direct bandgap materials, such as GaAs. This implies that the band-to-band carrier recombination in indirect bandgap materials is mainly nonradiative. For more details about this subject, the reader is referred to references 90–92.

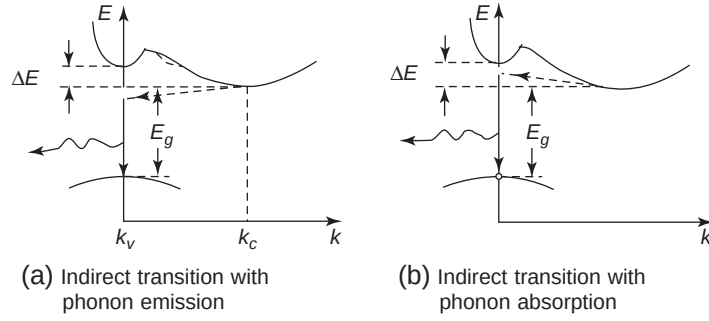


Figure 3-30 Optical emission processes in indirect bandgap semiconductors (a) involving emission of a phonon of energy ΔE_1 , and (b) involving absorption of a phonon of energy ΔE_2 .

For more information about nonradiative transition processes, see references 93 and 94.

3.3.4 Absorption and Dispersion

Suppose a monochromatic light beam of intensity I_i is incident perpendicularly on the flat surface of an absorbing material, as shown in Figure 3-31. This light beam will be partly reflected from the surface and will partly enter the material. The intensity of the light beam entering the material is thus

$$I(z) = I_i - I_r \quad (3-114)$$

where I_r is the intensity of the reflected light. Because of the absorption of the material, the intensity of the light will decrease with the distance it has traveled. The probability (or

the rate) of absorption in any element dz is constant, so the decrease of the light intensity is $-dI(z)/dz$, and it is proportional to the light intensity at z . Thus, we can write

$$-\frac{dI(z)}{dz} \propto I(z)$$

or

$$-\frac{dI(z)}{dz} = \alpha I(z) \quad (3-115)$$

where α is the absorption coefficient. Using the initial boundary condition $I(z) = I_o$ at $z = 0$, the solution of Equation 3-15 gives

$$I(z) = I_o \exp(-\alpha z) \quad (3-116)$$

The dimension of the absorption coefficient is cm^{-1} , and it depends on the wavelength of the light and the material.

The complex dielectric constant of dielectric materials is given by

$$\epsilon_r^* = \epsilon_r - j\epsilon_r' \quad (3-117)$$

where ϵ_r is the dielectric constant (or relative permittivity) and ϵ_r' is the loss factor (see The Complex Permittivity in Chapter 2). Since $u_r = 1$ for nonmagnetic dielectric materials, the complex refractive index n^* can be written as

$$n^* = n - jk = \sqrt{\epsilon_r^*} = (\epsilon_r - j\epsilon_r')^{1/2} \quad (3-118)$$

This equation leads to

$$\begin{aligned} n^2 - k^2 &= \epsilon_r \\ 2nk &= \epsilon_r' \end{aligned} \quad (3-119)$$

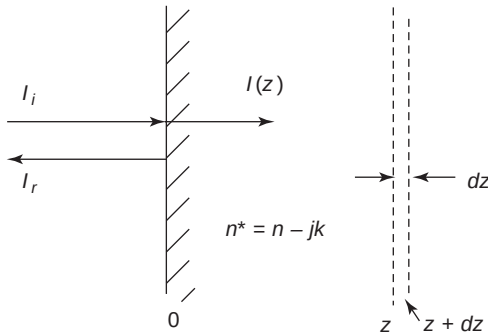


Figure 3-31 The incident I_i , the reflected I_r , and the transmitted $I(z)$ light beams from a transparent medium to an absorbing material.

From Equation 3-119 we can obtain n and k as a function of ϵ_r and ϵ'_r , and they are

$$n = \left\{ \frac{1}{2} \epsilon_r \left[1 + \left(\frac{\epsilon'_r}{\epsilon_r} \right)^2 \right]^{1/2} + \frac{\epsilon_r}{2} \right\}^{1/2} \quad (3-120)$$

$$k = \left\{ \frac{1}{2} \epsilon_r \left[1 + \left(\frac{\epsilon'_r}{\epsilon_r} \right)^2 \right]^{1/2} - \frac{\epsilon_r}{2} \right\}^{1/2} \quad (3-121)$$

For lossless materials, $\epsilon'_r = 0$ and $k = 0$, then $n = \sqrt{\epsilon_r}$.

Consider a plane-polarized light beam propagating in the z direction in the absorbing material. The velocity of the light wave inside the material is

$$v = \frac{c}{n^*}$$

or

$$\frac{1}{v} = \frac{n}{c} - j \frac{k}{c} \quad (3-122)$$

The electric field vector $F(z, t)$ of the propagating wave can be written as

$$\begin{aligned} F(z, t) &= F_o \exp[j\omega(t - z/v)] \\ &= F_o \exp[j\omega(t - nz/c)] \exp(-\omega kz/c) \end{aligned} \quad (3-123)$$

The term $\exp(-\omega kz/c)$ is a damping factor, representing the attenuation of the amplitude of $F(z, t)$ with z due to the absorption of the electromagnetic energy of the light wave by the material. Thus, the light intensity at point z can be determined by the following relation:

$$\frac{I(z)}{I_o} = \frac{F^2(z)}{F_o^2} = \exp(-2\omega kz/c) = \exp(-\alpha z) \quad (3-124)$$

Comparing Equation 3-124 to Equation 3-116, the absorption coefficient can be expressed in terms of wavelength and the absorption property of the material

$$\alpha = \frac{2\omega k}{c} = \frac{4\pi k}{\lambda} \quad (3-125)$$

In Section 3.1.2, we mentioned the dispersion phenomenon. Strictly speaking, there is no material free of dielectric loss; therefore, there

is no material free of dispersion. This means that there is no material with completely frequency-independent ϵ_r and ϵ'_r . In fact, the dispersion of ϵ_r and ϵ'_r is an intrinsic property of all materials, and all other properties must coexist with it. Also in Section 3.1.2, we mentioned that a dielectric solid can be considered an assembly of oscillators set into forced vibration by the external excitation force. The electric field vector of an electromagnetic wave of a light beam of frequency ω may cause the bound electrons or ions to be displaced from their equilibrium positions, giving rise to polarization. Equation 3-59 is the equation governing the displacement of the bound electrons or ions, and Equation 3-61 gives the complex dielectric constant resulting from this displacement by the electric field F_x .

In The Complex Permittivity in Chapter 2, we assumed that electronic and atomic polarization can immediately follow the electric field, and there is practically no time lag between P_e (or P_a) and F . This is true only at frequencies much lower than 10^{13} Hz. In the optical frequency region, the time lag becomes significant. In this case, there is a component of P_e (or P_a) with $\pi/2$ lagging from the applied electric field. This component contributes to the dielectric losses. Obviously, the damping force results in such losses due to mechanical friction. In the optical region, ϵ_r^* is still a complex quantity, which is given by Equation 3-61. From Equations 3-61 and 3-62, we obtain

$$\epsilon_r = n^2 - k^2 = 1 + \left(\frac{Nq^2}{m\epsilon_o} \right) \frac{\omega^2 - \omega_o^2}{(\omega^2 - \omega_o^2) + \omega^2 G^2} \quad (3-126)$$

$$\epsilon'_r = 2nk = \left(\frac{Nq^2}{m\epsilon_o} \right) \frac{\omega G}{(\omega^2 - \omega_o^2)^2 + \omega^2 G^2} \quad (3-127)$$

where N is the density of oscillators (atoms or molecules). These equations indicate that in the neighborhood of $\omega = \omega_o$ there is an absorption maximum and that n increases rapidly on decreasing ω through this region. With continuing decrease in ω , n will pass a maximum and then fall asymptotically as ω goes further away from ω_o . The resonance frequency ω_o is actu-

ally the frequency at which the AC conductivity $2\omega nk$ is maximum, or at which the absorption or the dielectric loss is the highest. The force constant or the damping factor G can be interpreted in this way: the AC conductivity is half the maximum value when $|\omega - \omega_0|$ is about $G/2$. Thus, G is about the bandwidth of the resonance curve, as shown in Figure 3-32.

As $\omega \rightarrow 0$, ϵ_r' and hence k approaches zero, n becomes

$$n_o = [1 + Nq^2 / m\epsilon_o\omega_o^2]^{1/2} \quad (3-128)$$

This is the undispersed refractive index, generally referred to as the refractive index of a material for $\omega \ll \omega_o$; n_o^2 is also referred to as the high-frequency dielectric constant. However, in general n_o is different from the static or low-frequency (or nonoptical) dielectric constant.

In Figure 3-33, the dielectric constant and the loss factor are plotted as functions of frequency. When the loss factor exhibits a maximum at ω_o , the material in this frequency region is opaque to the electromagnetic wave of frequency in this region. As ω rises from values lying below ω_o , the dielectric constant (or the refractive index) reaches a peak at ω_1 and then falls off rapidly to a value less than unity at ω_2 , whence it increases again with frequency ω and finally approaches unity, as shown in Figure 3-32. In the region from ω_1 to ω_2 , an increase in fre-

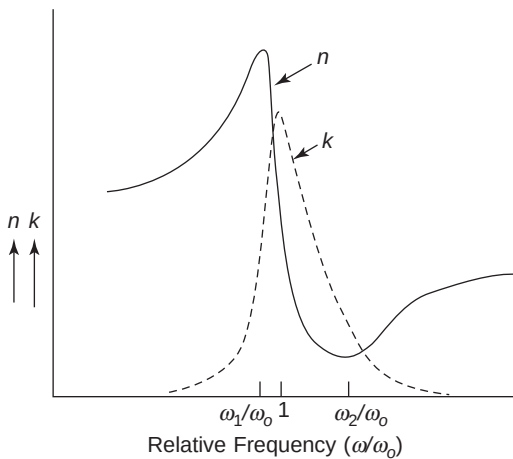


Figure 3-32 Refractive index n and extinction coefficient k as functions of frequency.

quency results in a decrease in refractive index and an increase in phase velocity. This type of dispersion is generally referred to as *anomalous dispersion*. In fact, dispersion introduced by the conductivity of the materials is anomalous. Anomalous dispersion occurs pronouncedly in gases. In liquids and solids, the general form of the dispersion curves is not greatly affected. It should be noted that Equations 3-126 and 3-127 are derived on the basis of a single oscillator, which described well the behavior of ϵ_r and ϵ_r' in the neighborhood of a single resonance frequency. Actually, a molecule is a complicated dynamic system possessing many natural frequencies, each affecting the reaction of the molecule to the incident field. Furthermore, the interaction between molecules in condensed materials also plays a role in affecting the reaction of the system to the incident field. The location of the natural frequency ω_o cannot be determined accurately by the classical approach. However, the development of the quantum mechanical theory of dispersion is based on the assumption that the interaction between the excitation field and the absorbing atoms can be represented by a set of linear oscillators, each of which has a resonance frequency corresponding to an optical absorption. Based on this theory, ϵ_r and ϵ_r' are given by

$$\epsilon_r = n^2 - k^2 = 1 + \sum_{i=1}^N \left(\frac{q^2 f_i}{m\epsilon_o} \right) \frac{(\omega^2 - \omega_i^2)}{(\omega^2 - \omega_i^2) + \omega^2 G_i^2} \quad (3-129)$$

$$\epsilon_r' = 2nk = \sum_{i=1}^N \left(\frac{q^2 f_i}{m\epsilon_o} \right) \frac{\omega G_i}{(\omega^2 - \omega_i^2) + \omega^2 G_i^2} \quad (3-130)$$

where f_i is the i th oscillator strength, measuring the magnitude of the contribution of this oscillator. For details, see references 91 and 95–98. Equations 3-129 and 3-130 are closely similar to Equations 3-126 and 3-127, indicating that the equations involving a set of linear oscillators give a more satisfactory agreement with experimental results over a wide range of frequencies.

It can be seen in Figure 3-33 that the frequencies corresponding to photon energies lower than 10^{19} Hz cannot excite the electrons

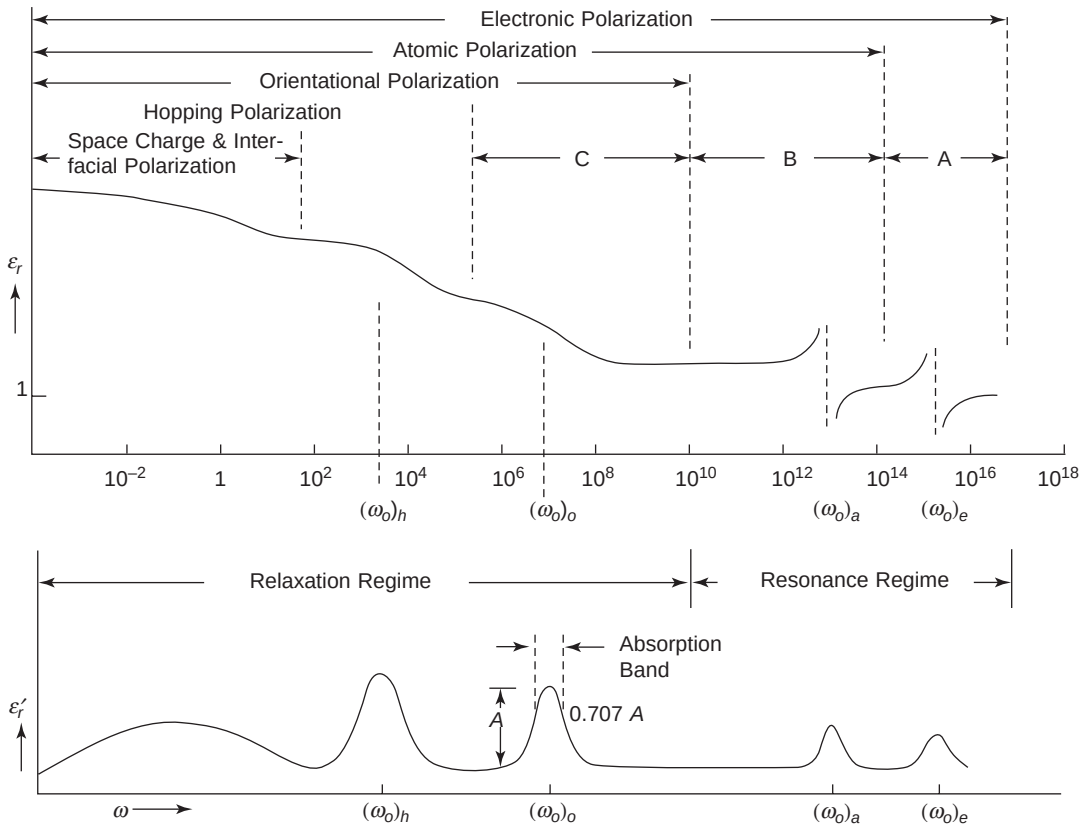


Figure 3-33 The relative permittivities (dielectric constant and loss factor) as functions of frequency.

in the inner shells; thus, these inner electrons will not contribute polarization to the material for frequencies lower than 10^{19} Hz. In region A, the polarization is due completely to electron polarization. In region B, it is due to the combination of electronic and atomic polarization, and so on. The frequency spectra of ϵ_r and ϵ'_r are self-explanatory.

3.3.5 The Franz–Keldysh Effect

For narrow bandgap materials, electrons in the valence band may be able to tunnel to the conduction band under a very high electric field. This is referred to as the Zener effect (see Chapter 8, Electrical Breakdown in Solids). The Franz–Keldysh effect is that the electron tunneling probability can be increased by absorption of photons from an illuminating

light beam.^{99–102} The wave function of a tunneling electron is

$$\psi_1 = u_1 \exp(jkx) = u_1 \exp(-|k_1|x) \quad (3-131)$$

where k_1 is an imaginary wave vector for the electron in the band gap.¹⁰³ This means that k_1 becomes a damping factor in the bandgap. If the wave function ψ_1 decays to a very low value before reaching the conduction band edge, the tunneling probability (or the electron transfer to the conduction band) is very small. This tunneling probability depends on barrier width, which is Δx , and barrier height, which is ΔE , as shown in Figure 3-34. Thus, we can write

$$qF\Delta x = \Delta E$$

or

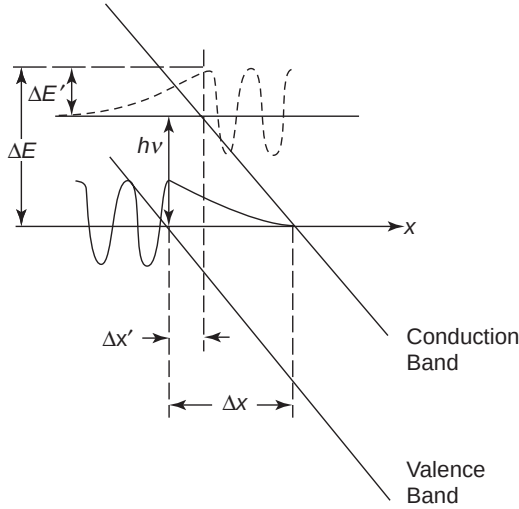


Figure 3-34 Schematic illustration showing the photon-assisted electron tunneling—The Franz-Keldysh effect.

$$\Delta x = \frac{\Delta E}{qF} \quad (3-132)$$

where F is the applied electric field. It is clear that Δx decreases with increasing applied field F .

The wave function of the tunneling electron before reaching the conduction band edge in the band gap is

$$\psi_2 = u_2 \exp[-|k_2|(\Delta x - x)] \quad (3-133)$$

The overlap of the two wave functions ψ_1 and ψ_2 determines the probability of tunneling. Without light illumination, the barrier width is Δx and the barrier height is ΔE .

However, with photons of energy, $h\nu$ from light illumination, the barrier width will change to

$$\Delta x' = \frac{\Delta E - h\nu}{qF} \quad (3-134)$$

The barrier height is also reduced by the illumination, as shown in Figure 3-34. This implies that the overlap of ψ_1 and ψ_2 is increased, thus increasing the tunneling probability. This phenomenon has been observed in Si, Ge, GaAs, and CdS.^{104–107}

Note that electron tunneling involves only the longitudinal component of the momentum. The

imaginary values of k_1 and k_2 occur only in the band gap in the longitudinal direction, that is, in the tunneling direction, which is also parallel to the applied field, and they become zero at the band edges. For optical transition, only the transverse (i.e., the vertical momentum) must be conserved. For more details about the treatment of electron tunneling involving photon absorption, see references 102, 108, and 109.

3.3.6 Formation and Behavior of Excitons

Absorption and reflectance spectra often show the structure for photon energies below the energy band gap, indicating that there are excited states of energy levels below E_c and above E_v . If the excited states are located at energy level E_{ex} , then the excited electron at this level is electrostatically bound with the hole in the valence band to form an exciton. The binding energy of the exciton is $E_c - E_{ex}$ referred to as a free electron and a free hole, and the absorption or excitation energy is $E_{ex} - E_v$, which is smaller than $E_c - E_v = E_g$. Excitons are unstable with respect to the ultimate recombination, in which the electron drops into the hole, producing either a photon or phonons.

Excitons can be formed in almost any insulating crystals, although some types of excitons are intrinsically unstable with respect to dissociation into free electrons and holes. Excitons can be generated by direct optical, indirect optical, and carrier injection (double injection) processes. In general, when the electron and hole group velocities are equal, the electron and hole may be bound by their coulombic attraction to form an exciton. Excitons can be classified into two groups, based on two different limiting approximations: One group is based on the tight-binding approximation, which was first investigated by Frenkel¹¹⁰ and Peierls,¹¹¹ and is generally referred to as Frenkel excitons; and the other group is based on the weakly binding approximation, which was first proposed by Wannier¹¹² and Mott,¹¹³ and is generally referred to as Wannier excitons. In the former, an excited state of a molecule is a state

in which an electron has been removed from the filled orbital and occupies a previously empty orbital of higher energy, leaving a hole in the original orbital (ground state); the excitation is confined within or near the molecule. In the latter, the molecules (or atoms) are packed so closely that interaction among molecules is strong, which reduces the coulombic interaction between the electron and the hole and hence increases their separation, as shown schematically in Figure 3-35.

In molecular crystals, the covalent binding of atoms within a molecule is much stronger than the van der Waals binding between molecules. This is indicated by the fact that electronic excitation lines of an isolated molecule appear also in the crystalline solid with the same molecules, with only a slight shift in frequency. At low temperatures, the excitation lines are quite

sharp, although there may be more structures to the lines in the crystal than in the isolated molecule, due to Davydov splitting.¹¹⁴ On this basis, the Frenkel model of excitons has been used extensively to explain luminescent phenomena.

However, Wannier excitons are most common in inorganic materials with covalent or ionic bonds, in which the interaction energy is large and the dielectric constant is high. In dealing with the motion of electrons in a perfect crystal, we usually ignore the mutual repulsion of the electrons. The mutual interaction of the electrons gives rise to scattering and leads to attraction between a hole and an electron, thus forming a Wannier exciton with excited states of various energy levels just below the conduction band. It corresponds, in a sense, to an excited state of an atom of the crystal being passed on to neighboring atoms by quantum mechanical resonance.¹¹²

For Frenkel excitons, the exciton radius is of the order of one lattice constant, while for Wannier excitons, the exciton radius is of the order of several lattice constants. This implies that for a tightly binding Frenkel exciton model, the atoms or molecules are well separated, such as those in molecular crystals. In this case, the electron and hole are close to one another and interact by way of coulombic and exchange energies. This means that the excited states of the Frenkel exciton can be characterized by atomic parameters. For the weakly binding Wannier exciton model, a dielectric medium is involved between the excited electron and the hole. This reduces the coulombic interaction between the electron and the hole by ϵ_r times, implying that the orbit of the excited electron encloses more atoms. This means that the excited states of the Wannier exciton must be characterized by coulombic interaction energy in a dielectric medium. Thus, the optical transition for large-radius Wannier excitons inside the solid should resemble the Rydberg transition in the hydrogen atom.

Ignoring, for simplicity, the motion of the center of mass of the electron and hole, the energy levels of the excited states of the Wannier exciton can be written as

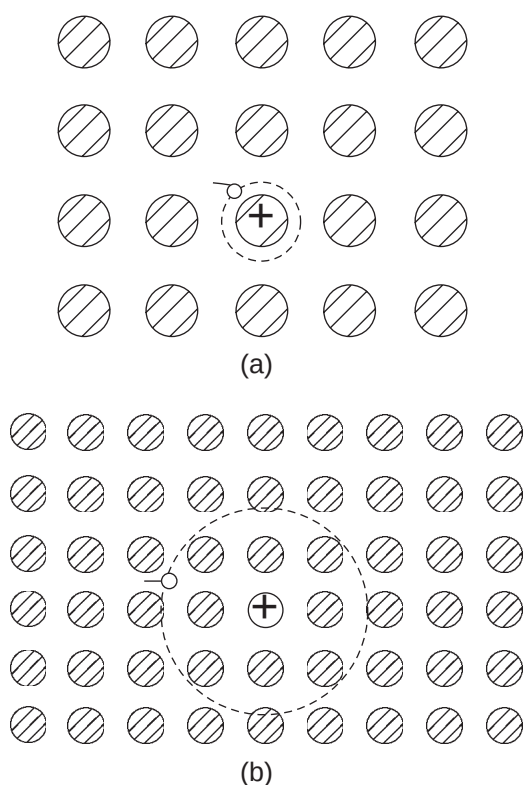


Figure 3-35 Schematic illustration of (a) the Frenkel exciton and (b) the Wannier exciton.

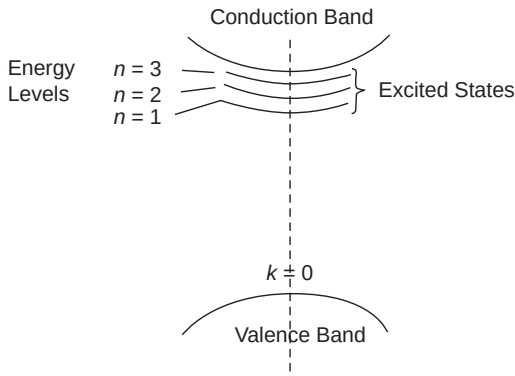


Figure 3-36 Schematic illustration of the energy levels of excited states of Wannier excitons.

$$E_{ex} = E_H \left(\frac{m_r}{m} \right) \left(\frac{\epsilon}{\epsilon_0} \right)^2 \frac{1}{n^2} \quad (3-135)$$

where m is the rest mass of the electron and m_r is the reduced mass of the electron and the hole, which is given by

$$m_r = \frac{m_e^* m_h^*}{m_e^* + m_h^*} \quad (3-136)$$

in which m_e^* and m_h^* are, respectively, the effective masses of the electron and the hole; ϵ_0 and ϵ are, respectively, the permittivities of free space and the material; $n = 1, 2, 3, \dots$ signifies the energy level; and E_H is the Rydberg energy, which is the energy of the ground state of the hydrogen atom, that is, when $n = 1$, which is given by¹¹⁵

$$E_H = \frac{mq^4}{8\epsilon_0^2 h^2} = -13.53 \text{ eV} \quad (3-137)$$

The energy levels of the excited states of the Wannier exciton are shown schematically in Figure 3-36. So, in inorganic dielectric materials or semiconductors, the excited states of the Wannier excitons appear as additional structure in the energy region below that of the fundamental absorption edge. The absorption spectrum of cuprous oxide (Cu_2O) at 77 K is a good example, illustrating the energy levels of the excited states of the Wannier exciton, as shown in Figure 3-37. The result is from Baumeister.¹¹⁶ The dielectric constant and the reduced mass of Cu_2O are 10 and $0.7m$, respectively. Similar

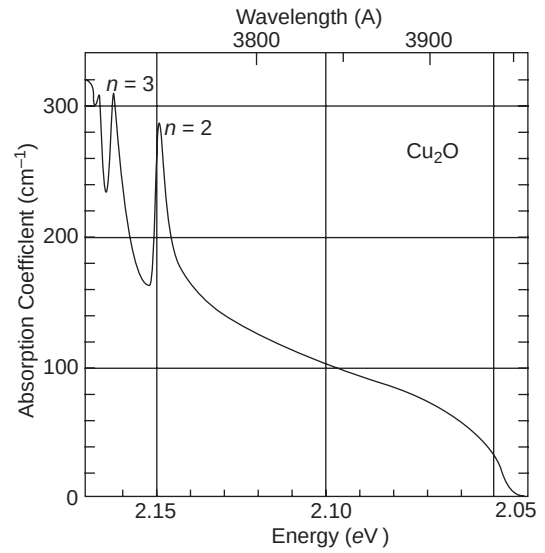
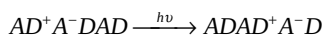


Figure 3-37 Absorption spectrum of an annealed thin Cu_2O film.

absorption spectra have also been observed in CdS , CdI_2 , PbI_2 , and AgI_2 .¹¹⁷ In general, the absorption lines created by Wannier excitons can be observed at very low temperatures, because at high temperatures the excitons may be thermally dissociated into free electrons and holes.

Between the small-radius Frenkel excitons and the large-radius Wannier excitons is an intermediate case, in which the excited electron remains correlated with the hole and both are located on the same molecule site. But the excited states occur when the excited electron is transferred to the nearest or next-nearest neighboring molecule site and still remains correlated with its parent hole. Such a correlated electron-hole pair with a spatial separation of one or two lattice constants is referred to as a charge-transfer (CT) exciton. In homomolecular crystals, consisting of only one kind of molecules, the absorption line of the CT exciton is very weak, since the energy involved in the optical transition is too close to that for Frenkel excitons. However, in heteromolecular crystals, consisting of more than one kind of molecules (such as CT complexes involving donors and acceptors), the absorption lines are much

stronger. The excited states correspond to the transfer of one electron from the donor (*D*) to the acceptor (*A*), implying that the optical transition involves the motion of an electron in the highest occupied orbital of the donor in the molecule of one kind to the lowest unoccupied orbital of the acceptor on the molecule of the other kind, such as



Polymers containing donor and acceptor groups may form donor-acceptor (DA) complexes with suitable monomers.

However, CT excitons can be considered the intermediates in carrier recombination. A CT exciton that is a nearest-neighbor pair of an electron and a hole, can move through the lattice and occasionally be dissociated by thermal excitation into a free electron and a free hole; then they recombine, producing luminescence. For more information about CT excitons, see references 118–120.

Most organic dielectric materials have weak bonding between molecules, like molecular crystals. In these materials, such as polymers and organic semiconductors, Frenkel excitons play a very important role in optical transition processes. In the following section, we shall discuss briefly the behavior of excitons with particular reference to Frenkel excitons.

Exciton Transport Processes

Excitation can transfer energy, but not charges, in wave packets of excitation. Excitons consisting of electron–hole pairs can move, implying that the energy can be transported by excitons a certain distance from where they are formed.

We mentioned earlier that when the electron and hole group velocities are equal, the coulombic attraction between the electron and hole may form an exciton. If both the electron and hole concentrations are high, the electron–electron and hole–hole coulombic repulsions are large, which tends to reduce the attractive coulombic interaction. The internal field created by the potential fluctuations of the band edges due to high concentrations of both

types of carriers tends to separate the electrons and the holes, causing the dissociation of excitons. However, if the internal field is due to a deformation potential, the direction of the force acting on the electron will be the same as that acting on the hole, thus causing the exciton to move as an entity from a large energy-gap region to a low energy-gap region.

Generally, there are three basic mechanisms of exciton energy transport:

Electromagnetic wave packet transport: The energy is transported by a polariton, which is an intimate mixture of a photon and an exciton. When a photon contributes one quantum of excitation to the electromagnetic flux, it will travel as a wave packet inside a crystal.^{120,121}

Hopping transport: If the exciton is self-trapped, it might jump to another site of the perfect lattice along a chain of molecules (e.g., anthracene molecules) until it falls into a trap (e.g., tetracene molecule), as in sensitized luminescence.^{122,123}

Long-range resonance transport: This transfer process is based on the dipole–dipole coupling mechanism, in which it is not necessary to have a chain of molecules to carry the energy. Suppose that an acceptor molecule *A* undergoes an absorption transition which coincides in energy with luminescent transition of a donor molecule *D*, then both may occur simultaneously without involving radiative emission. This implies that the excitation energy is transferred from *D* to *A* due to coupling between these two systems. In other words, the photon is absorbed by *A* before *D* has finished emitting.^{124,125}

In molecular crystals, the overlap between molecules is small and the electrons are highly localized. This implies that intermolecular interaction is small, and carriers or excitons do not move easily through the solid. However, there are two limiting cases generally used for analyzing exciton transport problems—coherent transport and incoherent transport:

Coherent Transport: In coherent transport, the time required for an exciton to transfer from

one molecule to another τ_{et} is much shorter than the displacement time required for molecules to change from the old to the new equilibrium positions τ_{md} under a change in the force of interaction between neighboring molecules upon excitation of one molecule in the solid. This means that the motion of an exciton is not accompanied by a local lattice deformation, and therefore, can be described using a band model because its mean free path is greater than the lattice spacing, and it can move coherently over several lattice spacings before being scattered. In general, coherent exciton motion may be expected to occur only in ultrapure organic crystals at low temperatures, at which both impurities and phonons are ineffective in limiting the exciton mean free path.

Incoherent Transport: The wavelength of the light that is absorbed to create excitons in a crystal is much larger than both the absorption length and the intermolecular separation in most cases. A typical example is $\lambda = 5000 \text{ \AA}$, $\alpha^{-1} = 100 \text{ \AA}$, and $a = 10 \text{ \AA}$. Therefore, it would be expected that some coherence in all excitons should exist at the time of their creation. It probably does exist for a very brief instant, but in most situations, the initial coherence associated with the photon's spatial extent is washed out by vibrational broadening.

Electron-phonon interaction is strong. For incoherent transport in the case of the slow exciton limit, the exchange energies of the exciton are small compared to phonon dispersion energies, and the transfer of the electron between adjacent molecules is rate-determining. In this slow exciton limit case, the hopping diffusion coefficient decreases with decreasing temperature. However, at low temperatures, coherent diffusion becomes important. In the case of the slow phonon limit, the electron exchange energies are large compared to the phonon dispersion energies, and the transfer of the electron between adjacent molecules is limited by the rate of the phonon transfer. In this case, the hopping diffusion coefficient increases with decreasing temperature.

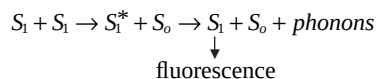
The overall diffusion coefficient is the sum of the coherent and incoherent contributions. However, it is generally agreed that coherent transport prevails at low temperatures and incoherent transport is dominant at high temperatures.

Exciton Interactions

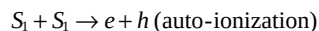
There are several possible ways in which an exciton can interact with another exciton or other particles. We shall discuss briefly various exciton interactions that are directly associated with photoluminescence and electroluminescence.

Free Exciton-Free Exciton Interactions

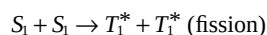
The collision of two free excitons will produce fluorescence or phosphorescence and phonons. Let us take anthracene crystals as an example. The singlet-singlet exciton interaction processes can be described by



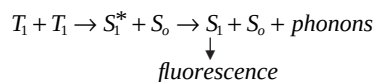
or



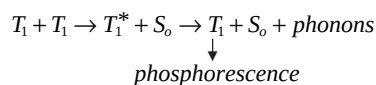
or



where S_1 and S_0 are, respectively, the excited singlet state (exciton) and the ground states; S_1^* and T_1^* are, respectively, the singlet and the triplet at high vibronic-electronic levels (hot excitons); and e and h are, respectively, the electrons and holes. The above processes are self-explanatory. Similarly, we can write the triplet-triplet interaction processes as



or



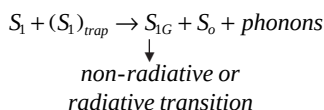
These interaction processes have been observed experimentally in anthracene and naphthalene by several investigators.^{126–129}

Free Exciton–Trapped Exciton Interactions

In real crystals, there are always traps capturing excitons. Exciton traps are sites capable of holding the energy that otherwise propagates through the lattice. These traps are generally localized and nonperiodic states in the crystals. Thus, the radiative transition rate is determined by the specific electronic structure of the trap site. The presence of traps changes the spectral energy distribution, especially the fluorescence and electroluminescence, and also changes the time dependence of exciton population and depopulation. In molecular crystals, three types of traps have been identified: guest molecules, such as tetracene doped in anthracene; defects or lattice imperfections—(structural defects); and self-trapping due to a lattice relaxation induced by excitons.^{130,131}

In general, the molecule that transfers the energy is referred to as the sensitizer, and the molecule that receives the energy is the activator. Either a host molecule or an impurity (guest) molecule can be the sensitizer. Energy transfer implies exciton migration through sensitizers, and the energy eventually given to an activator is referred to as exciton trapping.

Let us take free singlet–trapped singlet interaction as example. The interaction process can be written as



For more details about the singlet exciton energy transfer processes, the reader is referred to a comprehensive review [132].

Exciton–Charge Carrier Interactions

Again, we use anthracene as an example. The interactions of singlet excitons or triplet excitons with free electrons or holes, or with trapped electrons or trapped holes, can lead to their annihilation. When the interaction involves trapped electrons or trapped holes, it

will give rise to detrapping of trapped carriers, resulting in an increase of photoconduction. Such nonradiative destruction of excitons has been observed by several investigators.^{133,134} Obviously, exciton–charge carrier interactions lead to a decrease in both fluorescence intensity and exciton lifetime. However, such interactions provide an additional channel for nonradiative exciton decay. In other words, excitons can be quenched by charge carriers.^{133,134}

Exciton–Surface Interactions

There are two processes for quenching the mobile molecular excitons at the boundary between a molecular crystal and an electrode:

Charge transfer: An exciton can transfer an electron to an adjacent trapping center at the interface, producing a free hole in the crystal (for example, oxygen molecules adsorbed at the surface can act as electron trap centers).

Energy transfer: An exciton can transfer its energy to the acceptor molecules present at or adjacent to the surface of the molecular crystal. Usually, such an electron transfer is a rather slow process compared to the energy transfer between the crystal and an electrode.

A metal on the surface of a molecular crystal can influence the electronic states of the surface molecules in two ways: First, it can modify their energetic position; secondly, it can affect the lifetime of excited states. The first effect results from the discontinuity of the dielectric constant across the interface, which leads to a change in the molecular polarization energy. If the polarization is enhanced, surface molecules may act as traps for excitons. The second effect results from nonradiative transitions induced by metallic electrodes.

In general, the exciton quenching zone at the surface is very narrow (about 20 Å).¹³⁵ In the case of electroluminescence, the excitons are generated from electron–hole recombination, and the recombination zone is of the order of 10³ Å. Thus, in this case, exciton–surface interactions may not be so important, compared to other nonradiative transitions. However, in the

case of photoconduction, the dissociation of excitons at the boundary generates free carriers and is therefore one of the important processes for photocarrier generation. The behavior of excitons in contact with a metal surface has been studied theoretically and experimentally in some detail by several investigators.^{136,137}

Exciton–Photon Interactions

Exciton–photon interactions lead to photoionization of excitons. The photoionization of either singlet excitons or triplet excitons will produce free electrons and holes. This process of producing free electrons and holes by photoionization may compete with the exciton–exciton interactions, which may also produce free electrons and holes by autoionization mentioned earlier.

Exciton–Phonon Interactions

The effects of the interaction between Frenkel excitons and phonons arising from intramolecular vibrations can be summarized as follows¹³⁸:

- The main effect is the reduction of the bandwidth relative to the free exciton value; in general, the upper half of the band is compressed much more than the lower half.
- The effective mass of the exciton increases with increasing exciton–phonon coupling because exciton motion is retarded by nuclear displacements.
- The wider the free exciton band, the less effective is the exciton–phonon coupling.
- Exciton–phonon coupling affects the electronic transition intensity spectrum.

The intermolecular vibrations or lattice vibrations interact with phonons, causing a broadening and a shift of exciton energy levels.¹³⁸ Exciton–phonon coupling greatly affects exciton migration. In extremely pure crystals, the energy transfer rate constant increases with increasing temperature, indicating that exciton–phonon coupling is the dominant path-limiting mechanism. The nature and the effects of

exciton–phonon interactions are not yet fully understood. However, for details of the present state of knowledge in this area, see references 120, 138, 139.

3.4 Luminescence

Luminescence was known and studied long before the discovery of the quantum picture of atoms by Rutherford and Bohr. Luminescent spectra can provide information about the nature, structure, and excited states of an atom or a molecule in a solid. Luminescence and its related phenomena have been technically employed for various optical devices, including scintillation counters, electroluminescent devices, etc. The process of luminescence is the de-excitation of excited atoms or molecules (or the annihilation of excitons through recombination) by reemission of the absorbed energy as light. Whatever the form of the energy input to the luminescent materials, the final stage in the process is an electronic transition between two energy levels. There are many types of luminescence, depending mainly on the methods of excitation:

Photoluminescence: The excited atoms or molecules are produced by the absorption of photons.

Cathodoluminescence: The excited atoms or molecules are produced by the bombardment of high-energy electron beams. Many display devices are based on the cathodoluminescent process, employing a luminescent screen made of a layer of small phosphor granules adhering to a glass faceplate, as in cathode ray tubes and television screens.

Electroluminescence: Luminescence is due to the recombination of the electrons and holes injected from electrical contacts.

Triboluminescence (or sonoluminescence): Excited atoms or molecules are produced by mechanical disruption, such as friction, grinding, or sound and shock waves.

Thermoluminescence: Excited atoms or molecules are produced by heat, converting thermal energy to optical energy.

Chemiluminescence: Light emission is due to chemical reactions.

Bioluminescence: Excited molecules are produced by biochemical reactions or biological processes, including the motion of electrons, the vibrational energy of nuclei, and the rotational energy of molecules in the macromolecular system.

Ionoluminescence (or radioluminescence): Excited molecules are produced by α particles or ions.

We discussed thermoluminescence in Section 3.3.3. In this section, we shall confine ourselves to a discussion of photoluminescence and electroluminescence.

3.4.1 Photoluminescence

In general, after or during absorption of light photons, a luminescent material will be excited to excited states. The de-excitation will then produce photoluminescence. Photoluminescence is the emission of photons from electronically excited states. It is generally divided into two types—fluorescence and phosphorescence—depending on the nature of the excited states and the ground states. Suppose that a luminescent material is excited by optical excitation from the ground state to the excited state, as shown in Figure 3-38. If the return of the excited state to the ground state takes place by a spontaneous transition and emits light at a time of about 10^{-8} sec after excitation, as shown in Figure 3-38(a), this luminescent phenomenon is referred to as fluorescence. If, on the contrary, the emission of light takes place with the intervention of a metastable state, followed by the return to the excited state due to the addi-

tion of energy, as shown in Figure 3-38(b), the emission will persist for a period of time ranging from milliseconds to seconds after the excitation ends. This way of emitting light is referred to as phosphorescence. This also implies that for fluorescence, the rate of emission is 10^8 sec^{-1} and the fluorescence lifetime is 10^{-8} sec; while for phosphorescence, the rate of emission is about $1\text{--}10^3 \text{ sec}^{-1}$ and the phosphorescence lifetime is about $10^{-3}\text{--}1$ sec. The lifetime is the average period of time an excited electron remains in the excited state.

In the excited state of a singlet, the electron in the higher energy orbital has the opposite spin orientation as the electron in the lower energy orbital, that is, the two electrons are paired. In this case, the transition of the electron in the higher energy orbital to the lower one is quantum-mechanically allowed, thus resulting in the emission of fluorescence. In contrast, for a triplet state, these two electrons, one in the higher and the other in the lower energy orbitals, have the same spin orientation, that is, they are unpaired, so the return of the electron from the higher to the lower energy orbital is not allowed. This implies that a change in spin orientation is needed for the transition of an electron from the triplet excited state to the singlet ground state. Thus, depending on the de-excitation processes, the rate of light emission is small and hence the lifetime is long for phosphorescence.

Fluorescence

Singlet excitons can be generated not only directly by absorption of light but also indirectly by triplet–triplet annihilation, as discussed in the previous section. Since there is a great difference in lifetime between singlet and triplet excitons (e.g., for anthracene they are about 10^{-8} and 10^{-2} sec, respectively), fluorescence produced by singlet excitons generated directly by absorption of light or other external means is referred to as prompt fluorescence, and that produced indirectly through triplet–triplet annihilation as delayed fluorescence.

The absorption and emission of light in most organic crystals can be described by a simple

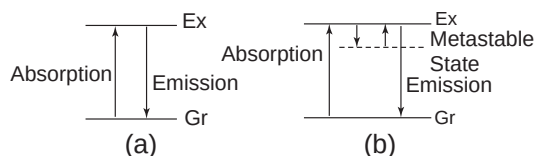


Figure 3-38 Schematic diagrams showing the absorption and emission processes of (a) fluorescence and (b) phosphorescence.

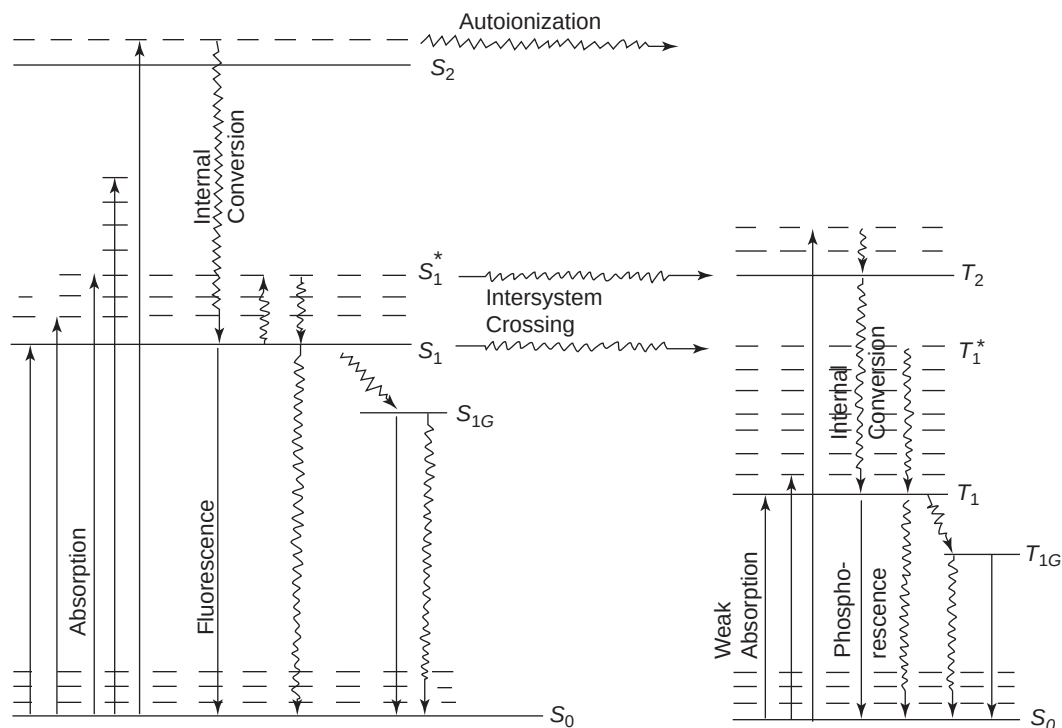


Figure 3-39 Schematic diagram showing the energy levels of various excited states. S_0 is the ground state of the singlet; S_1 and S_2 are, respectively, the first and second excited singlet states; S_1^* is the high vibronic level of S_1 ; S_{1G} is the trapped singlet state; T_1 and T_2 are, respectively, the first and second excited triplet states; T_1^* is the high vibronic level of T_1 , and T_{1G} is the trapped triplet state. $\rightarrow$ Radiative transition; $\rightsquigarrow$ Nonradiative transition; ----Vibronic levels; — Electronic levels.

energy level diagram, shown in Figure 3-39. Absorption and emission of photon energy by a bound electron occur between the allowed energy levels and must obey certain selection rules. Important first-order selection rules are summarized as follows:

The change of the orbital angular momentum quantum number ℓ must not exceed ± 1 , that is, $\Delta\ell = \pm 1$. The change of the resultant orbital angular momentum quantum number of all the electrons in an atom L must not be beyond 0, +1, that is, $\Delta L = 0, +1$ and $L = 0 \rightarrow L = 0$ is not allowed.

The change of the resultant spin angular momentum quantum number of all the electrons in an atom S must be zero, that is, $\Delta S = 0$.

The total angular momentum quantum number J , that is, the resultant of L and S so that $|L - S| \leq J \leq |L + S|$, may change by 1, 0, -1, except for $J = 0 \rightarrow J = 0$, which is forbidden.

The value of S is called the *multiplicity*. For a singlet, $S = 0$; for a doublet, $S = 1/2$; and for a triplet, $S = 1$.

All optical transitions between electronic levels are vertical, since this process occurs in about 10^{-15} sec, a time too short for significant displacement of nuclei based on the Franck-Condon principle.

In general, fluorescence is the result of a three-stage process in fluorescent materials, which are generally called the *fluorophors* or *fluors*. These three stages are

Excitation: The fluorophor must absorb a photon of energy sufficiently large to create an excited electronic singlet state.

Excited state lifetime: The excited state exists for a period of time of about 10^{-9} to 10^{-8} sec. During this short period of time, the fluorophor undergoes interactions with surrounding molecules, resulting in two possible changes: intersystem crossing to transfer some of its energy to the triplet state, or internal conversion to bring, by emission of phonons, the excited singlet state to the lowest level S_1 , which produces fluorescence.

Emission: Because part of the energy is taken away during the excitation lifetime and part is dissipated in nonradiative transition, the energy left over for fluorescence is lower; therefore, in the absorption and emission spectra, the wavelength of the emitted fluorescence is longer than that of the excitation (absorption). The difference between the excitation energy $h\nu_{EX}$ and the emission energy $h\nu_{EM}$ is referred to as the Stokes shift. Thus, the quantum yield for the fluorescence ϕ_f can be written as

$$\phi_f = \frac{\text{Number of photons emitted during fluorescence}}{\text{Number of photons absorbed during excitation}} \quad (3-138)$$

Fluorescence spectra vary widely, depending on the structure and the defects of the fluorophor and also on the solvent in which the fluorophor is dissolved. The typical fluorescence emission spectrum for perylene dissolved in benzene is shown in Figure 3-40. For perylene, the spectrum shows structure due to the individual vibronic energy levels of the excited and the ground states. The absorption spectrum appears at a lower wavelength than the emission spectrum because of the Stokes shift.¹⁴⁰

After optical excitation, the emitted fluorescence will decay with time. The decay process depends on whether it involves one body or two. The former, having the first-order kinetics, is referred to as the monomolecular or unimolecular mechanism; the latter, with the second-order kinetics, is referred to as the bimolecular mechanism. These terms are taken from chemical kinetics as synonymous for such processes. A monomolecular reaction involves only one molecule, while a bimolecular reaction

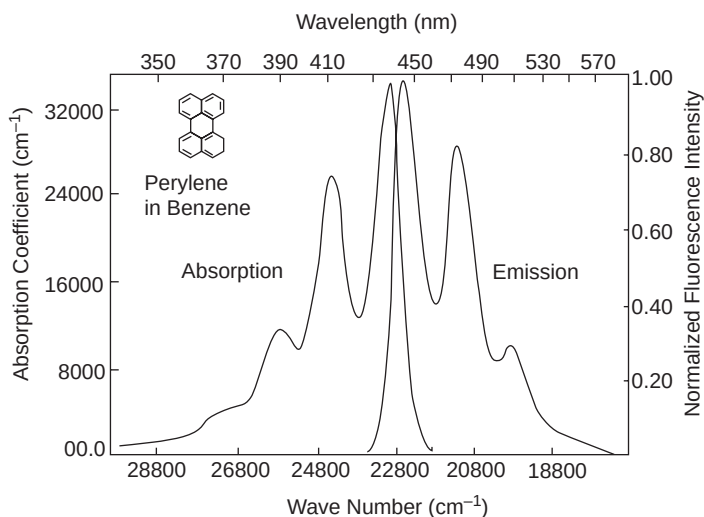


Figure 3-40 Absorption and fluorescence emission spectra of perylene in benzene.

involves the combination of two molecules. In photoconducting luminescent materials, there occur in all cases monomolecular and bimolecular recombinations. These two kinds of recombination processes are now briefly described.

Monomolecular or Unimolecular Mechanism

The rate of decrease of the number of excited electrons n is proportional to n . Thus, we can write

$$\frac{dn}{dt} = -\alpha n \quad (3-139)$$

where α is a constant that represents the probability for the annihilation of the excited electrons for luminescence. The solution of Equation 3-139 gives

$$n = n_0 \exp(-\alpha t) \quad (3-140)$$

where n_0 is the initial value of n , that is, the value just before the excitation ends. Luminescence intensity $I = -dn/dt$, so I can be expressed as

$$I = I_0 \exp(-\alpha t) = \alpha n \quad (3-141)$$

When a molecule inside a crystal is excited or ionized, the excited or emitted electron is still close to the molecule. It is possible that such an electron will undergo geminate recombination, but the yield would be very low and depends markedly on the electric field. Generally, the emitted electron tends to diffuse away from its parent molecule, as suggested by Onsager.¹⁴¹ However, if the excited electron comes from the molecules inside the crystal and returns to the ground state of the molecules, as in the return of excited singlet state S_1 to its ground state S_0 , the process is monomolecular.

Bimolecular Mechanism

The rate of decrease of the number of excited electrons n is proportional to n^2 , since the probability for the annihilation of the excited electrons for luminescence (or recombination) is also proportional to the number of available centers for the recombination. Thus, we can write

$$\frac{dn}{dt} = -\beta n^2 \quad (3-142)$$

where β is a constant related to the probability of the annihilation of excited electrons. The solution of Equation 3-142 yields

$$n = \frac{n_0}{(1 + n_0 \beta t)} \quad (3-143)$$

In this case, n decreases hyperbolically with time. The luminescence intensity is

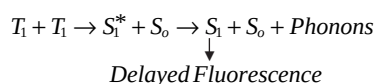
$$\begin{aligned} I &= -\frac{dn}{dt} = \beta n^2 \\ &= \frac{I_0}{(1 + \sqrt{I_0 \beta} t)^2} \end{aligned} \quad (3-144)$$

The luminescence decays more rapidly as the excitation intensity is increased.¹⁴²

Recombination processes may involve one body, two bodies, or three bodies. The so-called *bodies* may mean the *quasi-particles* that obey the Fermi–Dirac statistics, such as electrons and holes. The recombination that involves one free carrier at a time, such as indirect recombination through a recombination center (e.g., an electron captured by a recombination center and then recombined with a hole, each process involving only one carrier), is generally referred to as monomolecular recombination. The recombination that involves two free carriers simultaneously, such as direct band-to-band recombination, is generally referred to as bimolecular recombination. The recombination that involves three free carriers simultaneously, such as a three-body collision in the Auger intrinsic recombination (in which one electron in the conduction band recombines with a hole in the valence band and the energy released is taken up by a third particle-electron), is generally referred to as trimolecular recombination or three-body recombination (or simply as Auger or impact recombination).

So far, we have discussed prompt fluorescence. As shown in Figure 3-39, some light may be absorbed in the excitation of triplets to their excited states, particularly for the light photons of lower energy—lower than the lowest energy required for creating excited singlets S_1 but higher than that for creating excited triplets

T_1 , although the absorption is very low due to the fact that the optical transition (absorption and emission) is generally forbidden by the first-order selection rules. However, there is always some way to circumvent the selection rules, so optical transition in triplets is never zero.¹ It is possible that the interaction of two triplet excitons leads to indirect creation of an excited singlet by triplet-triplet annihilation, resulting in delayed fluorescence emission.^{143,144} The interaction process can be simply described as follows:



The delay time is approximately of the order of the lifetime of the excited triplet state. The delayed fluorescence intensity is proportional to I_{ex}^2 , where I_{ex} is the intensity of the excitation light and depends on the traps in the crystal.¹¹⁸ Furthermore, the interaction between excitons and charge carriers always results in quenching of delayed fluorescence.^{133,145}

Fluorescence spectra under various temperatures and applied electric fields provide a great deal of information about the structure of the material, including structural and chemical defects, as well as charge carriers. Therefore, fluorescence spectroscopy has been widely used for investigating the properties of biological macromolecules.^{140,146,147}

Phosphorescence and Phosphors

Short-persistent fluorescence emission involves transition between states with the same multiplicity, whereas long-persistent phosphorescence involves delays in the quasi-metastable states having different multiplicities from the ground state. In this case, there is still a low but appreciable probability of a spontaneous transition to the ground state, producing the so-called *beta* phosphorescence.¹ Metastable states are presumed to be the lowest triplet state in the molecule in which it is possible for an electron to absorb thermal energy to raise the system to a nonmetastable singlet state, whence it can make a delayed fluorescence transition to

the singlet ground state, producing the so-called *alpha* phosphorescence.

Phosphorescence may be considered a temporary storage of potential luminescence energy in the form of trapped excited electrons or electrons in metastable states. Electrons in metastable states are not allowed to make radiative transitions to lower, unoccupied states by the first-order selection rules. However, electrons in metastable states may be raised to higher levels, where radiative transitions to lower levels are allowed. In this respect, metastable states are similar to traps. An electron in a particular trap may have to be thermally excited to another, higher level, where it may be able to travel to meet a hole appearing nearby to make a radiative transition.

Metastable states acting as traps are responsible for the persistence of phosphorescence. A trapped electron may remain in a trap for some time. It may be released from the trap for a while and then be retrapped in another trap. In this case, the ultimate phosphorescence will be further delayed. In general, thermal excitation is required to release a trapped electron from the trap, as shown in Figure 3-38(b). The thermal energy comes from phonon interaction with the surroundings. The number of escapes per second for the thermal release of a trapped electron depends on the energy level of the trap and temperature, which, according to the general thermal activation law, can be expressed as

$$p = s \exp(-E_t/kT) \quad (3-145)$$

where E_t is the energy level of the trap below the conduction band (or the excited state), and s is a constant approximately equal to the rate of natural fluorescence emission, which is of the order of 10^8 sec^{-1} . It can be seen that p decreases with increasing depth of the traps and with decreasing temperature. As the delay time for phosphorescence emission after the original excitation τ is equal to $1/p$, so the delay time τ increases with increasing depth of the traps and with decreasing temperature. For example, for Cu-doped ZnS crystals, $E_t = 0.65 \text{ eV}$ below the conduction band edge, the delay time is about three minutes at 18°C and becomes about one day at -50°C .¹⁴²

Materials exhibiting phosphorescence are referred to as phosphors. Since most organic luminescent materials have instability problems at elevated temperatures, low absorption power, and short afterglow, most phosphors are inorganic materials. Artificial inorganic phosphors are not only much more efficient, versatile, and stable than naturally occurring minerals, they are also better defined and hence more suitable for practical applications. Generally, artificial phosphors depend on the presence of some impurity ions, called *activators*, inside the crystal for persistent luminescence. Such impurity ions usually substitute some of the host ions in the crystal lattice. If the charge of the doped activator ions is not identical to that of the host ions, then it is necessary to introduce other impurity ions, called *coactivators*, into the crystal to balance the charges. The persistence of the luminescence is directly related to the lifetime of the excited state of the activators. For atomic electric dipole transition, the persistence time is of the order of 10^{-8} sec. But for phosphorescence transition, the persistence time can be much longer, as mentioned earlier. For phosphors, the activators and the coactivators are usually considered acceptorlike electron traps and donorlike traps, respectively.

There are two types of luminescence:

Characteristic luminescence: Excitation, emission, and persistence depend on the activators.

Noncharacteristic luminescence: Excitation, emission, and persistence depend on the action of both the host atoms in the crystal

lattice and the activators, as well as the coactivators.

For characteristic luminescence, an electron in the ground state can be excited directly to the activator, and the electron in the activator can make a direct transition to the ground state, producing luminescence. For noncharacteristic luminescence, activators are generally chosen with multiplicity different from that of the ground state, so the direct transition from the activator to the ground state is not allowed. In this case, optical absorption first creates electron-hole pairs in the host lattice. The excited electrons will quickly be trapped in the activators and the holes in the coactivators. The trapped electrons may remain in the traps for some time before being released thermally. The time that an electron spends in the trap depends on the depth of the trap energy level below the conduction band and temperature, governed by Equation 3-145, as mentioned earlier.

For phosphors involving activators only (i.e., when the charge of the activator is identical to that of the host atom it replaces), some of the thermally released electrons may make a transition to the valence band, producing luminescence, and some may go on to be retrapped before being detrapped again to make a luminescent transition, as shown in Figure 3-41(a).

For phosphors involving both activators and coactivators, the thermally released trapped electrons may have a better chance to recombine with holes trapped in the coactivators, as shown in Figure 3-41(b). It should be noted that transition between donors and acceptors can be very efficient. Many phosphors have activators

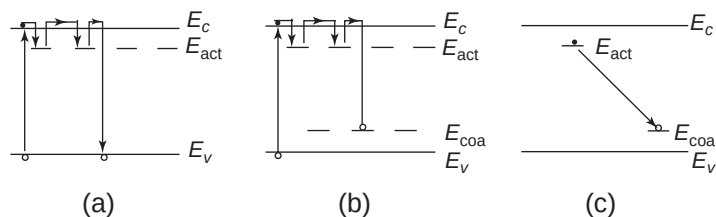


Figure 3-41 Schematic diagrams showing the optical absorption and emission process in (a) phosphors involving activators only, (b) phosphors involving activators and coactivators, and (c) phosphors involving donor-accepted recombination.

as acceptors and coactivators as donors. These two impurity atoms can be thought of as an impurity molecule formed by their coulombic interaction, as shown in Figure 3-41(c). The electron in the activator can recombine with the hole in the coactivator, producing luminescence

$$h\nu = E_g - (E_{\text{act}} + E_{\text{coa}}) + \frac{q^2}{4\pi\epsilon r} \quad (3-146)$$

where r is the separation between the donor and the acceptor, and ϵ is the permittivity of the phosphor. Obviously, the emission spectrum would be broad, due to the large range of the possible impurity atom separations. In fact, the emission of red light from GaP is due to Zn substituting for Ga and O substituting for P in the GaP-doped crystal, with Zn and O forming a donor-acceptor complex.¹⁴⁸

Group II–VI compounds are an important class of luminescent materials. ZnS is one of these materials that has been extensively studied. In general, the electronic properties of group II–VI compounds are between those of group I–VII and group III–V compounds. For example, the energy band gap of ZnS is about 3.7 eV, as compared to about 9.4 eV for KCl and 1.4 eV for GaAs. Regarding the bonding of atoms, the ZnS lattice is about 75% ionic, compared to predominantly ionic KCl and predominantly covalent GaAs.

Most widely used phosphors are made of group II–VI compounds activated by suitable impurities as activators and coactivators. As mentioned earlier, the charge balance is important when choosing impurities to replace the host atoms. For example, ZnS:Cu means ZnS activated by Cu activators, Cu⁺ to replace Zn²⁺ creates a charge imbalance problem. So, it is necessary to introduce another impurity ion, such as Ga³⁺, into the crystal as coactivator to compensate for the charge difference. In this case, one Cu⁺ and one Ga³⁺ replace two Zn²⁺. Alternatively, a halide ion, such as Cl[−], can be used to replace S^{2−}. In this case, the Cu⁺ and Cl[−] pairs do not produce any charge imbalance problem.

On the basis of the basic principles for phosphors given here, the properties of phosphors, including persistence and the luminescence

spectrum, can be tailored to suit any particular applications by selecting suitable impurities for activators and coactivators. For example, the phosphors used for the color displays involve the production of three primary colors: blue, green, and red. The phosphors commonly used for this purpose are ZnS:Ag for blue, Zn_xCd_{1−x}S:Cu for green, and Y₂O₂S:Eu,Tb for red.⁵ For more details about phosphors and luminescence in general, see references 103, 142, and 148–150.

3.4.2 Electroluminescence

Electroluminescence in a broad sense is the phenomenon of converting electrical energy into light energy. This implies that luminescence can be produced by electrical activation. There are two major types of electroluminescence. One type does not involve carrier injection from electrical contacts; this type is generally referred to as classical electroluminescence. The other type involves carrier injection and is generally referred to as injection electroluminescence.

Classical Electroluminescence

Without involving charge carriers injected externally through electrical contacts, this type of electroluminescence is sometimes referred to as intrinsic electroluminescence. In this case, the excited electrons must be produced by electrical activation. The idea of intrinsic electroluminescence was first proposed and developed in 1936 by Destriau, who used a thin layer of phosphor powder suspended in an insulating medium.¹⁵¹ The early structure of this type of electroluminescence (EL) device is very simple. The phosphor powder (usually ZnS:Cu phosphor particles) is suspended in a transparent insulating medium and sandwiched between two electrodes, one of which is transparent. Later, instead of using the phosphor-powder dispersion method (which has many disadvantages, such as short life, low stability and brightness), the techniques of producing thin phosphor films by vacuum deposition or

other methods were developed.¹⁵²⁻¹⁵⁴ We shall discuss briefly the performance of such thin phosphor film EL devices, but before doing so, we would like to mention two possible mechanisms that may be responsible for intrinsic electroluminescence: electron tunneling and impact ionization.¹⁵⁵

Electron Tunneling Mechanism

If the phosphor film is thin and the applied electric field is sufficiently high, electrons at the acceptor level may be able to tunnel to the conduction band and subsequently recombine with the hole (empty acceptor) nearby, producing luminescence, as shown in Figure 3-42(A).

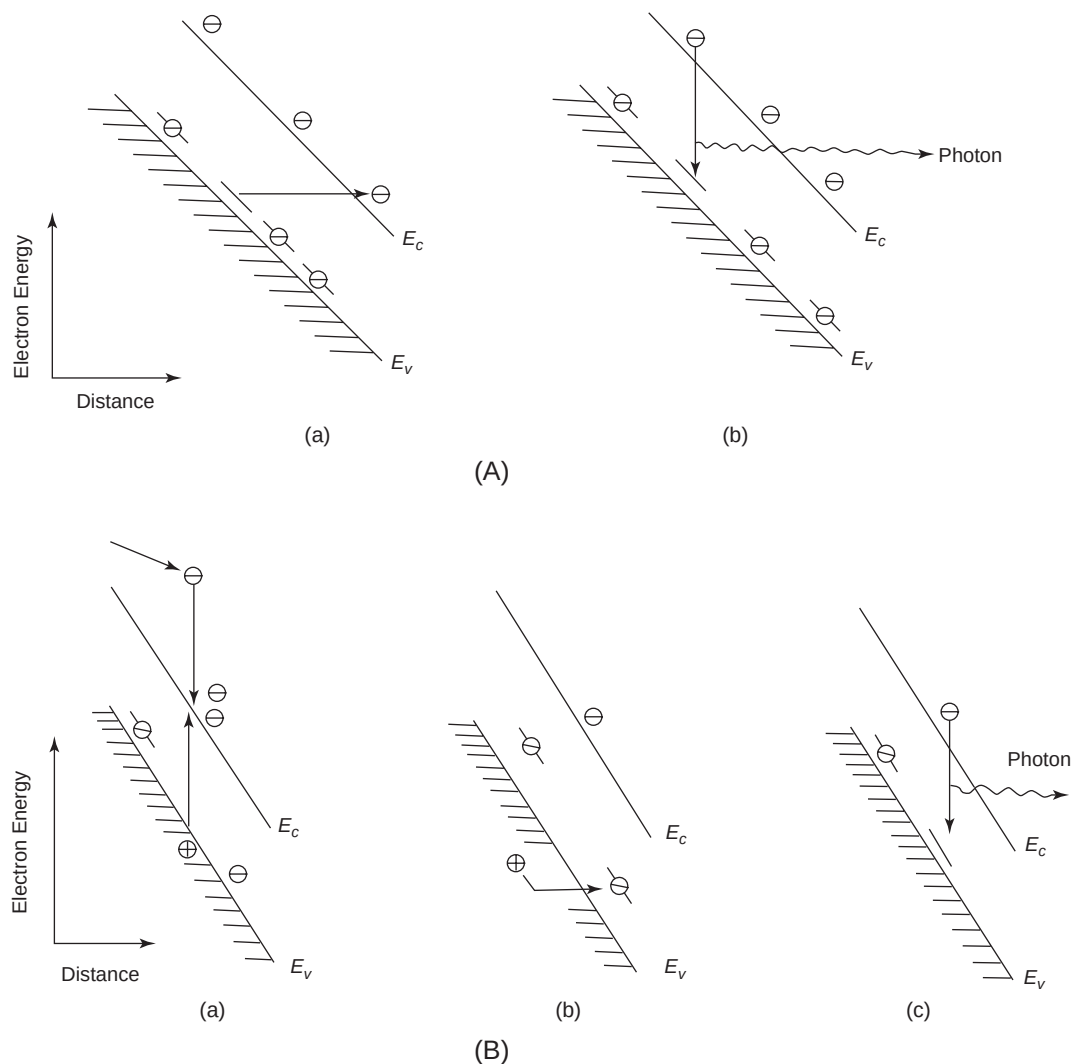


Figure 3-42 Schematic diagrams showing possible mechanisms for producing electroluminescence, (A) involving tunneling: (a) an electron in the acceptor state may tunnel to a state in the conduction band of the same energy, and (b) this electron may then fall into an empty acceptor state, resulting in radiative emission; (B) involving impact ionization: (a) an electron moving in the high electric fields may acquire sufficient energy to produce an electron-hole pair by impact ionization, (b) the hole may recombine with an electron in the acceptor state, and (c) the electron may then fall into the empty acceptor state just created by (b), resulting in radiative emission.

Impact Ionization Mechanism

Electrons in the conduction band that are accelerated under a sufficiently high field may acquire enough energy to cause impact ionization in the material, producing electron-hole pairs. The holes will quickly be trapped at the acceptor sites nearby, while the electrons in the conduction band will move until they find suitable empty acceptor sites and then recombine with holes there, producing luminescence, as shown in Figure 3-42(B). This phenomenon has been observed in the ZnS:Mn phosphor film EL devices.

The typical thin-film EL device is the LUMOCEN (LUMinescence from MOLECULAR CENters). This EL device consists of phosphor with special luminous centers, made of rare earth-halide molecules.¹⁵³ Figure 3-43 illustrates schematically a representative structure of a LUMOCEN device. Normally, ZnS phosphor is used as the host phosphor material and TbF₃ as the luminous centers. The SnO₂ film serves as the transparent electrode. However, this structure has not been developed further for practical applications because of many inherent shortcomings. Several better thin-film EL devices have been developed, including ZnS:Mn thin-film EL devices, radio-frequency sputtered ZnS:Mn, Cu thin-film EL devices, ZnS:Mn thin-film EL devices with double insulating layer structure, etc.^{154,156}

We use the ZnS:Mn thin-film EL device with a double insulating layer structure as an example to show the typical performance of classical electroluminescence devices. Figure 3-44(a) illustrates schematically the fundamental structure of the EL active film sandwiched between two insulating layers. These two insu-

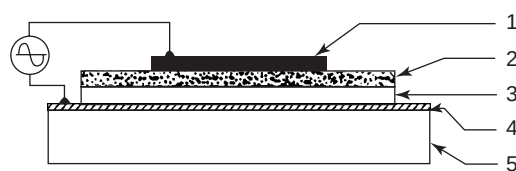


Figure 3-43 Schematic diagram showing the basic structure of a LUMOCEN device. 1: metal electrode (Al); 2: ZnS:TbF₃ film (1500 Å); 3: HO₂ film (3000 Å); 4: SnO₂ film (2000 Å); 5: glass substrate.

lating layers prevent leakage current from flowing through the EL film and keep the applied electric field for producing luminescence across the active element of ZnS:Mn phosphor. The materials for the insulating layers must have a high breakdown strength and a high dielectric constant, so that they do not take away a large portion of the applied field, and they must be transparent. Usually, Y₂O₃, Si₃N₄, or Al₂O₃ is used for insulating layers.

The general performance of this device is good and has a long life of operation. This device emits bright, yellowish-orange light under an AC field. The luminescence spectrum is not affected by the amplitude of the driving AC field and frequency. The typical brightness-voltage characteristics of this device are shown in Figure 3-44(b). The brightness depends strongly on the applied voltage in the lower voltage region and tends to saturate in the higher voltage region. The brightness-voltage characteristics shift to higher voltages as time elapses, but the saturation value does not change. This shift is not due to the degradation of the device but to a kind of stabilization process similar to the thermal annealing process. After the completion of such a stabilization process, the device becomes extremely stable for three years, without any change in brightness. The stabilization process can be accelerated by elevating the temperature and applying a suitable voltage on the device at the same time. For example, if a voltage in the saturation region is applied across the device at 200°C, the time required to complete the stabilization process is within one hour.¹⁵⁶

Injection Electroluminescence

Injection electroluminescence is sometimes referred to as extrinsic electroluminescence, because the electrons and holes undergoing radiative recombination are supplied externally by carrier injection through electrical contacts. Light emission from electrical contacts is, of course, very similar to light emission from p-n junctions, on which most light-emitting diodes (LED) and semiconductor lasers are based. However, for many luminescent materials, such

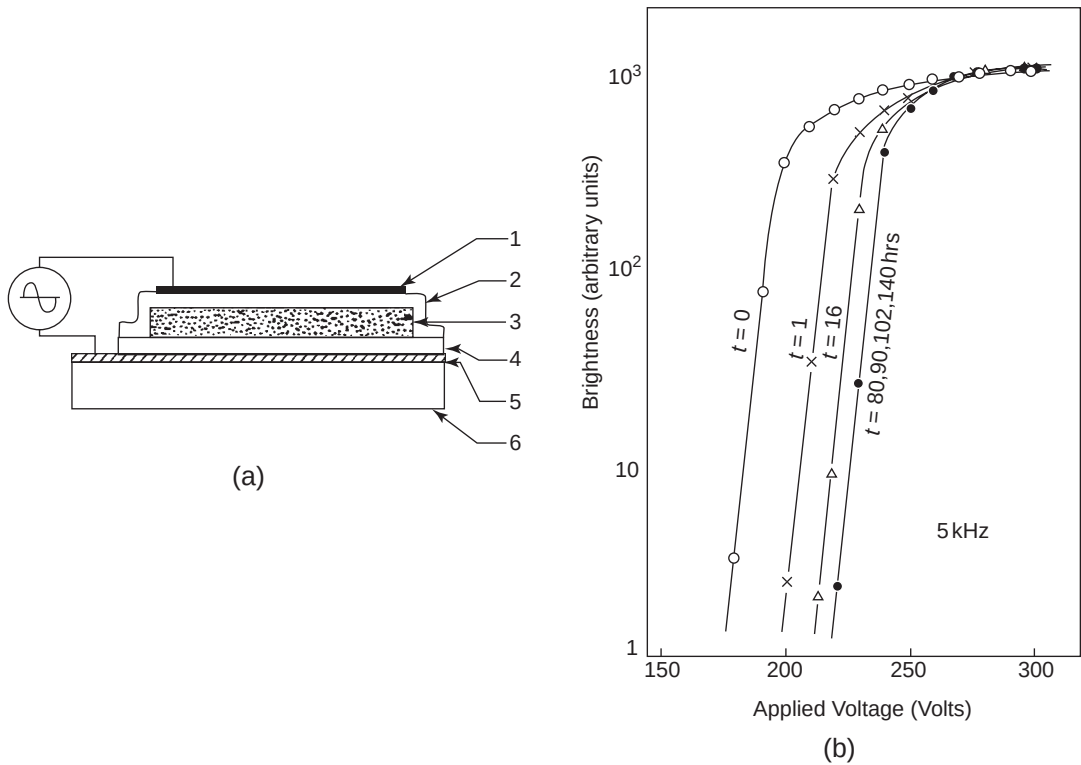


Figure 3-44 (a) Schematic diagram showing the ZnS:Mn thin-film electroluminescence device with a double-insulating layer structure, 1: metal electrode (Al); 2: insulating layer (2000 Å); 3: active layer ZnS:Mn film (5000 Å); 4: insulating layer (2000 Å); 5: SnO₂ transparent electrode; 6: glass substrate, and (b) the brightness-applied voltage characteristics of the ZnS:Mn thin-film device.

as molecular crystals and II–VI compounds, p–n junctions are not feasible. In such cases, other methods of carrier injection must be used. In the following sections, we shall discuss briefly some commonly used methods of carrier injection.

Carrier Injection from Electrical Contacts

Filamentary Charge-Carrier Injection in Solids in Chapter 7 discusses the theory of filamentary double injection in solids. It is likely that multiple current filaments may simultaneously exist between two parallel plane electrodes. In such a case, we can always consider that within a domain of radius r_d is enclosed only one current filament and that the total current between the plane electrodes may be expressed as

$$I_T = I_{\text{domain1}} + I_{\text{domain2}} + \dots = \sum_n I_n \approx HI \quad (3-147)$$

This means that the total current can be represented by the current in one domain I multiplied by a constant H . We can also assume that the current is uniformly distributed within the area of πr_d^2 , provided that r_d is small enough to satisfy this condition.

In molecular crystals—for example, in undoped and doped anthracene—both the electron and hole mobilities are generally small and the recombination rate constant is large, resulting in a small space charge overlap. Thus, the simultaneous injection of electrons and holes from the contacting electrodes will produce two-carrier space charge limited currents within the filament, and lead to electroluminescence when two types of carriers meet and recombine radiatively.

After the onset of electroluminescence in a molecular crystal with a fixed emission spectrum, the electroluminescent brightness is gov-

erned by the external quantum efficiency,¹⁴⁸ η_q , given by

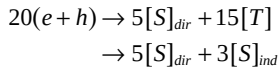
$$\eta_q = \eta_i \eta_g \eta_e = \eta_{int} \eta_e \quad (3-148)$$

where η_i is the carrier injection efficiency, which is the ratio of the current due to minority carriers to the total current. If it is assumed that J_n is the current due to minority carriers, then

$$\eta_i = J_n / (J_n + J_p) \quad (3-149)$$

η_g is the light generation efficiency; $\eta_{int} = \eta_i \eta_g$ is the internal quantum efficiency, which is a function of the total current density and temperature of the electroluminescence specimen; and η_e is the light extraction efficiency, which is defined as the ratio of power loss due to the light transmission within the electroluminescence specimen to the total power losses, which consist of both the losses in the bulk and on the surface. This can be considered fixed for a given specimen.

For double injection, the recombination of the injected electrons with the injected holes in the molecular crystal will yield singlet and triplet excitons. It is generally accepted that the singlet excitons producing fluorescence are partly generated directly by electron-hole recombination and partly generated indirectly by triplet-triplet recombination in pairs, according to the following relation^{157,158}:



where e and h represent, respectively, the electron and the hole; $[S]_{dir}$ and $[S]_{ind}$ represent, respectively, the singlet excitons produced by the direct and the indirect processes; and $[T]$ represents the triplet excitons. It is the efficiency of generating $[S]_{dir}$ and $[S]_{ind}$ and their subsequent population in the crystal that governs the electroluminescent intensity, but the threshold voltage is mainly governed by local field effects on the electrode surfaces. Since there is a great difference in lifetime between the singlet and the triplet excitons in molecular crystals (for example, in anthracene they are 10^{-8} and 10^{-2} sec, respectively), the total electroluminescence consists of prompt

electroluminescence due to $[S]_{dir}$ and delayed electroluminescence due to $[S]_{ind}$ and exhibits time constants corresponding to both of these decays. In the steady state, however, electroluminescence is the combination of these two.

Through a semitransparent planar electrode, Hwang and Kao have observed that electroluminescence in anthracene occurs within a single filament, and its brightness decreases with increasing distance from the center of the filament in a manner similar to the variation of the current density with radial distance described in Filamentary Charge-Carrier Injection in Solids in Chapter 7.¹⁵⁹ As the applied voltage is increased, multiple filaments are observed and the overall electroluminescent brightness increases. Since the normal parallel-plane electrodes have sharp edges, the filaments are generally formed near the edges, because the field is higher there.¹⁵⁹

Electroluminescent intensity (or brightness) is dependent on the injection current density. At a high injection level, bimolecular decay is dominant, so the electroluminescent brightness is proportional to the injection current. At a low injection level, monomolecular decay becomes dominant, so the electroluminescent brightness is proportional to the square of the injection current.^{131,159,160} Some typical results for the total brightness B_T as a function of the injection current I for anthracene are shown in Figure 3-45. The experimental results are from Williams and Schadt¹⁶¹; Schwob and Zschokke-Granacher¹⁶³; Kawabe, Masuda, and Namba¹⁶⁴; and Hwang and Kao.¹⁵⁹

If the current-voltage (I - V) characteristics are known, the relationship between electroluminescent brightness and applied voltage can easily be deduced. It is most likely that prior to the onset of observable electroluminescence, the current is dominated by one type of charge carrier from one electrode, the other being a neutral or blocking contact. But when the applied voltage $V > V_{th}$, the space charge built up near the blocking contact becomes sufficient to turn on the carrier injection; therefore, the current increases sharply to another regime. In this regime, both types of carriers, holes and electrons, play almost equally important roles in the conduction current, and both electrodes

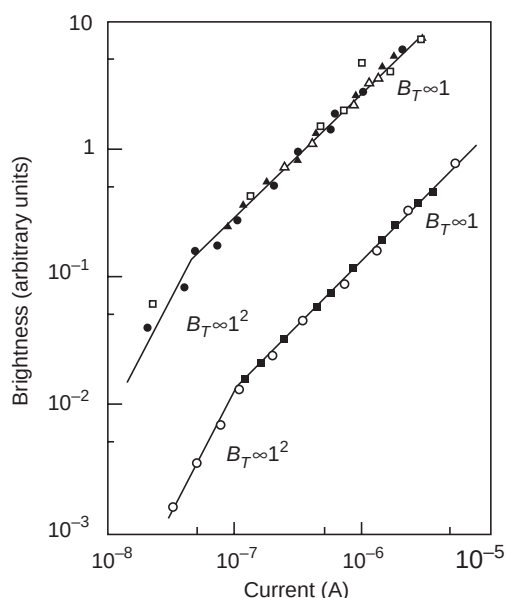


Figure 3-45 Electroluminescent brightness as a function of current for (a) undoped anthracene crystals and (b) anthracene crystals doped with tetracene. Solid curves are based on the theory and experimental results are from Williams and Schadt,¹⁶¹ ▲, Mehl and Funk, □,¹⁶² Schwob and Zschokke-Gränacher, ○,¹⁶³ Kawabe, Masuda, and Namba, △,¹⁶⁴ and Hwang and Kao, ■.¹⁵⁹

under such a condition can be assumed to be ohmic contacts, but the effects of field-enhanced detrapping and reabsorption of electroluminescence are excluded. If these effects are taken into account, it would be expected that I and hence B_T would increase more rapidly with V . This trend of B_T - V relationship has been observed in anthracene.¹⁵⁹

Electroluminescent intensity is also dependent on temperature. The temperature-dependent phenomenon may be explained in terms of three processes:

- Exciton-trapped exciton interactions
- Exciton-carrier interactions
- Exciton-surface state interactions, which control the electroluminescent intensity and are temperature-dependent

The surface states at the interface between the contacting electrode and the anthracene crystal

quenches singlet excitons, and the quenching rate decreases with increasing temperature. However, the effect of surface states may be very small compared to those of processes 1 and 2,¹⁴⁵ and for most cases, process 3 may be ignored.

If the temperature for peak electroluminescent brightness is defined as the brightness characteristic temperature T_b , then it is possible that for temperatures lower than T_b , process 1 is dominant, and for temperatures higher than T_b , process 2 becomes important. T_b can be thought of as the characteristic temperature of these processes, at which the singlet exciton-attempt-escape frequency is equal to the carrier-singlet exciton reaction rate.¹⁵⁹

For temperatures lower than T_b , brightness increases with increasing current. Since B_T is proportional to I in the high-injection case, the temperature dependence of B_T can be explained in terms of the temperature dependence of I .

For temperatures higher than T_b , brightness decreases although the current still increases with increasing temperature, and electroluminescence disappears at a certain temperature, depending on the applied voltage. It has been experimentally observed that the interaction of singlet excitons¹⁶⁵ or of triplet excitons¹³³ with charge carriers quenches fluorescence. The change of temperature may not greatly affect the carrier injection from the electrodes, but it would greatly affect the ratio of the free carrier density to the total carrier density, which includes both free and trapped carriers.

Typical temperature-dependent electroluminescence results are shown in Figure 3-46. The experimental data are from Hwang and Kao.¹⁵⁹ It should be noted that the nature of exciton-trapped exciton interactions and carrier-exciton interactions is still not fully understood. However, the argument here serves, at least, to explain qualitatively this temperature-dependent phenomenon.

Electroluminescence consists of two components: the fast component (generally referred to as prompt electroluminescence) and the slow component (generally referred to as delayed electroluminescence). The "fast" light transient

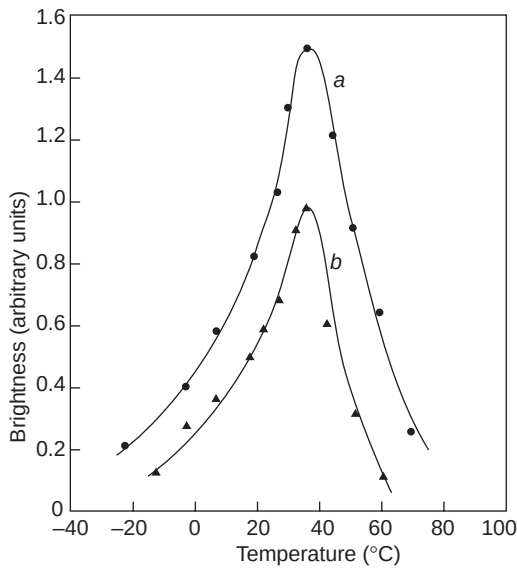


Figure 3-46 Electroluminescent brightness as a function of temperature for undoped anthracene crystals of about 1 mm in thickness for (a) 1.2 kV applied voltage and (b) 1.0 kV applied voltage.

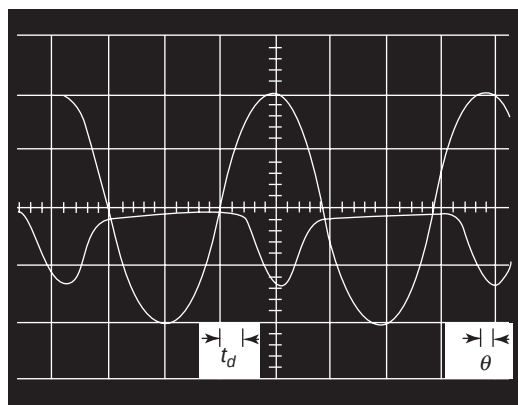
marks the time when the two leading carrier fronts meet in the specimen, while the “slow” light transient involves triplet-triplet annihilation and electron and hole detrapping processes. The time dependence of the “fast” current transient enables the determination of the carrier recombination rate constant, and the “slow” current transient can be used to monitor any change of exciton generation rate that may arise from a decrease of mobile carriers due to trapping.

The steady state (or DC) electroluminescence spectrum is independent of the electrode material but depends on temperature and crystal preparation. Under pulse voltage conditions, however, a carrier-injection mechanism of a normally blocking contact becomes apparent. Electroluminescence first appears at the time, after the application of a voltage pulse, corresponding to the transit time of $d^2/(\mu_n + \mu_p)V$ in cases of both electron-injecting and hole-injecting contacts. But in cases of one electron-injecting contact and one normally hole-blocking contact, electroluminescence appears at the time corresponding to the transit

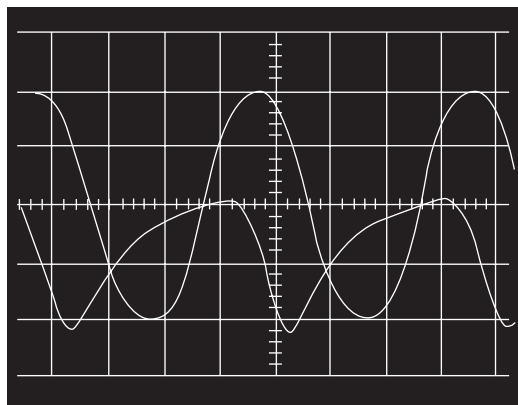
time of electrons alone, that is, on the arrival of the electron space charge front at the anode to enhance the hole injection. In cases with two carrier-injecting contacts, electroluminescent intensity is proportional to the current, irrespective of the current level, while in cases with only one carrier-injecting contact and one normally blocking contact, the relationship between the electroluminescent intensity and the current depends on the current level.¹⁶¹

Electroluminescent intensity depends on the frequency under applied sinusoidal AC fields. Using a sodium electrode as the electron-injecting contact and a silver electrode as the hole-injecting contact (a silver electrode is a hole-blocking contact at low fields and can become a hole-injecting contact only at high fields), Kunkel and Kao have studied electroluminescence under continuously sinusoidal AC fields at various temperatures.¹⁶⁶ The typical wave forms of the AC voltage applied across the specimen and the corresponding electroluminescence produced in it are shown in Figure 3-47. It can be seen that electroluminescence appears only when the silver electrode is at the positive voltage half-cycles, indicating that neither the sodium electrode injects holes nor the silver electrode injects electrons, that there is a time delay (t_d) between the time when the voltage is applied and the time when the electroluminescence appears, and that there is also a phase shift (θ) between the peak of the applied voltage and the peak of the electroluminescent intensity.

Peak electroluminescent brightness as a function of frequency for various temperatures is shown in Figure 3-48. Brightness decreases monotonically with increasing frequency. Brightness increases with increasing temperature, reaches a peak at a certain critical temperature, and then decreases with increasing temperature in a manner similar to that for DC electroluminescence, discussed earlier. The delay time (t_d) shown in Figure 3-47 is also frequency-dependent, as shown in Figure 3-48 in a dashed curve. The phase shift θ also increases with increasing frequency; it is not noticeable at 20 Hz but becomes significant at 5000 Hz. The leading portion of the wave form of



(a)



(b)

Figure 3-47 Oscilloscope traces illustrating the waveform of AC voltages and that of corresponding electroluminescence, which appears only when the silver electrode is at the positive-voltage half cycle, t_d is the time delay and θ is the phase shift. (a) frequency: 500 Hz, temperature: 40°C, vertical scale: voltage, 400 V/div; brightness, 0.5 V/div; horizontal scale: time, 0.5 msec/div; (b) frequency: 5000 Hz, temperature: 20°C, vertical scale: voltage, 400 V/div; brightness, 0.1 V/div; horizontal scale: time, 0.05 msec/div.

electroluminescence is similar in shape to that of the applied voltage for all frequencies, but the tailing portion is different. There is a significant exponential tail developed on the trailing edge at high frequencies, as shown in Figure 3-47(b). Usually, when the applied voltage reaches the zero point, the electroluminescence wave form becomes exponential, since beyond this point light output results predominantly from delayed fluorescence.

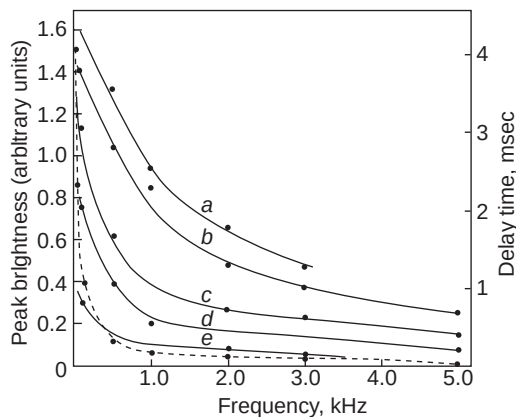


Figure 3-48 The peak electroluminescent brightness (solid curves *a–e*) and the delay time (dashed curve) as a functions of frequency of the applied sinusoidal AC voltage of 1600 V peak to peak at various temperatures, *a*: 20°C, *b*: 30°C, *c*: 40°C, *d*: 0°C, *e*: –20°C. The delay time is measured at 40°C.

Carrier Injection through P–N Junctions

Most LEDs and semiconductor lasers are based on the principle of p–n junctions under forward-bias conditions. Several p–n junction structures formed by luminescent materials produce injection electroluminescence.¹⁶⁷ In this section, we will briefly describe some of the most commonly used p–n junction devices for electroluminescence.

P–N Homojunction Devices—Since group IV elements, such as Si and Ge, are not luminescent materials, and group II–VI compounds cannot be readily doped to form n- or p-type materials, group III–V compounds are the major luminescent materials for p–n junctions. For example, p–n junctions are readily formed in GaAs by doping with Si as donors by replacing Ga, or as acceptors by replacing As. Similarly, p–n junctions can be formed by diffusing Zn into pulled crystals of n-GaAs.

P–n junctions can be made as step (or abrupt) junctions, in which the p-type semiconductor is uniformly doped with acceptor impurities and the n-type semiconductor is uniformly doped with donor impurities up to the metallurgical junction. P–n junctions can also be made as

graded junctions, in which the concentration of the doped acceptor and donor impurities are distributed either linearly or following a particular distribution function up to the metallurgical junction. However, when a p–n junction is formed, the requirement of the constant Fermi level throughout the device in thermal equilibrium results in the formation of a depletion region due to the transfer of electrons and holes across the junction, as shown in Figure 3-49. A full discussion of the p–n junctions is beyond the scope of this book. For details, see the standard references, such as reference 168. Here, we will use abrupt p–n junctions to illustrate the basic concept of p–n junctions' production of light emission.

The basic normal p–n junction is shown in Figure 3-49(a). For example, GaAs has an energy bandgap of 1.43 eV corresponding to the wavelength of 861 nm. The most probable energy of the electrons in the conduction band is $kT/2$; thus, the wavelength of the emitting light due to band-to-band transition is slightly shorter than 861 nm. Furthermore, the light produced will be attenuated due to self-absorption during its travel to the surface of the semiconductor (i.e., the semiconductor–air interface). The light intensity will be reduced by $\exp(-\alpha\ell)$, where ℓ is the length of the semiconductor. Reaching the surface, part of the light will be reflected according to the reflection coefficient R given by Equation 3-57. So, the external quantum efficiency η_q can be written as

$$\eta_q = \eta_{\text{int}}(1 - R)(1 - \cos\theta_c)\exp(-\alpha\ell) \quad (3-150)$$

where $\theta_c = \sin^{-1}(1/n)$, n is the refractive index of the semiconductor, and η_{int} is the internal quantum efficiency, which is given by

$$\eta_{\text{int}} = \frac{\text{Radiative Transition}}{\left(\text{Radiative Transition} + \text{Nonradiative Transition} \right)} \quad (3-151)$$

In general, all recombination processes involve both radiative and nonradiative transitions. For radiative transition, the effective carrier lifetime is τ_R , and for nonradiative transition it is τ_{NR} . The internal quantum efficiency can also be expressed as

$$\begin{aligned} \eta_{\text{int}} &= \frac{1/\tau_R}{1/\tau_R + 1/\tau_{NR}} \\ &= \frac{1}{1 + (\tau_R/\tau_{NR})} \end{aligned} \quad (3-152)$$

Thus, to achieve high η_{int} , the τ_{NR} must be made as long as possible. However, although η_{int} may be close to 100%, the actual η_q in practical LEDs is generally less than 10%.

Since radiative transition is not limited to band-to-band recombination, depending on the dopants and their concentration, radiative transition may take place between the conduction band and the acceptor level or between the valence band and the donor level. In this case, the peak of the light emission spectrum may not occur at 861 nm but at a longer wavelength. For example, the peak of the electroluminescent emission spectrum for an Si-doped GaAs LED occurs at wavelengths between 910 and 1020 nm.

In general, light emission intensity is proportional to forward-biased current I at high current levels, that is, under high forward-bias conditions, in which the radiative transition occurs mainly outside the depletion region due to diffusion current. At low current levels, that is, under low forward-bias conditions, the radiative transition takes place mainly in the depletion region; hence, the light emission intensity is proportional to I^2 due to the recombination current in the depletion region.

To reduce the reabsorption (or self-absorption) of the emitting light in the semiconductor, p–n junction LEDs usually have one side much more heavily doped, as shown in Figure 3-49(b), or both sides heavily doped, as shown in Figure 3-49(c), since the emitting photons have an energy of E_g lower than the reabsorption energy level, which is $E_g + \Delta E$. When both sides of the junction are heavily doped, the n-side becomes n^+ degenerate material and the Fermi level lies inside the conduction band. Similarly, for the p^+ degenerate material on the p side, the Fermi level lies inside the valence band. In this case, when the junction is forward-biased with a voltage nearly equal to E_g/q , the numbers of electrons and holes injected across the junction are sufficient

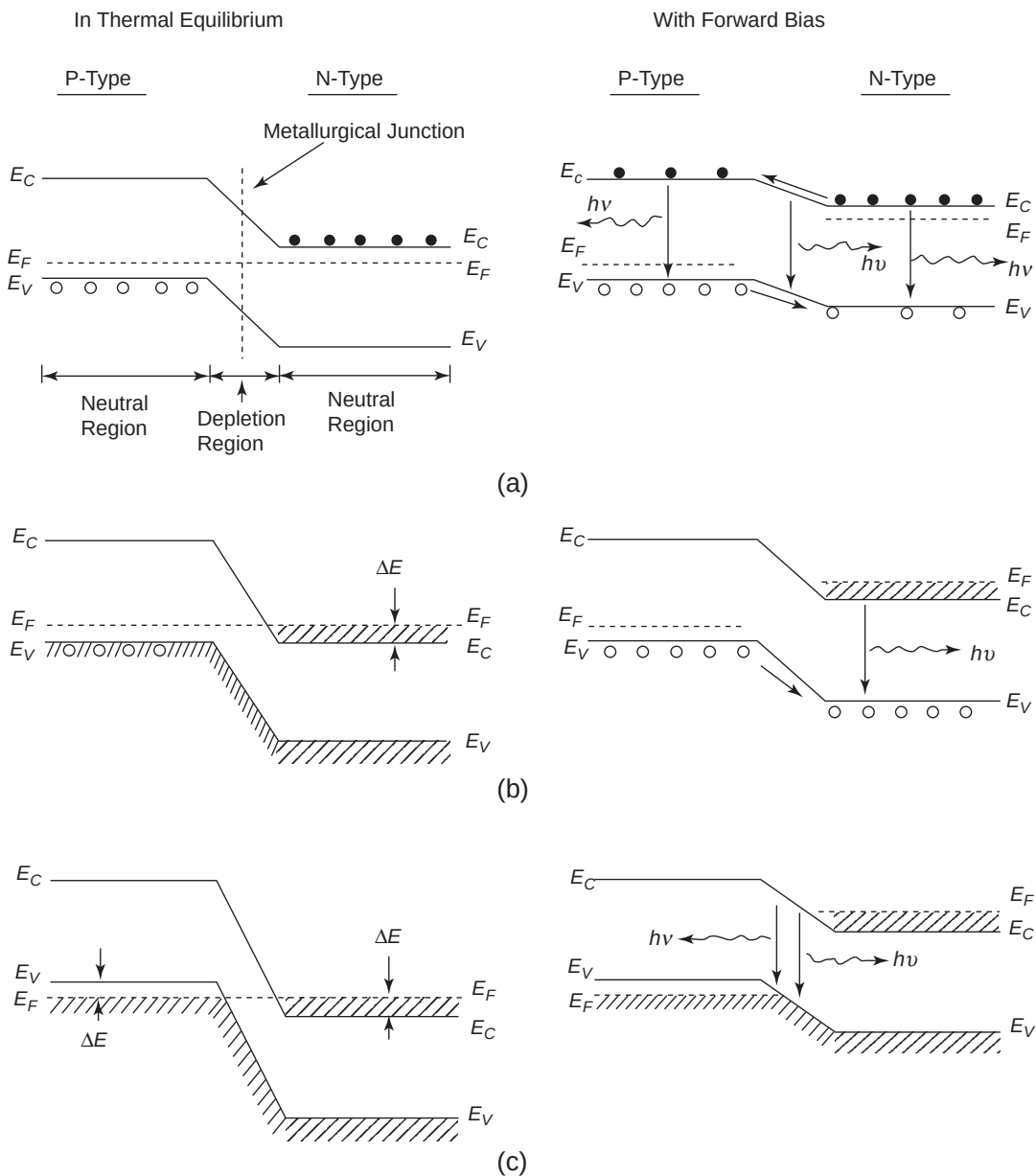


Figure 3-49 Schematic diagrams illustrating p-n junctions for injection electroluminescence, (a) the normal p-n junction, (b) the p-n junction with the n-side heavily doped, and (c) the p-n junction with both the n-side and the p-side heavily doped.

to create a population inversion in a narrow region, called the *active region*. The thickness of the active region is approximately of the order of the minority carrier diffusion length. The radiation produced by recombination in

that region may interact with valence electrons and be absorbed, or interact with electrons in the conduction band, thereby stimulating the production of further photons of the same energy. If the injected carrier concentration

becomes large enough, the stimulated emission can exceed the absorption, so the optical gain can be achieved in the active region, resulting in laser oscillations.

P-N Heterojunction Devices—When the p-type material and the n-type material, having different energy bandgaps, form a p-n junction, such a junction is called the *p-n heterojunction*, as shown in Figure 3-50(a). When a forward-bias voltage across the junction is sufficient to flatten the valence band edges, as shown in Figure 3-50(b), holes are injected freely into the n-type semiconductor without a barrier. But it is difficult for the minority electrons to be injected into the p-type semiconductor because of the large barrier height. Obviously, radiative recombination will then occur in the lower-gap semiconductor, as shown in Figure 3-50(b). For example, in the p(GaAs)-n(GaSb) heterojunction, the injection electroluminescence occurs at about 0.7 eV, which is near the energy bandgap of n-GaSb. The larger-gap semiconductor side, such as p-GaAs is, in general, transparent to the radiation generated in the lower-gap semiconductor and therefore serves as a window for transmitting the radiation.

In practice, heterojunctions suffer from interfacial problems. In order to make a compromise with this interfacial difficulty, p-n heterojunctions can be made from different compositions of miscible alloys having similar lattice constants, such as $\text{Al}_{1-x}\text{Ga}_x\text{As}$. This alloy can be

made either n-type or p-type by adjusting the value of x .¹⁶⁸

3.5 Photoemission

Photoemission is generally referred to as emission of electrons into a vacuum from a solid, or injection of electrons or holes from a solid into another solid, resulting from an interaction between photons and a solid. The photoemission of carriers from a solid into a vacuum is referred to as external photoemission, while that from a solid into another solid (such as from a contact electrode to a semiconductor) is referred to as internal photoemission. Photoemission has been extensively studied as a tool to study energy levels, particularly ionization energies of molecular crystals,^{169,170} as well as to obtain information about the injecting contact, potential barriers, surface states, and electronic structures of solids.^{171–173} In this section, we shall discuss photoinjection from a contacting electrode to a crystalline solid and photoemission from a crystalline solid to a vacuum.

3.5.1 Photoemission from Electrical Contacts

An extensive compilation of experimental data on this subject is available in the literature for inorganic semiconductors,^{171,174} as well as for organic semiconductors.¹⁷³ Metallic contacts

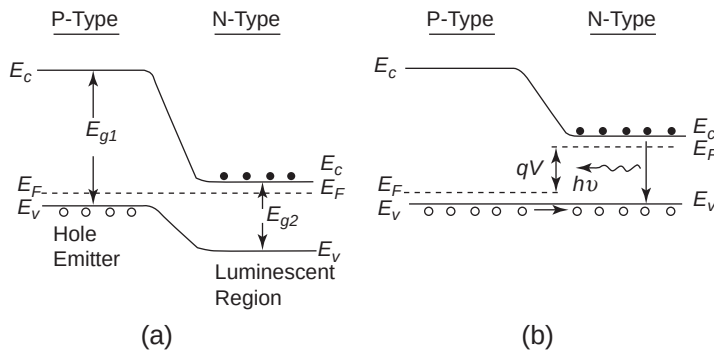


Figure 3-50 Schematic diagrams illustrating a p-n heterojunction for injection electroluminescence: (a) in thermal equilibrium and (b) with forward bias to bring the valence band edges of both sides to the same level.

can be roughly classified into two groups: ohmic contacts and blocking contacts.¹⁷⁵ An ohmic contact can be considered a reservoir of carriers that are always ready to supply as many carriers as needed. Usually, at a given field, the ohmic contact can supply more carriers than the bulk material can carry. Thus, the current is bulk-limited. Further increase in current-injection by photoexcitation does not affect the current; therefore, no photoemission current can be observed.

A blocking contact can inject only a very few carriers—much less than the bulk material can carry. Thus, the current is contact limited. If a light of energy $h\nu \geq \phi_B$ (where ϕ_B is the potential barrier height—the difference between the Fermi level of the metal and the conduction band edge of the semiconductor, as shown in Chapter 6, Figure 6-4) illuminates the metallic contact, photoinjection from the contact will take place. The photoemission, without taking into account the effects of scattering and relaxation, can be written, according to the Fowler theory^{176,177} as

$$J_{ph} = C(h\nu - \phi_B)^2 \text{ for } h\nu \geq \phi_B \quad (3-153)$$

This equation is valid only for $\phi_B \geq 0.5$ eV, that is, if $h\nu - \phi_B$ is greater than some multiple of kT ($>6kT$) or $h\nu \leq 1.5\phi_B$. Beyond this range, the assumptions for the simple Fowler theory are no longer valid.¹⁷⁸ Furthermore, Equation 3-153 is valid only for photoinjection from a metal to a semiconductor having wide energy bands (with bandwidths larger than 0.5 eV). Most metals, inorganic semiconductors, and insulators can satisfy these conditions. For materials and conditions for which Equation 3-153 is valid, the measurements of J_{ph} as a function of $h\nu$, and then the extrapolation of the plot of $J_{ph}^{1/2}$ versus $h\nu$ to $J_{ph} = 0$, yields ϕ_B . However, Equation 3-153 cannot be applied to narrow energy band materials, such as most low-mobility organic semiconductors (e.g., anthracene). In narrow energy band materials, an electron injected into the narrow band can diffuse only a few angstroms (about 5 Å) and would then be captured in a bound state of the coulomb image potential (see Figures 6-12 and 6-13). There-

fore, an injected electron must have a sufficient momentum perpendicular to the surface before it can escape from the image force to enter the conduction band.

From a Metal into Wide Bandwidth Semiconductors

This is the most common case, and it has been treated in detail by investigators.¹⁷⁷⁻¹⁷⁹ As mentioned before, in this case, photoemission measurements would directly yield the barrier height of the blocking contact ϕ_B . For semiconductors for which the image-force effect is negligible, the photoemission method is reasonably accurate for determining ϕ_B . A typical example is given in Figure 3-51 for a gold contact to a CdS crystal specimen, in which $\phi_B = 0.7$ eV is obtained for this contact potential barrier height. The experimental data are from Goodman.¹⁸⁰ A similar agreement to the Fowler theory has also been reported for Si and for other inorganic semiconductors.¹⁸¹ By measuring the threshold energy $h\nu_0 = \phi_B$ for electron emission and that for hole emission, the sum of these two threshold energies gives the energy band gap of the semiconductor. For example, the threshold energy for electron emission from Au to n-GaP is 1.30 eV¹⁸² and that for hole

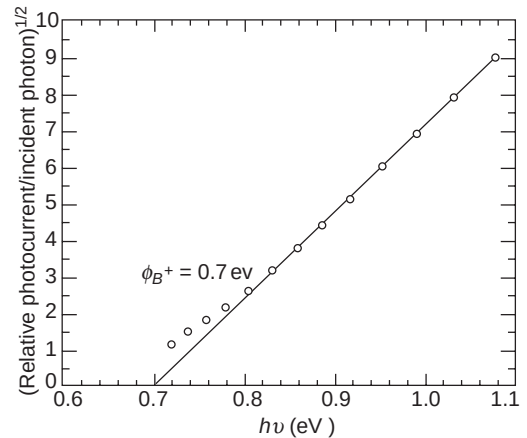


Figure 3-51 $(\text{Relative photocurrent per incident photon})^{1/2}$ as a function of photon quantum energy for an evaporated gold contact to a cadmium sulfide crystal specimen.

emission from Au to p-GaP is 0.72 eV.¹⁸³ The sum of these is 2.02 eV, which is about the energy band gap of GaP, which is 2.18 eV.¹⁸⁴ It should be noted that the barrier heights determined by means of photoemission measurements depend on the specimen's surface conditions. The effects of surface states have been discussed by several investigators.^{180,181}

For large bandgap materials, the Fowler plot at $h\nu$ close to $h\nu_0$ is usually not linear, which means $(J_{ph})^{1/2}$ is not proportional to $(h\nu - \phi_B)$, or deviates from the linear Fowler plot. Figure 3-52 shows a typical $(J_{ph})^{1/2}$ versus $h\nu$ curve for an MIM (metal-insulator-metal) structure; the results are from Kadlec and Gundlach.¹⁸⁵ The asymmetry of Al-Al₂O₃-Al ($\phi_1 > \phi_2$) shown in

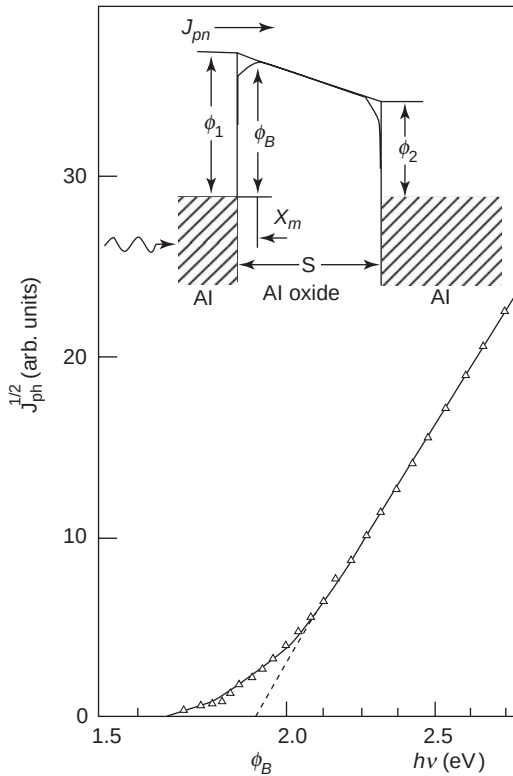


Figure 3-52 (Photoemission current)^{1/2} as a function of incident photon quantum energy for an Al (1000 Å) – Al₂O₃ (30 Å) – Al (200 Å) structure with the 200 Å-thick top electrode illuminated. The corners of the barriers are rounded off due to the image force, as shown in the inset.

the inset of Figure 3-52 is probably due to the different prehistories of the two interfaces. This nonlinear behavior or deviation has been observed by many investigators,^{186,187} and makes it difficult to determine accurately the value of ϕ_B . Several factors may be responsible for this deviation¹⁷⁸:

- The electron distribution may be smeared out about the Fermi level, but the deviation is too large to be explained by this effect.
- The scattering of electrons in the conduction band may play a role, but it is likely that it can be important for thick specimens and not for films thinner than 50 Å.
- The quantum mechanical transmission coefficient, which affects the photocurrent, is not equal to zero for $E_x < \phi_B$, because some of the electrons can tunnel through the potential barrier, and is not equal to 1 for $E_x > \phi_B$, because some of the electrons are reflected, where $E_x = mv_x^2/2$ and v_x is the electron velocity in x direction.
- It is possible that the barrier height is not uniform over the whole area of the interface.

It should be noted that the barrier height may depend on specimen thickness and light intensity, possibly due to space charge effects.¹⁸⁸ Obviously, because of the image force effect, the barrier height is field dependent. Berglund and Powell¹⁸⁹ have derived the expression for the photoemission current versus applied voltage characteristics for an MIM structure. It is given by

$$J_{ph} \propto \left\{ h\nu - \phi_B - \left[\frac{q}{4\pi\epsilon S} \times (qV + \phi_1 - \phi_2) \right]^{1/2} \right\}^p \exp\left(-\frac{x_m}{\lambda}\right) \quad (3-154)$$

where S is thickness of the specimen, x_m is the distance from the illuminated electrode to the location of the maximum barrier height, λ is the mean free path, and p is a parameter depending on the kind of electron excitation and lies $1 < p < 3$. If the insulator is too thin, the incident light illuminating the first electrode can

also reach the second electrode (opposite to the first) and excite carriers there. Consequently, the photoemission current is composed of two types of currents (electrons and holes) under an applied field. Unless the component from the opposite electrode is negligibly small, the photoemission consisting of two types of carriers would influence the thickness and $h\nu$ dependence of J_{ph} . Furthermore, the electron–electron and electron–phonon interactions also play an important role in this dependence. All these effects on the internal photoemission—as well as the application of various internal photoemission experiments for the determination of barrier height, potential barrier shape, charge distribution in the insulator, topographical distribution of the barrier height, trap distribution in the barrier, and the mean free paths and energy losses of the carriers—have been critically reviewed by Kadlec and Gundlach.^{178,186,190}

From a Narrow Bandwidth Emitter into Wide Bandwidth Semiconductors

A degenerate n-type or p-type silicon or similar semiconductor can act as a narrow bandwidth emitter. Photons with energies $h\nu_0$ or $h\nu_1$ can excite electrons from the conduction band, and those with energy $h\nu_2$ can excite electrons from the valence band of a degenerate n-type semiconductor into the conduction band of an insulator, as shown in Figure 3-53. For a given photon energy $h\nu$, the excited electrons are distributed over only a narrow band of the width

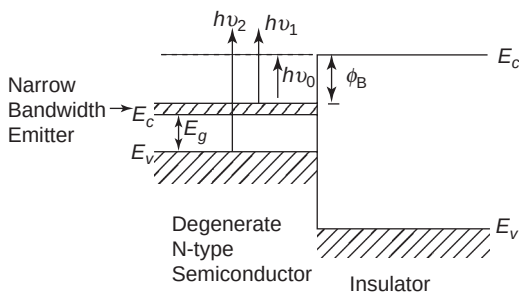


Figure 3-53 Various possible transitions for photoemission of electrons from a degenerate n-type semiconductor into an insulator.

within a few kT of the bottom of the conduction band, far narrower than those excited from the metal. Thus, within the range of photon energies $\phi_B \leq h\nu \leq \phi_B + E_g$, the photoemission current (or the quantum yield for photoemission) as a function of photon energy is simply

$$J_{ph} = C(h\nu - h\nu_0) \quad (3-155)$$

For the Si-SiO₂ system, the experimental results fit Equation 3-155 if the photon energies $h\nu$ are less than $\phi_B + E_g$, and then fit Equation 3-153 (rather than Equation 3-155) when $h\nu > \phi_B + E_g$, because at higher photon energies, the electrons excited from the valence band become dominant, as shown in Figure 3-53.

From a Metal into Narrow Bandwidth Crystals

The narrow band can be considered a delta function. This means that within the band, the energy distribution of excited holes (or excited electrons) is uniform and implies that the quantum yield for photon energies above the threshold $h\nu_0$ is independent of photon energy, as shown in Figure 3-54(a). In organic crystals, there are usually several narrow bands separated by roughly equally spaced levels, due to molecular vibrations. For example, in anthracene, the molecular vibration has frequencies corresponding to energies of approximately 0.2 eV, but the electronic bandwidth is about 0.02 eV. For clarity, only two such narrow valence bands are shown in Figure 3-54(b). Each band gives rise to a step-function of photoemission current, starting at different photon energies. The combination of these step-functions gives a staircase quantum yield Y , and the derivative $dY/d(h\nu)$ gives a clear picture of energy-level splitting by molecular vibration, as shown in Figure 3-54(b).

Figure 3-54(b) gives only the basic concept of the behavior of narrow bands. In fact, the quantum yield of the i th narrow band should be proportional to $(h\nu - \phi_{Bi})$, since the number of excited carriers that can surmount the potential barrier ϕ_{Bi} is proportional to $(h\nu - \phi_{Bi})$, as shown in Figure 3-54(c). See also the top energy diagram of Figure 3-54(b). The total quantum

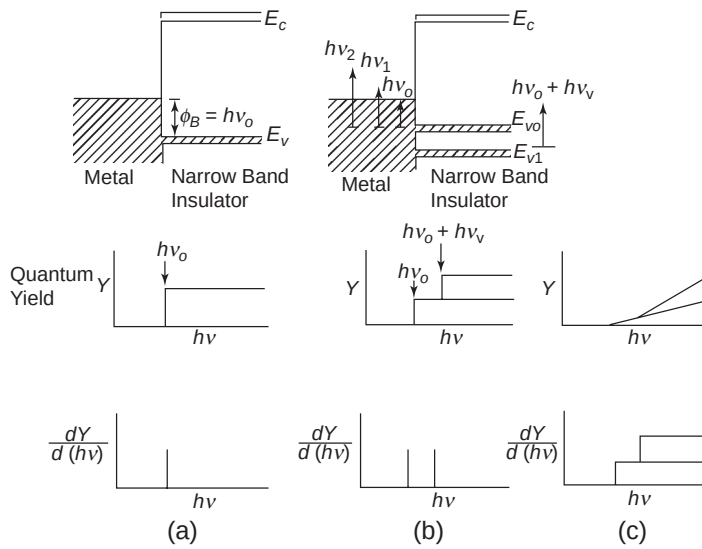


Figure 3-54 Schematic diagrams illustrating photoemission from metal into narrow-band crystals: (a) with one narrow band when the acceptance angle is not considered; (b) with two narrow bands when the acceptance angle is not considered; (c) with two narrow bands when the acceptance angle is taken into account.

yield Y from all the vibrationally split levels should be the sum of the linear ramps, as shown in Figure 3-54(c). In this case, the derivative $dY/d(h\nu)$ gives a series of step-functions, starting at different photon energies. For photoemission from a metallic contact to an organic crystal, the quantum yield Y should follow the relation schematically shown in Figure 3-54(c) and $Y \propto (h\nu - \phi_B)^2$, rather than those shown in Figure 3-54(b) with $Y \propto (h\nu - \phi_B)$.

Several investigators have reported that the quantum yields for the photoemission of holes into organic semiconductors, measured as functions of photon energy using various metallic electrodes, are in good agreement with the theoretical prediction discussed here.^{171–174,191,192} Obviously, these arguments for photoemission of holes from metal to narrow band crystals apply equally well to cases of photoemission of electrons from metal to narrow-band crystals. It should also be noted that, from the arguments given earlier in this chapter in Photoemission from Electrical Contacts, it is possible to estimate the photoemission spectrum in cases of a narrow bandwidth emitter into crystals with a series of narrow bands.

3.5.2 Photoemission from Crystalline Solids

In general, photoemission from crystalline solids means volume photoemission resulting from optical absorption in the bulk, which should be distinguished from surface photoemission resulting from optical absorption at the surface. The contribution from excitation of surface states is small, because the total number of surface states is small compared to the number of states in the bulk that can participate in photoemission processes. Therefore, photoemission is mostly a volume effect. The photoemission quantum yield increases with increasing specimen thickness and reaches a saturation value when specimen thickness exceeds the penetration (or absorption) depth or the escape depth. If specimen thickness is smaller than penetration depth, photoemission may exhibit a roughly periodic nature of its variation with specimen thickness, going through a minimum at thicknesses equal to an odd number of quarter-wavelengths and through a maximum at thicknesses equal to an integral number of half-wavelengths of the

incident light.¹⁹³ To simplify matters, such an interference effect is ignored in the following discussion, and photoemission is assumed to be a bulk phenomenon.

Photoemission consists of three steps:

1. Excitation of an electron to a high-energy state by the absorption of a photon
2. The scattering of the excited electron on its way to the vacuum-solid interfaces
3. The escape of the electron over the potential barrier at the surface of the solid

Surface states may affect volume photoemission indirectly through band bending. However, for the sake of simplicity, we will neglect the band-bending effects. Figure 3-55 illustrates the three steps, the band diagram, and the relative energy levels for photoemission from an organic crystal into a vacuum. The light intensity of the excitation with photon energy $h\nu$ at the same distance x from the vacuum-solid interface is

$$I(\nu, x) = I_o(1 - R)\exp(-\alpha x) \quad (3-156)$$

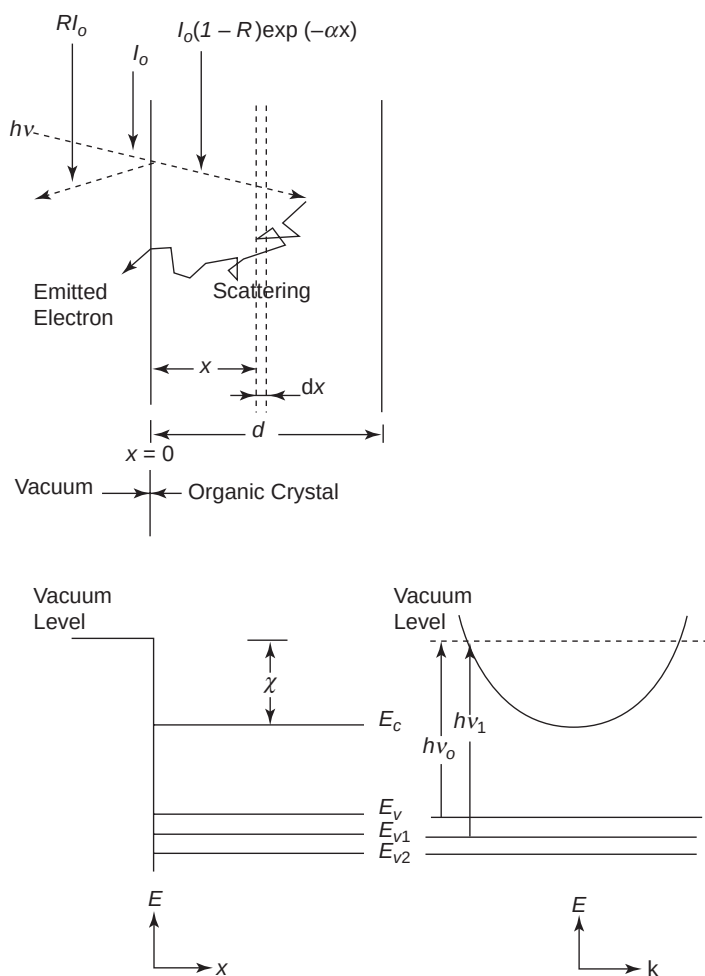


Figure 3-55 Schematic diagrams illustrating excitation, scattering, and escape of an excited electron, the hole bands E_v , the vibrationally excited hole bands E_{v1} , E_{v2} , and so on. The conduction band is assumed to be parabolic, and the valence bands are assumed to be almost independent of k vector. Localized states in the energy gap due to imperfections are not shown for the sake of clarity.

where I_o is the light intensity in vacuum; R and α are, respectively, the reflection and the absorption coefficients of the crystal, which depend on light wavelength. $I(v, x)$ can be defined as the number of photons of energy $h\nu$ arriving at x per second.

Before going to organic crystals, we will take a simple inorganic semiconductor as an example to illustrate the procedures of deriving an expression for the photoemission yield Y . Supposing that the energy of incident photons $h\nu$ is larger than $E_g + \chi$, then the concentration of photocarriers generated within dx at x is

$$\begin{aligned} dn &= P\alpha I(v, x)dx \\ &= P\alpha I_o(1 - R)\exp(-\alpha x)dx \end{aligned} \quad (3-157)$$

where P is the probability that a photon absorbed will excite an electron to a high-energy state. The energy $h\nu > (E_g + \chi)$ may excite some electrons to a high-energy state above the vacuum level, but it may also excite some to a level below the vacuum level if the electrons being excited are located much lower than E_v in the valence band. Clearly, only those above the vacuum level can lead to photoemission. Furthermore, even though some excited electrons at x may have energies above the vacuum level, they will lose part of their energies by the scattering processes during their motion toward the vacuum-solid interface. Thus, by the time they reach the interface, some may have lost so much energy that their energies become lower than the vacuum level, and they will not be able to escape over the potential barrier. The probability for an excited electron at x to escape to contribute to photoemission can be written as¹⁹⁴

$$P_{\text{esc}}(v, x) = B(v)\exp(-x/L_{\text{esc}}) \quad (3-158)$$

where $B(v)$ is constant and L_{esc} is the escape depth. Both $B(v)$ and L_{esc} are dependent on electron energy. The total photocarriers emitted from the crystal specimen are then

$$\begin{aligned} n &= \int_0^\infty P\alpha I_o(1 - R)\exp(-\alpha x)B\exp(-x/L_{\text{esc}})dx \\ &= \frac{P\alpha B I_o(1 - R)}{\alpha + 1/L_{\text{esc}}} \end{aligned} \quad (3-159)$$

The quantum yield is defined as

$$Y = \frac{n}{I_o(1 - R)} = \frac{PB}{1 + 1/\alpha L_{\text{esc}}} \quad (3-160)$$

Since B and L_{esc} are functions of electron energy, they must be directly related to scattering mechanisms and absorption processes (direct transition, without a change in momentum, or indirect transition, involving a change in momentum). Kane has analyzed theoretically the quantum yield versus photon energy characteristics near the threshold for a general band structure and for a variety of photocarrier generation and scattering mechanisms involving volume and surface states (volume and surface scattering processes) in semiconductors.¹⁹⁵ The initial fast-rising portion of the $Y - h\nu$ curve can be expressed in the form

$$Y = A(h\nu - E_{\text{th}})^S \quad (3-161)$$

where A and S are constants and E_{th} is the threshold energy required for photoemission. For intrinsic and lightly doped semiconductors $E_{\text{th}} = \chi + E_g$. Depending on the photocarrier generation and scattering mechanisms, S varies from $S = 1$ to $S = 5/2$.¹⁹⁵ For cases that do not involve surface scattering, the values of S are as follows:

$S = 1$ for direct transition without volume scattering

$S = 2$ for direct transition involving elastic scattering

$S = 5/2$ for indirect transition with or without elastic scattering

In practical cases, photocarriers suffer both volume and surface scattering. However, that $S = 1$ in the high-energy region and $S = 3$ in the region near the threshold has been experimentally observed in III-V compounds, such as InSb, GaSb, InAs, and GaAs.¹⁹⁶ Kane's theory, however, depends on the shape of the edge of the valence band. Therefore, it cannot be applied to organic crystals with narrow valence bands.

In organic molecular crystals, excitons are generally the intermediate states in the photocarrier generation processes. Also, the molecu-

lar vibrations split each valence band into many well-separated narrow bands with an energy separation of one vibrational quantum, $h\nu_v$. The electron escape probability P_{esc} involves both the scattering probability and the surface transmission probability. Near the threshold, P_{esc} may be considered due mainly to surface transmission probability, because scattering probability varies slowly in this energy region.¹⁷⁰ On the basis of this consideration, several investigators^{170,197,198} have derived a relation between the quantum yield and the photon energy near the threshold, which is given by

$$Y \propto (h\nu - E_{\text{th}})^S \quad (3-162)$$

with $\frac{1}{2} \leq S \leq 3$. This expression is similar to Equation 3-161. However, for a large number of molecular crystals, the experimental results on the quantum yield near the threshold are in agreement with Equation 3-162 with $S = 3$, such as anthracene, tetracene, and pentacene.^{170,199} It should be noted that, because many factors may affect the value of S , it is often possible to obtain fits to the experimental data versus $h\nu$ with S other than 3. When photon energy is in the above-threshold region, the value of S may become $1/2$. It is likely that the extrapolation of the plot of Y^2 as a function of $h\nu$ to $Y^2 = 0$ yields the value of E_{th} corresponding to the first maximum in the photoelectron energy distribution spectrum (from the most probable position of the first valence band to the vacuum level), while the cubic law extrapolation gives the value of E_{th} corresponding to the onset of photoemission.¹⁹⁸ For more details about photoemission in molecular crystals, see references 198–200.

Effects of Surface Conditions

In the preceding analysis, the energy bands are assumed to be flat up to the surface. In fact, there are always surface states present at the surface that induce space charge near it, causing the band to bend up or down, as shown in Figure 3-56. Gases such as oxygen adsorbed on the surface may result in a change of the order of 1 eV in E_{th} . If the space charge causes the

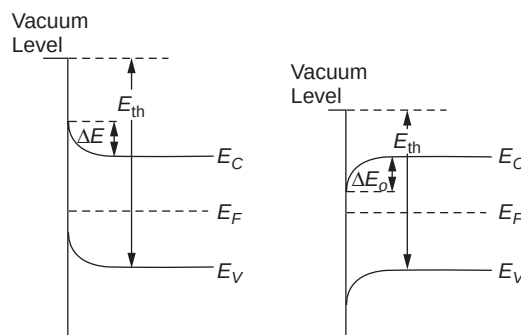


Figure 3-56 The effect of band bending on the threshold for photoemission.

band to bend up by ΔE the threshold energy for photoemission $E_{\text{th}} = |E_{\text{th}}|_{\text{flat band}} + \Delta E$. Conversely, if the band is bent down by ΔE then $E_{\text{th}} = |E_{\text{th}}|_{\text{flat band}} - \Delta E$. To include the effects of surface conditions, we must know the variation of E_v with x measured from the vacuum level. Of course, the analysis for the $Y - h\nu$ relation will be very mathematically involved.

Effects of Defects in Crystals

Obviously, defects in a crystal will affect the photoemission spectrum. It is likely that the adsorbed gases in the crystal lead to the formation of deep traps, particularly in the surface layers of the solid. Electron acceptors, such as O_2 , produce electron traps, while electron donors, such as H_2 , produce hole traps. Local changes in polarization energy and impurities with low ionization energy present in the crystal lattice would cause changes in the photoemission spectrum. In general, the defects tend to broaden the low energy tails, causing a decrease in the threshold. Thus, in studying photoemission from solids, there are two intricate problems: the determination of the threshold photon energy, and the identification of the origin of the emitted electrons. For example, the values of E_{th} change as oxygen is incorporated into tetracene,¹⁹⁴ or for alkali metals such as caesium, potassium, and sodium doped into anthracene crystals.²⁰¹

Energy Distribution of Emitted Electrons

The energy distribution of the emitted electrons can be measured by several techniques. The most commonly used technique is the retarding potential method. This method simply employs a bias, acting as a retarding potential V_R , which collects only those emitted electrons with a kinetic energy E_k greater than the retarding potential energy qV_R . By sweeping the retarding potential over a range of bias, a complete energy distribution curve (EDC) can be obtained.^{202,203} All energies above the vacuum level (see Figure 3-55) are available as kinetic energy for the emitted electrons. So, the incident photon energy E_p must be at least equal to the energy between the valence band edge E_v and the vacuum level E_{vac} . Thus, the minimum retarding potential $V_{R(min)}$ that can completely block the photoemission current should be

$$V_{R(min)} = (E_p - E_{vac})/q \quad (3-163)$$

From the EDC, it is possible to determine the band structure in the crystal,²⁰⁴ as well as to obtain information about the density of states in the valence and the conduction bands.

Figure 3-57 shows schematically the variation of the photoemission current I and the number of emitted electrons per unit times per unit energy N with retarding potential V_R . For the latter, N can be expressed as

$$N = \frac{1}{q} \frac{dI}{d(qV_R)} = \frac{1}{q^2} \frac{dI}{dV_R} \quad (3-164)$$

It can be seen that $N = 0$ when $V_R = V_{R(sat)}$ and $V_R = V_{R(min)}$. N increases with V_R from $V_{R(sat)}$, as expected, since the attenuation of the absorbed optical energy varies exponentially with the depth from the surface of the solid specimen. So, a greater fraction of excited electrons lies in the region close to the surface. Fewer electrons are excited deeper inside the crystal; these electrons will find greater difficulty in emerging because their kinetic energy is close to zero. Thus, electrons excited closer to the surface will be greater in number and their energies close to the maximum kinetic energy. The peak of the $N - qV_R$ curve occurs at $qV_R = E_A$, imply-

ing that the loss of electron energy for $qV_R < E_A$ results in a decrease in the number of emitted electrons with kinetic energies less than E_A . There are many possible causes of the loss of electron energy. For example, the loss may be due to the excitation of bound molecular states by the electrons before they emerge from the crystal. The width of the peak at E_A could be due to a variety of causes; for example, not all of the emitted electrons travel in the direction perpendicular to the surface of the collecting electrode [see the insert of Figure 3-57(a)] due either to scattering, to the excitation of bound molecular states, or to a particular distribution in the intrinsic density of states. Furthermore, the vibrational modes of the molecular ions, as well as the fluctuation of intermolecular electronic polarization, may also cause N to decrease for qV_R greater and smaller than E_A . The second peak at $qV_R = E_B$ may be due to the structure of the crystal. Depending on the incident photon energy and the structure of the crystal, the $N - qV_R$ curve may exhibit one peak, two peaks, or one peak with several shoulders. For example, for the inorganic cesiated GaAs crystal, the $N - E_k$ curve exhibits one peak if it is excited by photons of about 1.4 eV. But at photon energies greater than 1.7 eV, the valleys can take electrons. If the photoemission time is comparable to the intervalley relaxation time, then the crystal will exhibit two peaks in the $N - E_k$ curve.²⁰⁵ For the organic anthracene crystal, there are many shoulders before the occurrence of the peak in the $N - E_k$ curve.¹⁷⁰ A wide variety of phenomena can be studied using the experimental data on the photoemission of electrons with a definite kinetic energy.

Multiquantum Processes

So far, our discussion has been limited to phenomena based on one-quantum processes. For example, when an anthracene crystal is excited by light of the energy $h\nu < 5\text{ eV}$ there is still photoemission below the ionization threshold, which is 5.7 eV. But in this case, the quantum yield Y varies as the square of the incident photon energy. This phenomenon was first

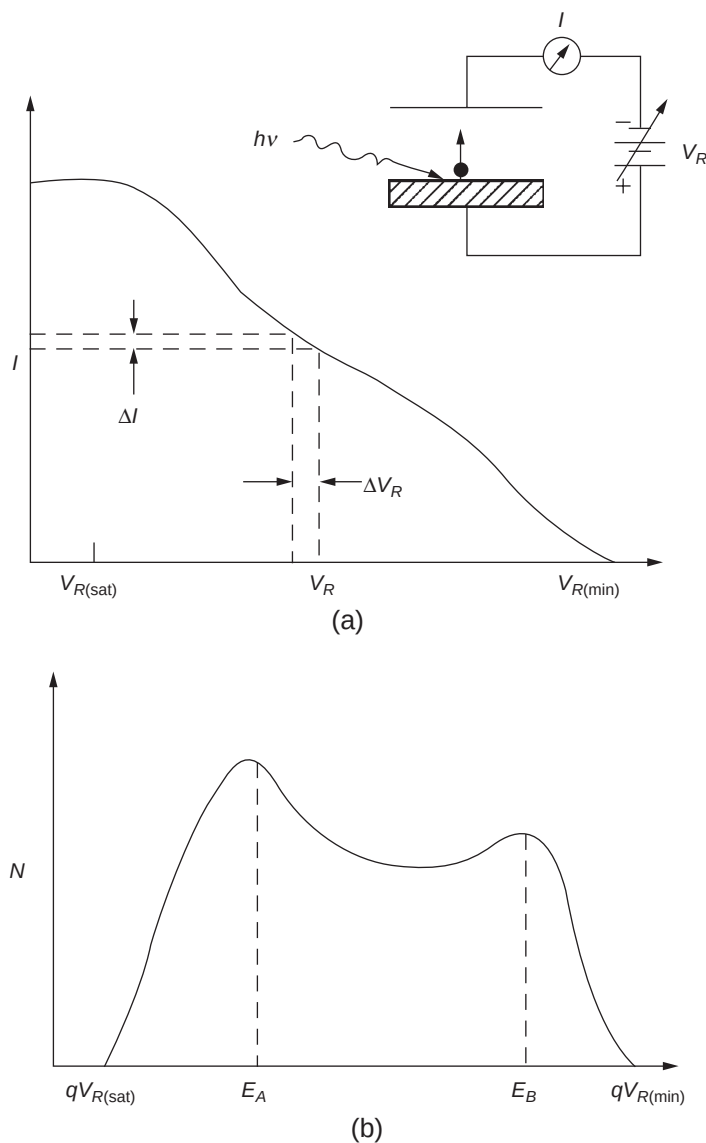


Figure 3-57 Schematic diagrams illustrating (a) the variation of the photoemission current with the retarding potential and the simple set-up for this measurement and (b) the corresponding variation of the number of emitted electrons per unit time per unit energy N with their kinetic energy represented by qV_R .

discovered by Pope et al.²⁰⁶ It can only be explained by the double-quantum processes.

If the rate of photoemission is proportional to light intensity, excitation is a one-photon process; if it varies quadratically with light intensity, excitation is a two-photon process. The mechanisms that may contribute to this

double-quantum process are direct two-photon excitation, exciton photoionization, exciton–exciton collision ionization, exciton–conduction electron ionization, and conduction electron ionization. The excitons involved could be neutral singlet or triplet excitons or ionic excitons.¹⁶⁹

3.6 Photovoltaic Effects

Photovoltaic effects generally refer to phenomena resulting from the conversion of light energy into electrical energy. This conversion process can be considered the reverse of electroluminescence. However, the criteria for achieving the conversion are different. For electroluminescence, only radiative recombination produces light, and the carriers involved in producing light can be free or bound carriers. But for photovoltaic effects, optical absorption must produce free carriers, and the carriers involved in generating photovoltages must be free (mobile) carriers.

A photovoltaic device is a photodiode in which electron-hole pairs can be generated by photon absorption, and they will be separated by a force due either to the diffusion of these photogenerated carriers with different mobilities in the bulk, to the contact potential (or diffusion potential) associated with the contact between two different materials, or to photosynthesis. We may detect this charge separation in two ways. If the device under photon radiation is left on an open circuit, then the potential between two terminals can be measured as the open-circuit voltage. This is known as the photovoltaic mode of operation. On the other hand, if the device under photon radiation is short-circuited, then an external current can be measured under the short-circuit condition or through a small resistance. This is known as the photoconductive mode of operation.

Photovoltaic effects in solids may therefore be caused by

Bulk photovoltaic effects: A photovoltage arises due to the diffusion of nonequilibrium photogenerated carriers with different electron and hole mobilities in the bulk of the solid.

Contact potential photovoltaic effects: A photovoltage arises due to the potential barrier at the interface between two different materials, such as the Schottky barrier at the metal-semiconductor or metal-insulator contacts; the p-n junction between a p-type and an n-type semiconductor; or the p-i-n

structure, with an insulator between a p-type and an n-type semiconductor.

Photosynthesis photovoltaic effects: A photovoltage arises due to the photosynthesis of a dye sensitizer and an electrolyte.

Anomalous photovoltaic effects: A photovoltage arises due to a combination of several mechanisms, such as the Dember effect in microregions, photovoltaic effects at p-n junctions, Schottky barriers or strains at grain boundaries.

Any photovoltaic effects occurring in a solid involve

- Light absorption in the solid
- Mobile charge carriers generated by the absorption
- An internal discontinuity or nonuniform distribution of impurities or defects giving rise to the creation of an internal electric field to separate the two types of carriers
- Electrical contacts

In this section, we shall discuss briefly various photovoltaic effects.

3.6.1 Bulk Photovoltaic Effects

As early in 1931, Dember observed an electric potential developed across a cuprous oxide specimen when subjected to illumination with a strong light in the absorption region of this material. Later, this phenomenon was referred to as the *Dember effect*²⁰⁷ and was soon found also in diamond and zinc sulfide,²⁰⁸ in Ge and Si,^{91,209,210} in CdS,²¹¹ and in organic crystals.²¹² This phenomenon is attributed to the diffusion of photogenerated electrons and holes with different mobilities. In general, the spectral distribution of the photovoltage amplitudes correlates closely to the absorption spectrum, and the change in light intensity changes only the photovoltage amplitude; it does not affect this correlation. Photovoltaic behavior is sensitive to the surface condition of the device. Adsorption of oxygen on the crystal surface may reduce greatly the photovoltage amplitude.

The Dember Effect

The light-induced creation of a photovoltage in the bulk of a solid is referred to as the Dember effect. Conditions necessary for this effect to occur are nonuniform illumination, which gives rise to a concentration gradient of the photo-generated carriers, and photogenerated carriers with different mobilities. Under these conditions, when the surface of a solid specimen is illuminated with a light of high intensity, a high concentration of photogenerated electron-hole pairs will be produced near the illuminated surface, as shown in Figure 3-58(a).

In general, the mobility of electrons is higher than that of holes for most inorganic semiconductors (usually, the reverse is true for most organic semiconductors). If this is the case, electrons will diffuse more rapidly than the holes,

leaving a positive space charge near the illuminated surface and a negative charge within the bulk of the specimen. The internal electric field set up by this charge distribution tends to oppose the electron flow and to assist the hole flow, so the carriers drift in a manner that reduces the space charge. For low-resistivity materials, the Debye length is small, implying that the length over which a charge imbalance is neutralized by the majority of carriers is small. In other words, the overall internal field created is considerably smaller than would be obtained for high-resistivity materials. Thus, for high-resistivity materials, the excess photogenerated carrier concentration is larger than the thermal equilibrium carrier concentration. For this case, the open-circuit photovoltage, usually called the *Dember voltage*, can be derived on the basis of large signal theory. It is given by

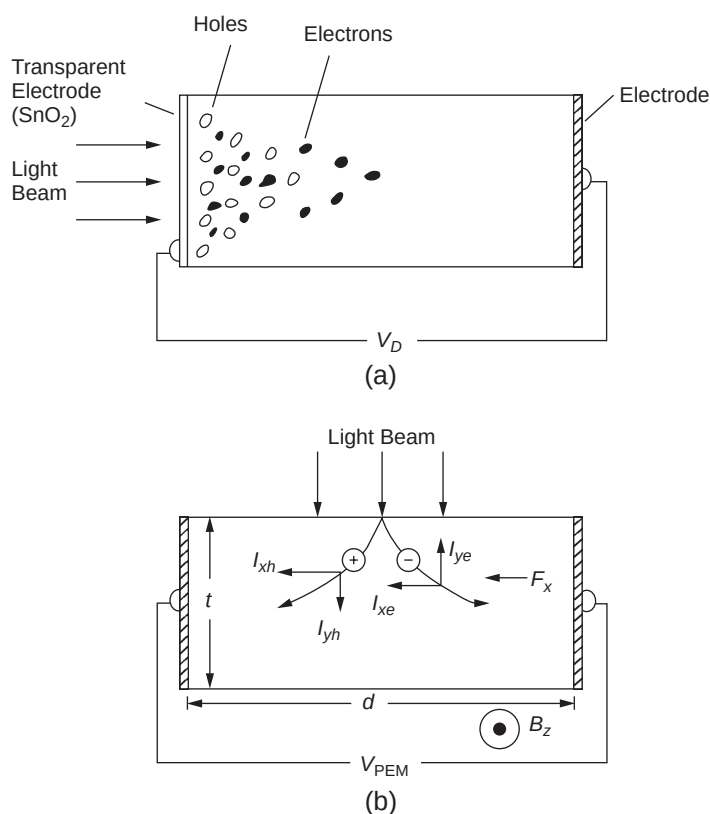


Figure 3-58 Schematic diagrams illustrating the basic arrangements for measurements of (a) the Dember voltage V_D and (b) the photoelectro-magnetic (PEM) voltage V_{PEM} .

$$V_D = \frac{kT}{q} \frac{\mu_n - \mu_p}{\mu_n + \mu_p} \ln \left[1 + \frac{(\mu_n + \mu_p) \Delta n}{\mu_n n_o + \mu_p p_o} \right] \quad (3-165)$$

where μ_n and μ_p are the mobilities of electrons and holes, respectively; n_o and p_o are the concentrations of thermally equilibrium electrons and holes; and $\Delta n = \Delta p$ is the photogenerated excess electron concentration (equal to the excess hole concentration).⁹⁹ It should be noted that there are always traps in solids, and that the electrical contacts to the solid specimen are usually not ohmic, resulting in the formation of a surface potential barrier. The electric field associated with such a barrier would partially separate the electrons and holes, which are injected to the specimen due to optical illumination in the vicinity of the surface. This would create a photovoltage indistinguishable from the Dember voltage. However, the Dember voltage is usually very small, and the Dember effect has so far received little attention because of its low efficiency for use as solar cells. Most photovoltaic effects are based on other, more efficient ways of photo-generation of free carriers and their separation for the creation of a high internal electric field.

The Photoelectro-Magnetic (PEM) Effect

The photoelectro-magnetic (PEM) effect, which is also called the photomagneto-electric (PME) or the magneto-photovoltaic (MPV) effect, was originally discovered in cuprous oxide by Kikoin and Noskov in 1934²¹³ and later studied by many investigators.^{91,214,215} The PEM effect is illustrated schematically in Figure 3-58(b). When a slab of photoconductor is illuminated perpendicularly with light within the intrinsic absorption band of the material, electron-hole pairs are generated in a layer close to the illuminated surface. These photo-generated electrons and holes will set up a concentration gradient and diffuse in the direction of the illumination. If a magnetic field B is applied in the z direction transversely to the diffusion current, the electrons and holes will be deflected in the opposite x direction, resulting in the creation of an electric field in the x direc-

tion. This will, in turn, produce J_{sc} if the two electrodes are short-circuited, or produce an open-circuit photovoltage V_{PEM} between the two electrodes, which is just sufficient to prevent the current flow.

The photovoltaic phenomenon, unlike the Dember effect, occurs even if $\mu_n = \mu_p$. Based on a simple phenomenological theory, an approximate expression for the short-circuit current per unit width of the slab in the magnetic field direction has been derived,^{69,91} and it is given by

$$J_{sc} = IB[2kT\tau q\mu_n\mu_p(\mu_n + \mu_p)]^{1/2} \quad (3-166)$$

The open-circuit photovoltage is given simply by the product of the resistance of the slab and J_{sc} , so that the PEM open-circuit photovoltage can be expressed as

$$V_{PEM} = -J_{sc}d/\sigma t \quad (3-167)$$

where I is the light intensity, expressed in quanta (photons) per second per unit area absorbed by the slab; B is the magnetic field in flux density; τ is the lifetime of the photogenerated carriers; σ is the conductivity of the slab; and d and t are, respectively, the length and the thickness of the slab. It can be seen from Equations 3-166 and 3-167 that both J_{sc} and V_{PEM} increase linearly with light intensity and magnetic field. It is interesting to note that for the PEM, J_{sc} depends on $(\tau)^{1/2}$, while for photoconduction, the current usually depends linearly on τ . It should also be noted that J_{sc} and V_{PEM} depend on surface recombination, which has been ignored in deriving these approximate expressions. However, for the validity of the expressions, the thickness t of the slab must be small compared to the carrier diffusion length, so bulk recombination can be neglected.

For the configuration shown in Figure 3-58(b), the direction of the illuminating light and that of the magnetic field are mutually perpendicular. This arrangement will develop a voltage V_x , which creates an open-circuit circulating current lying completely in the x - y plane perpendicular to the magnetic field. So, there is no net mechanical force produced due to this interaction if the field is homogeneous.

However, if the applied magnetic field is tilted in the x - z plane instead of being perpendicular to the x direction, then the B_x component will interact with the open-circuit circulating current in the slab, creating a torque. This phenomenon is called the *photomechanical effect*.²¹⁴ Measuring PEM torques instead of PEM photovoltages, no electrodes are required and thus, the effects due to electrical contacts can be eliminated. This is an important advantage when using the PEM effect for studying bulk properties of a material.

3.6.2 Contact Potential Photovoltaic Effects

To produce a photovoltage in a device, it is necessary to have mobile photogenerated electron-hole pairs and an internal electric field to separate the two types of carriers and to enable them to flow to the external contacts of the device. The contact potential at the interface between two different materials provides an efficient means for this requirement. Since the contact potential is associated with the bending of the energy bands near the interface, the contact potential's magnitude and sign depend strongly on the surface states at the interface.

Schottky Barrier Photovoltages

There are three possible photo-effects that can take place at the Schottky barrier:

Process 1: Light absorbed in the metal will raise electrons in the metal to energy levels high enough to surmount the barrier, as shown in Figure 3-59(a). The threshold for this to occur is the measure of the barrier height ϕ_B . Electrons with sufficient energy to enter the solid will make it acquire negative charges, thus creating a photovoltage across the barrier.

Process 2: Light with $h\nu \geq E_g$ can generate electron-hole pairs in the depletion region. The high electric field in the depletion region will separate the photogenerated carriers, as shown in Figure 3-59(a), resulting in a photovoltage between the metallic electrode and the bulk of the solid.

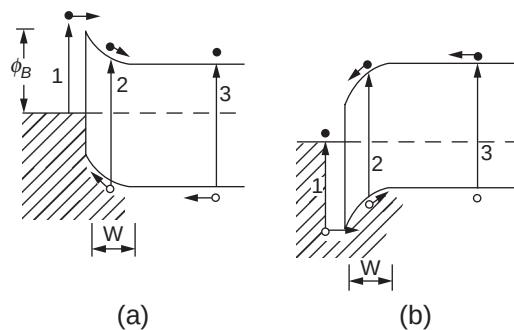


Figure 3-59 Energy band diagrams for (a) the n-type semiconductor with an electron Schottky barrier and (b) the p-type semiconductor with a hole Schottky barrier, and three carrier generation processes for photovoltaic effects.

Process 3: Light with $h\nu \geq E_g$ may also be absorbed deep in the bulk of the solid and generate electron-hole pairs there. One type of carrier (minority carriers) diffuses to the junction, as shown in Figure 3-59(a). This also contributes to the total photovoltage across the barrier.

In general, the contribution from process 1 is small because the electrons require momentum conservation (phonons) for their crossing.²¹⁶ The major contribution to the photovoltaic effects is from processes 2 and 3. Process 1 is, in fact, similar to photoemission, which was discussed in Section 3.5.1. Electrons injected from a metal into a semiconductor experience a force directed away from the contact. Of course, the probability of the occurrence of photoemission depends on the thickness of the metal film that is the illuminated electrode. If the electron attenuation length is sufficiently long compared to the thickness of metal film, then the photo-response is proportional to $(h\nu - \phi_B)^2$.²¹⁷ Processes 2 and 3 occur when $h\nu \geq E_g$. A typical curve showing the spectral dependence of the photoresponse is shown in Figure 3-60. Thus, measurements of the threshold energies at which these processes occur may be used to determine the barrier height ϕ_B and the energy gap of the semiconductor E_g .¹⁸¹ When the photon energies

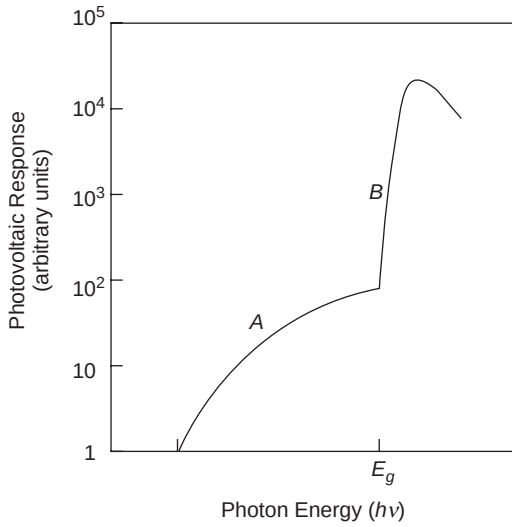


Figure 3-60 Spectral dependence of photovoltaic response. A: photoemission from the metal. B: the band-to-band transition.

are much larger than E_g , the electron-hole pairs are generated very close to the interface between the illuminated electrode and the semiconductor because of the high absorption there. The decrease of photoresponse at high photon energies is due to the strong surface recombination. For $h\nu \geq E_g$, and on the assumption that the energy band diagram of the Schottky barrier is as shown in Figure 3-59(a) without interface states, the photocurrent due to process 2 (photocarriers generated in the depletion region) for the exciting light of wavelength λ is given by²¹⁸

$$J_{dr} = qT(\lambda)I(\lambda)[1 - \exp(-\alpha w)] \quad (3-168)$$

and the photocurrent due to process 3 (photo-generated holes collected from the bulk) is given by²¹⁸

$$J_p = \frac{qI\alpha L_p}{(\alpha^2 L_p^2 - 1)} T \exp(-\alpha w) \times \left(\alpha L_p - \frac{(SL_p/D_p)\{\cosh[(d-w)/L_p] - \exp[-\alpha(d-w)]\} + \sinh[(d-w)/L_p] + \alpha L_p \exp[-\alpha(d-w)]}{(SL_p/D_p)\sinh[(d-w)/L_p] + \cosh[(d-w)/L_p]} \right) \quad (3-169)$$

where $T(\lambda)$ is the transmission of the metal film into the underlying semiconductor, $I(\lambda)$ is the incident photon flux, α is the absorption coefficient, S is the surface recombination velocity, d is the semiconductor specimen thickness, L_p is the diffusion length of holes, and w is the width of the depletion layer and is given by Equation 6.27 (see Chapter 6). The total photocurrent is the sum of J_{dr} and J_p .

Several effects that would affect the photocurrent are not included in the derivation of Equations 3-168 and 3-169: the surface states at the metal-semiconductor interface, the reflection from the illuminated electrode surface, the image potential, and the interfacial dielectric layer. To include these effects would make the analysis intractable.

If the photocurrent under the short-circuit condition is

$$J_{sc} = J_{dr} + J_p \quad (3-170)$$

then when the two electrode is not short-circuited, but connected with a resistive load with the resistance R between them, the photocurrent collected externally can be written as

$$J_{ph} = J_{sc} - J_o[\exp(qV_{ph}/kT) - 1] \quad (3-171)$$

in which the second term is the normal forward-bias current at the voltage V_{ph} for a Schottky barrier diode and J_o is the reverse bias saturation current, which is

$$J_o = AT^2 \exp(-\phi_B/kT) \quad (3-172)$$

where A is the Richardson constant, which is normally equal to $120 A \text{ cm}^{-2} \text{ K}^{-2}$, and ϕ_B is the Schottky barrier height.²¹⁹ When the photocurrent J_{ph} is flowing, a photovoltage $V_{ph} = J_{ph}R$ will be developed across the diode, producing a forward-bias current in the direction opposite to the photocurrent. Thus, the open-circuit voltage can be deduced simply by setting $J_{ph} = 0$ in Equation 3-171, which is

$$\begin{aligned}
 V_{oc} &= \frac{kT}{q} \ln \left(\frac{J_{sc}}{J_o} + 1 \right) \\
 &= \frac{kT}{q} \ln \left(\frac{J_{sc}}{J_o} \right)
 \end{aligned}
 \quad (3-173)$$

since $J_{sc} \gg J_o$.

MIS Solar Cells

If Si or any covalent bonded crystal is used as the semiconductor for solar cells, dangling bonds exist at the surface of the crystal due to the interruption of the perfect periodicity of the crystal lattice. These dangling bonds create surface states, which may overlap the states in the valence and the conduction bands, but only those within the forbidden gap play a role in the behavior of the metal–semiconductor contacts. The dangling bonds tend to capture electrons to complete the bonds. Theoretically, the number of surface states created by dangling bonds should be approximately equal to $(10^{23} \text{ cm}^{-3})^{2/3} = 10^{15} \text{ cm}^{-2}$. But the experimental data show the surface state density being of the order of 10^{12} cm^{-2} , implying that there exists mutual saturation of unsaturated bonds of neighboring atoms, and that there may be an oxide layer of about 20 Å in thickness formed on the semiconductor surface due to the preparation and exposure of the surface to the atmosphere containing oxygen before the deposition of the metallic electrode.²¹⁹ This would result in the formation of an MIS structure with a very thin insulating layer (I) between metal (M) and the semiconductor (S), as shown in Figure 3-61.

The insulating layer (oxide) screens the semiconductor surface from the metal and helps saturation of unsaturated bonds. The open-circuit photovoltage V_{oc} decreases with increasing J_o , as shown in Equation 3-173. J_o for Schottky barrier diodes is a few orders of magnitude higher than for p–n junction diodes, so V_{oc} for Schottky barrier solar cells is significantly smaller than that of p–n junction solar cells. J_o represents the majority carrier thermionic current flowing in the direction opposite to the photogenerated current. The thin insulating layer between the metal and the semiconductor also serves to reduce J_o , thus increasing the magnitude of V_{oc} .²²⁰ The thin insulating layer with a thickness of the order of 20 Å will be no

barrier for the photogenerated minority carriers. They can tunnel through a thin layer.

Photovoltaic Behavior of P–N Junctions

A simple p–n junction diode is shown in Figure 3-62. In general, a very thin semitransparent metallic film (usually gold film of 50–100 Å in thickness) is vacuum-deposited on one side of the semiconductor surface for the light illumination; the front contact, with a thick metallic grid, is then deposited. To minimize surface reflection at the metal–air interface, an antireflective film is coated on the surface, as shown in Figure 3-62(a). For such simple p–n junction diodes, the current-voltage characteristic for the abrupt junction case is given by¹⁶⁸

$$J = J_o [\exp(qV/kT) - 1] \quad (3-174)$$

where J_o is the reverse bias saturation current, which is

$$J_o = \frac{qD_n n_{po}}{L_n} + \frac{qD_p p_{no}}{L_p} \quad (3-175)$$

where D_n and D_p are, respectively, the diffusion coefficients for minority electrons and minority holes; L_n and L_p are, respectively, the diffusion lengths of minority electrons and minority holes; n_{po} and p_{no} are, respectively, the concentrations of the minority electrons in the p-type and the minority holes in the n-type semiconductors. It can be seen from Figure 3-62(b) that an electric field exists in the space charge (depletion) region even without bias. Incident photon illumination creates electron–hole pairs in the space charge region, as well as in the neutral regions of the n-type and p-type semiconductors. For the latter, however, only those electrons and holes within the diffusion length from the edge of the depletion region have a chance to diffuse to the depletion region. These carriers in the depletion region will be swept out, producing the photocurrent J_{ph} in the reverse-bias direction, as shown in Figure 3-62(c) and (d).

Similar to Schottky barrier solar cells, the photocurrent will produce a voltage drop across the load resistance R , which produces a forward bias across the p–n junction. This forward-bias voltage V_{ph} in turn produces a forward-bias current J , as given by Equation

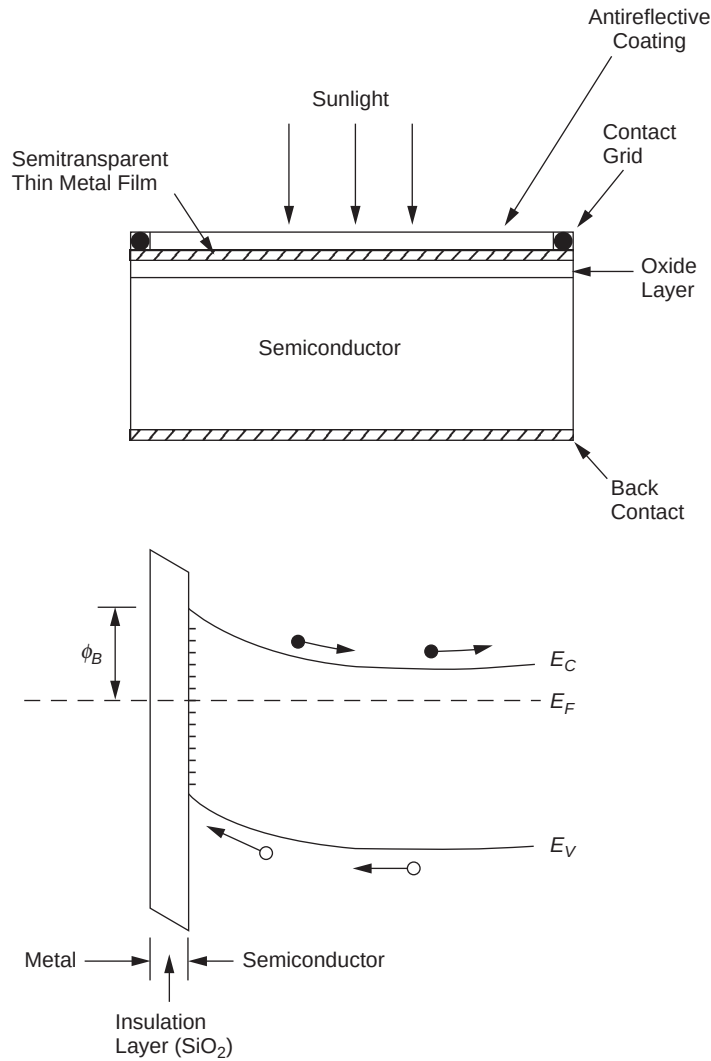


Figure 3-61 Schematic diagrams showing the basic structure of an MIS solar cell and its energy band diagram.

3-174. Thus, the photocurrent under short-circuit condition is

$$J_{sc} = J_{ph}(V_{ph} = 0) \quad (3-176)$$

With a load resistance R instead of short-circuiting, the net photocurrent becomes

$$J_{ph} = J_{sc} - J_o[\exp(qV_{ph}/kT) - 1] \quad (3-177)$$

There are two limiting cases. One is when $R = 0$, $V_{ph} = 0$, we have J_{sc} , which is the highest photocurrent that can be obtained. The other is when $R \rightarrow \infty$, then $J_{ph} = 0$. From Equation 3-177, the open-circuit photovoltage V_{oc} can be written as

$$V_{oc} = \frac{kT}{q} \ln\left(\frac{J_{sc}}{J_o} + 1\right) \quad (3-178)$$

This equation is similar to Equation 3-173, but J_o for p-n junctions is much smaller than J_o for Schottky barrier diodes.

The power delivered to the load is

$$P = J_{ph}V_{ph} = \{J_{sc} - J_o[\exp(qV_{ph}/kT) - 1]\}V_{ph} \quad (3-179)$$

By setting $dP/dV_{ph} = 0$, we can obtain the condition under which the maximum power can be delivered to the load, which is

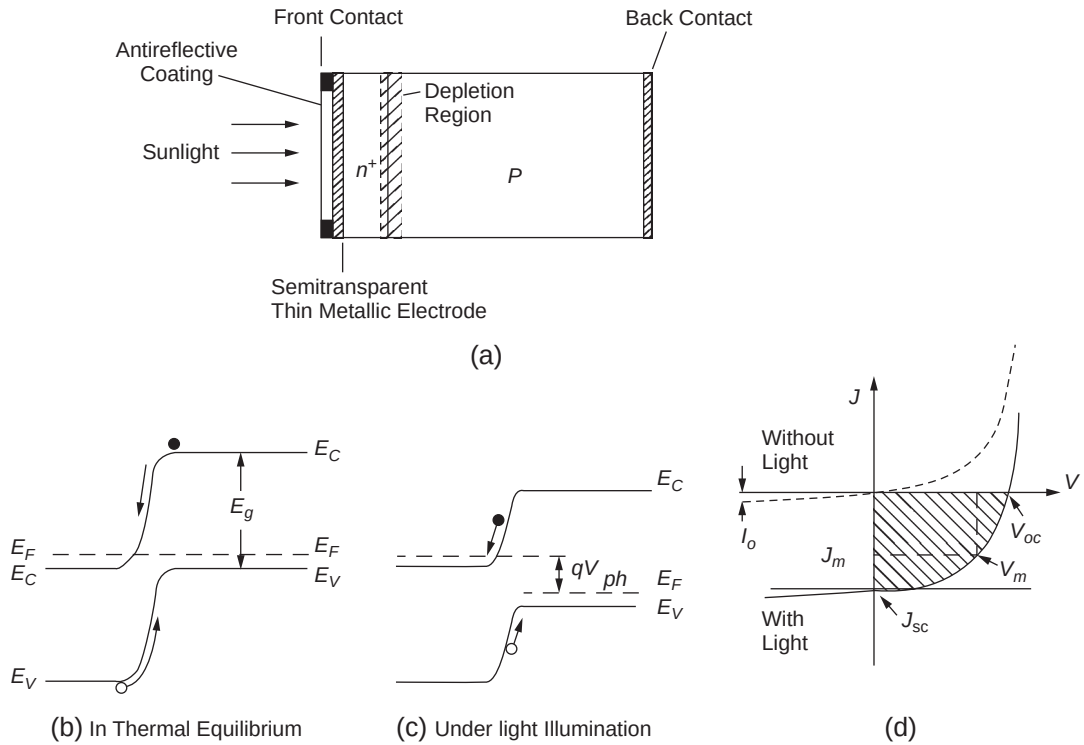


Figure 3-62 Schematic diagrams showing (a) the basic structure of an n^+ - p junction solar cell, (b) the energy band diagram in thermal equilibrium, (c) the energy band diagram under light illumination generating a photovoltage V_{ph} , and (d) the $J - V$ characteristic.

$$\left(1 + \frac{qV_m}{kT}\right) \exp\left(\frac{qV_m}{kT}\right) = 1 + \frac{J_m}{J_o} \quad (3-180)$$

where V_m and J_m are, respectively, the photovoltage and the photocurrent, at which the power delivered to the load is maximal. V_m and J_m can be determined by the trial-and-error method. The maximum power is $J_m V_m$ product, which is the area of the rectangle formed within the boundary of the $J_{ph} - V_{ph}$ curve, as shown in Figure 3-62(d). The energy conversion efficiency is given by

$$\eta = J_m V_m / P_{in} \quad (3-181)$$

where P_{in} is the incident optical power. The maximum possible photocurrent is J_{sc} , and the maximum possible photovoltage is V_{oc} . The fill factor is defined as

$$\text{Fill Factor} = \frac{J_m V_m}{J_{sc} V_{oc}} \quad (3-182)$$

This simple analysis shows the basic principle of an ideal p-n junction solar cell. For practical p-n junction solar cells, one side of the semiconductor (n- or p-side, but usually the n-side) is heavily doped to the n^+ level, so that the Fermi level of the n^+ side is inside the conduction band because of its degeneracy. This would make it easy for photons with the energy of the order of E_g to reach the depletion region. The heavy doping of the n^+ side also makes it easy to produce an ohmic contact with the deposited thin semitransparent metallic electrode for the admission of the incident light. The energy conversion efficiency depends on the wavelength of the incident photons. A photon with energy greater than E_g will contribute to the solar cell output power, but a fraction of the energy (typically $h\nu - E_g$) will be dissipated as heat, while a photon with energy less than E_g will not contribute much to the solar cell output power.

The following list points out the limitations of photovoltaic solar energy converters, along with some possible improvements in their efficiency:

Reflection losses: The surface of the n^+ side semiconductor should be texturized to produce pyramids (roughnesses) before metallization. After the deposition of the thin semitransparent electrode, a thin antireflective coating is then deposited to provide further antireflection. A good antireflective coating should have a refractive index of about two and should be as transparent as possible in the solar energy spectrum. The optimal thickness is of the order of $800 \text{ \AA}$. Typical materials for antireflective coating are TiO_2 , Ta_2O_3 , and Si_3N_4 . Several methods can be used to deposit the antireflective coating, such as vacuum evaporation, spin-on or spray-on techniques (spinning or spraying a liquid containing TiO_2 or Ta_2O_3 on the surface), or screen-printing techniques.²²¹

Collection losses: Only those electron-hole pairs generated within the diffusion length from the edge of the depletion region have a chance to contribute to electrical energy. Thus, the larger the diffusion length, the better the conversion efficiency. But both the diffusion coefficient and the lifetime of the minority carriers decrease with increasing concentrations of the doping impurities or unwanted defects, so the diffusion length is limited and cannot be increased much.

Voltage factor: The open-circuit voltage V_{oc} is smaller than E_g/q simply because $E_{Fn} - E_{Fp} < E_g$. Only when the injection level becomes extremely high may V_{oc} approach the value of E_g/q , but such a high injection level can never be reached from sunlight.

Effect of series resistance: For p-n junctions, series resistance is the sum of the contact resistance and the bulk resistance of the semiconductor. The voltage drop across the series resistance reduces the effective potential across the junction, increases the internal power dissipation, and decreases the fill factor. In general, the contact resistance can be reduced to a negligible value, but the bulk resist-

ance depends on the doping level. A heavily doped semiconductor reduces the minority carrier lifetime and diffusion length. So, the doping level and the junction depth must be adjusted to arrive at the optimal compromise. The effect of series resistance R_s on the $J_{ph} - V_{ph}$ characteristics is shown in Figure 3-63. The experimental result of the GaAs p-n junction is from Janny, Loferski, and Rappaport.²²²

Thus, taking R_s into account, Equation 3-177 must be written as

$$J_{ph} = J_{sc} - J_o(\exp[q(V_{ph} - J_{ph}R_s)/kT] - 1) \quad (3-183)$$

and Equation 3-178 as

$$V_{oc} = \frac{kT}{q} \ln\left(\frac{J_{sc}}{J_o} + 1\right) \quad (3-184)$$

which is the same as Equation 3-178, implying that series resistance does not affect the open-circuit voltage.

Incomplete absorption: A small absorption coefficient means a deep penetration of the incident photons, implying that most of the photons eventually transmit through the material because of insufficient interaction between the photons and the bound electrons inside the material. Thus, the absorption properties of photovoltaic material determine, to a great extent, the amount of the incident radi-

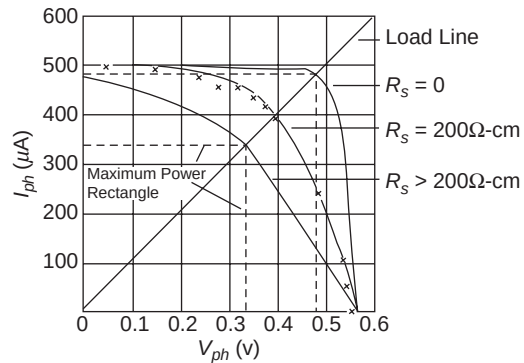


Figure 3-63 The effect of the series resistance R_s on the photocurrent (I_{ph}) – photovoltage (V_{ph}) characteristics. Solid lines are theoretical curves; xxx represent the experimental data from a GaAs photovoltaic device.

ation that can be converted into electrical energy. The absorption coefficient of a material depends on the wavelength of the incident radiation. The absorption of light in a semiconductor is determined by several mechanisms, of which the two most important are fundamental absorption and free carrier absorption. *Fundamental absorption* refers to the absorption by which a photon with energy equal to or larger than E_g will generate one electron-hole pair with the excess energy, $\Delta E = h\nu - E_g$, if any, being dissipated as thermal energy and thus lost for photovoltaic conversion. *Free carrier absorption* refers to the absorption by free carriers to raise their energy level in the conduction or the valence band without contribution to energy conversion. Total absorption coefficient $\alpha(\lambda)$ is defined as the reciprocal of the distance for the energy to fall by a factor of e . Thus, $\alpha(\lambda)$ is

$$\alpha(\lambda) = \alpha_f + \alpha_{fc} \quad (3-185)$$

where α_f and α_{fc} are, respectively, the fundamental and the free carrier absorption coefficients. The absorption coefficient $\alpha(\lambda)$ depends strongly on whether the semiconductor is crystalline or amorphous, and in the case of a crystalline semiconductor, whether it is a *direct* or *indirect* gap semiconductor. In a crystalline semiconductor, both energy and momentum must be conserved when an electron makes a transition from the valence band to the conduction band. In a direct gap semiconductor, such as GaAs, optical transition does not need a change in momentum, so it has a large absorption coefficient. In an indirect gap semiconductor, such as Si, an optical transition for an electron with the energy even equal to E_g is only possible if, at the same time, a suitable phonon is available to help the electron to satisfy the momentum conservation for the transition. In this case, the probability of light absorption is much smaller; hence, Si has a much lower absorption coefficient.

Possible ways to improve conversion efficiency: Sunlight consists of photons having different energies and associated wave-

lengths. For p-n junctions with a constant energy gap, the maximum conversion efficiency for Si crystal p-n junction solar cells is 21.6%, and that for GaAs is 23%. Several methods may lead to the improvement of conversion efficiency:

- We may use a proper doping impurity grading to increase the absorption of those photons with energies lower than E_g .
- We may use multitransition materials with two or more trapping levels in the energy band gap, as shown in Figure 3-64.
- We may use suitable heterojunction structures, including nSi-pGaAs, pGa_{1-x}Al_xAs-pGaAs-nGaAs, etc.^{223,224} A typical heterojunction solar cell is shown schematically in Figure 3-65. The basic principle is that the two junctions can occur in materials of different energy gaps, so that each can make efficient use of a different portion of the solar spectrum. Usually, the large energy gap material is on top and absorbs the short wavelength radiation first. The longer wavelength radiation is captured by the smaller energy gap material.

Reduction of cost per unit power generated by solar cells:

Thin film solar cells have been developed to reduce fabrication cost and increase the exposure area of the solar cell to incident sunlight. The most promising material for high conversion efficiency and low-cost thin film solar cells is cadmium telluride (CdTe). The cell consists of a heterojunction between II-IV compounds, such as between p-type copper telluride (Cu₂Te) and n-type CdTe. The behavior is very similar to a heterojunction between p-type CdTe and n-type CdS solar cell. The conversion

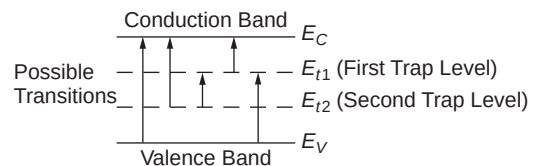


Figure 3-64 Schematic illustration of a material having multitransition levels in the energy band gap for increasing the light absorption in different wavelengths.

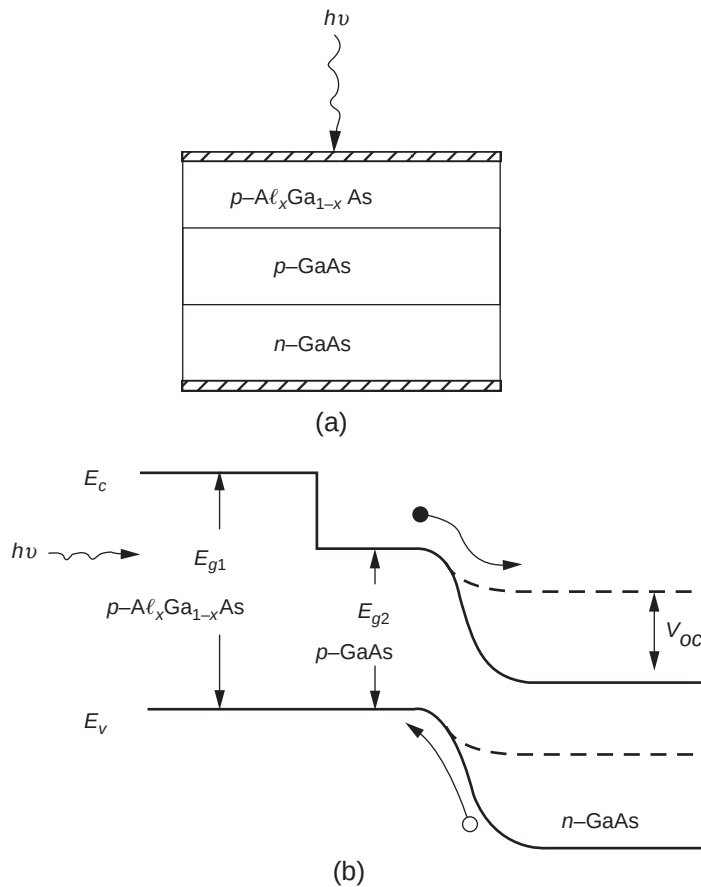


Figure 3-65 Schematic diagrams illustrating (a) the basic structure of $p\text{Al}_x\text{Ga}_{1-x}\text{As-pGaAs-nGaAs}$ heterojunction solar cell, and (b) the energy band diagram for this solar cell with $E_{g1} > E_{g2}$.

efficiency of such II–VI compound thin film solar cells has reached about 17%.²²⁵ Most thin films are polycrystalline. For further cost reduction, amorphous silicon-based solar cells may provide the solution. This will be discussed in the next section.

PIN Structures for Amorphous Si-Based Photovoltaic Devices

It is important to understand the basic concept of why amorphous silicon ($a\text{-Si}$) has the potential for low-cost photovoltaic devices and why a p-i-n structure must be used for such devices.

Amorphous materials have no long-range order; the concept of crystal momentum is not

applicable, and they are not conserved in optical transition as they are in crystals. The lack of long-range order results in two important features:

- The absorption coefficient in the solar light energy spectrum for $a\text{-Si}$ is about an order of magnitude higher than that for crystalline Si at the maximum solar light wavelength near 5000 Å.
- A high concentration of defects exists in the energy gap.

In crystalline Si photovoltaic devices, the p-n junction structure is generally used, in which the space charge (depletion) region width is generally between 0.1% and 1.0% of the device

thickness. Almost all of the carriers are photo-generated in the neutral (field-free) regions outside the space charge region; therefore, the minority carrier lifetime must be longer than the time required for them to travel to the space charge region edge, so that they can be collected at the terminals.

In amorphous Si photovoltaic devices, the carrier collection process is quite different. Amorphous Si has a high absorption coefficient but a low mobility, implying a low diffusion length. In order to eliminate the dangling bonds for obtaining good electronic and optical properties, *a*-Si must be incorporated with hydrogen and become *a*-Si:H. Generally, the optical gap and the diffusion length decrease with increasing impurity content. This implies that the p-n junction structure is not suitable for *a*-Si:H-based photovoltaic devices. It is necessary to adopt a p-i-n structure to provide an undoped i-layer in which an internal field is created so that the carriers photogenerated in the i-layer are strongly attracted toward the regions in which they are the majority carriers. The thicknesses of the three layers are typically $0.5\text{ }\mu\text{m}$ for the i-layer and $0.01\text{ }\mu\text{m}$ for both the n- and the p-layers. The i-layer normally refers to an insulating layer, but here, it refers to an intrinsic or undoped *a*-Si:H layer. Because of the large internal field, photogenerated carriers can be collected even though their lifetime is small, since the collection mechanism does not depend strongly on carrier lifetime. It is a drift transport throughout almost the entire device, thus yielding a good collection efficiency. In *a*-Si:H p-i-n structures, diffusion outside the field region is less important.

Figure 3-66 illustrates the basic structure of an *a*-Si:H p-i-n homojunction solar cell and the energy band diagrams at thermal equilibrium and under solar radiation. In the crystalline Si p-n homojunction, the diffusion current is larger than the drift current, whereas in the *a*-Si:H p-i-n homojunction, the drift current is larger than the diffusion current. This may be the basic difference between these two types of solar cells. In *a*-Si:H p-i-n solar cells, the carrier transport is characterized by the drift length, which is $\mu F\tau$, where F is the internal

field in the i-layer and μ and τ are the carrier mobility and lifetime, respectively. For good collection efficiency, the drift length must be larger than the i-layer thickness.

However, the high absorption in the front p-type *a*-Si:H or n-type *a*-Si:H thin layer reduces the light penetration into the i-layer. To overcome this disadvantage, a heterojunction structure has been developed by employing a wide bandgap, boron-doped *a*-SiC:H layer as a front window, as shown in Figure 3-67. It is important to increase the light input to the i-layer and also to increase the absorption of the long wavelength portion of the solar light spectrum. The latter can be achieved by using a narrow bandgap material such as *a*-SiGe:H behind the i-layer.²²⁷ The advantage of using *a*-SiC:H is that the optical gap of this material can be adjusted from 1.8 eV to 2.8 eV simply by adjusting the value of x in the gas mixture $[\text{SiH}_{4(1-x)} - \text{CH}_{4x}]$ when fabricated by a high-frequency glow discharge technique in a plasma reaction chamber, the optical gap increasing with increasing x . Unlike crystalline materials, where there is no lattice-matching problem, *a*-Si:H can form junctions with any material.

By carefully reducing the unwanted impurities using an ultrahigh vacuum (UHV) chamber of 10^{-9} torr, Nakano et al.^{228,229} have achieved a conversion efficiency of 11.7% for a cell of 1 cm^2 and 9.72% for a cell module of 100 cm^2 size of a p-i-n heterojunction structure, as shown in Figure 3-67. The i-layer usually consists of a large amount of unwanted impurities (10^{19} to 10^{21} cm^{-3}) such as O, N, C. These impurities have a strong influence on film properties, space charge density, and light-induced degradation. The concentrations of such impurities depend on the method used for film fabrication. It has been found that there are two types of light-induced degradation—one related to impurities and the other to structural defects—and that the reduction of the oxygen and SiH_2 contents will result in improvement of conversion efficiency.^{228,229}

For more details about the p-n junction and amorphous p-i-n junction photovoltaic devices and solar cells, see references 168, 218, 221, 226, and 230–236.

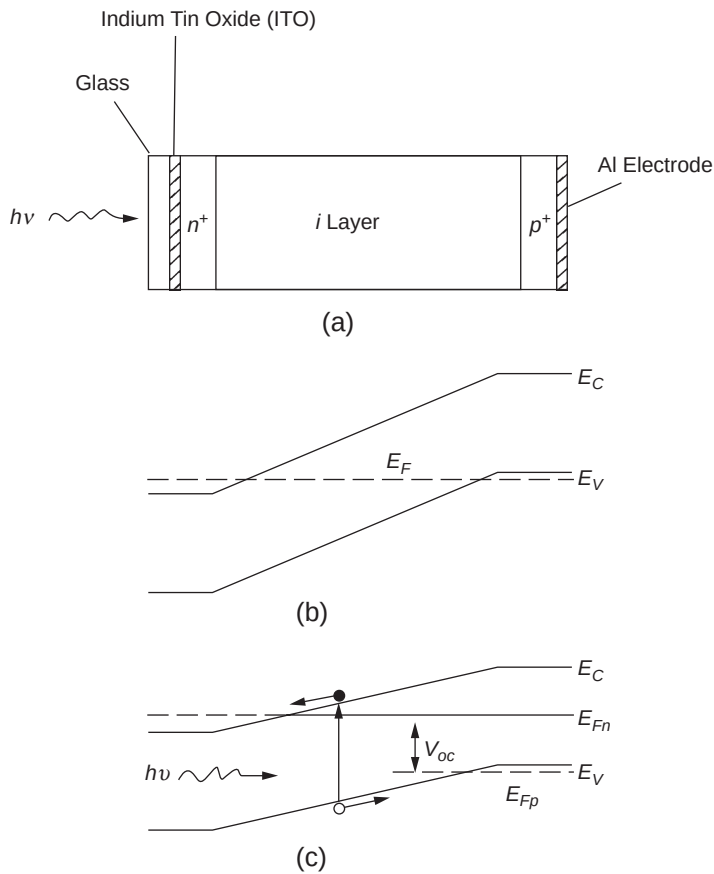


Figure 3-66 Schematic diagrams showing (a) the basic structure of an amorphous silicon a -Si:H p-i-n solar cell, (b) the energy band diagram at thermal equilibrium, and (c) the energy band diagram under light illumination.

3.6.3 Photosynthesis Photovoltaic Effects

In general, photosynthesis refers to a process utilizing light (usually sunlight) to convert carbon dioxide and water into carbohydrates and oxygen. This usually occurs in green plants. In living systems, chlorophyll and the lipid membrane are liable to change, but they are continuously renewed to maintain a dynamic stability. For photovoltaic devices, we may use a similar photosynthesis process, but we must use a more stable sensitizer to react with light in a manner similar to chlorophyll or the lipid membrane. In inorganic materials, which are generally more stable than organic materials, it has been found that transition metal complexes are very stable and have good

light absorption properties. Titanium dioxide (TiO_2) is one of these inorganic materials. It is a semiconductor with a large band gap, but it does not absorb visible light. So, a thin TiO_2 film consisting of TiO_2 particles of a few nanometers size, coated with a monolayer of charge transfer dye, will form a sensitizer. Usually, the film is about $5\mu\text{m}$ thick, deposited on a transparent conductive layer of fluorine-doped thin oxide, as shown in Figure 3-68(a). In the case of n-type TiO_2 film, electrical current is produced when light absorbed by the dye molecules generates electrons, which are then injected into the conduction band of the semiconductor TiO_2 . The minority carriers (holes in the present case) do not participate in the photovoltaic conversion. However, to complete the

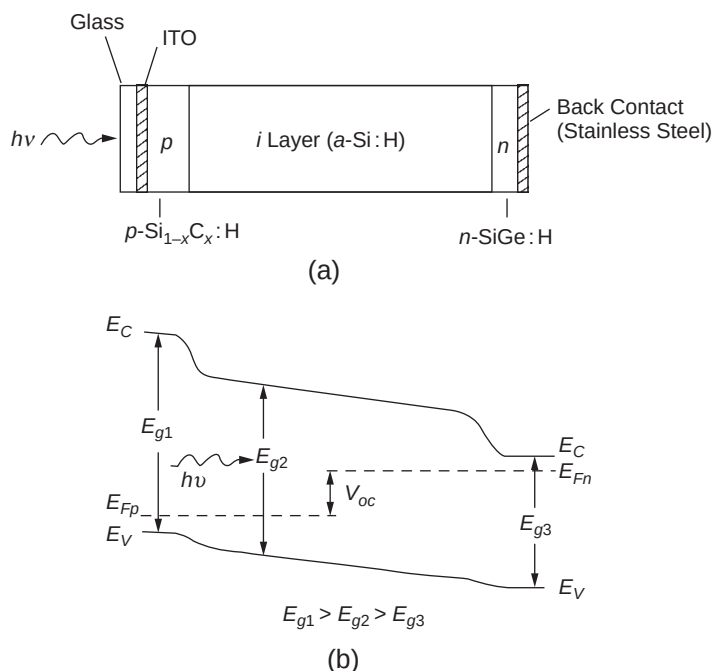
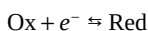


Figure 3-67 Schematic diagrams showing (a) the basic structure of an amorphous $p\text{Si}_{1-x}\text{C}_x\text{:H}$ – $i\text{Si:H}$ – $n\text{SiGe:H}$ heterojunction p–i–n solar cell and (b) the energy band diagram for this solar cell with $E_{g1} > E_{g2} > E_{g3}$ under light illumination.

current flow circuit, the dye must be regenerated. This is done by the transfer of the electrons from redox species in the electrolyte to the dye. The counter electrode will collect the electrons from the semiconductor and then return them to the electrolyte.

In the electrolyte–dye region, the electrolyte is reducing (*Red*) when there is electron transfer from the electrolyte to the dye, because *reducing* means an increase in negative charge or a gain of electrons in the dye. Similarly, the electrolyte is oxidizing (*Ox*) when there is electron transfer from the dye to oxidized species in the electrolyte, because *oxidation* means a decrease in negative charge or a loss of electrons in the dye. A redox reaction may be simply expressed by



The Red species is equivalent to an occupied electron state in the valence band, whereas the Ox species is equivalent to an empty electron state in the conduction band. The Red and Ox species are separated by an energy gap, thus

energy is required for an electron to transfer from the former to the latter. This energy gap is analogous to the energy band gap in an intrinsic semiconductor.

Figure 3-68(b) illustrates schematically how electrons are photogenerated and then injected from the dye to the TiO_2 semiconductor under sunlight radiation. The electrons are transported from the semiconductor to the counter electrode and then return to the electrolyte, which has transferred electrons to recombine with holes photogenerated in the dye, thus completing the electron flow circuit. This device operates entirely on majority carriers (electrons in the present case), in contrast to conventional p–n junctions, whose photovoltaic performance depends almost entirely on minority carriers and their diffusion lengths.

Between the counter electrode and the sensitizer is an electrolyte containing a redox couple, that is, iodine and iodide. The electrons reduce iodine to iodide ions, which diffuse from the counter electrode to the sensitizer, where they regenerate the cations in the dye;

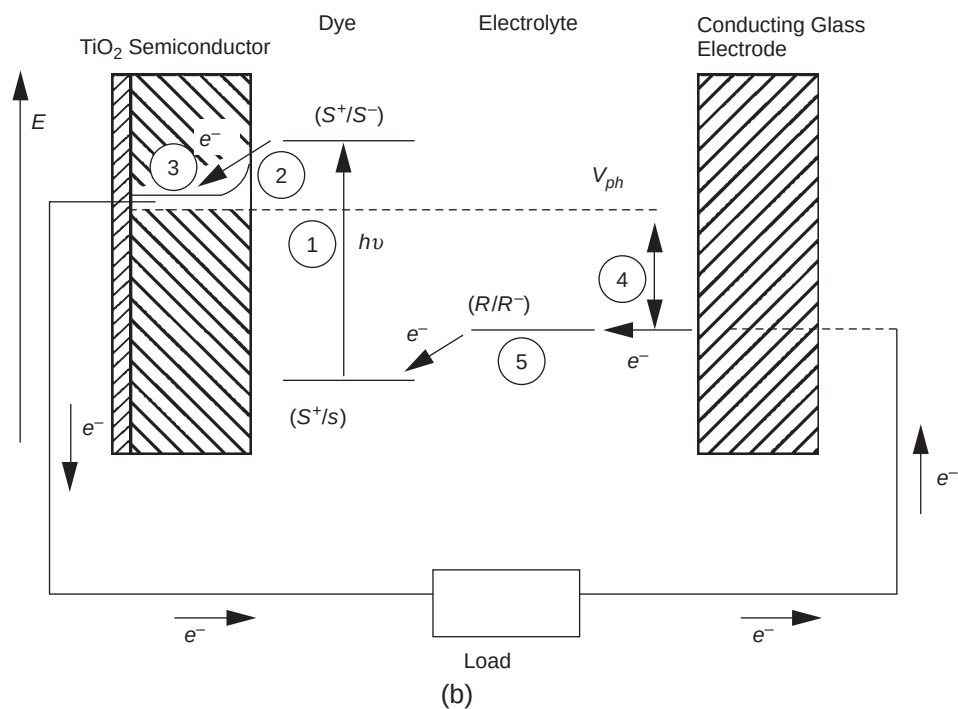
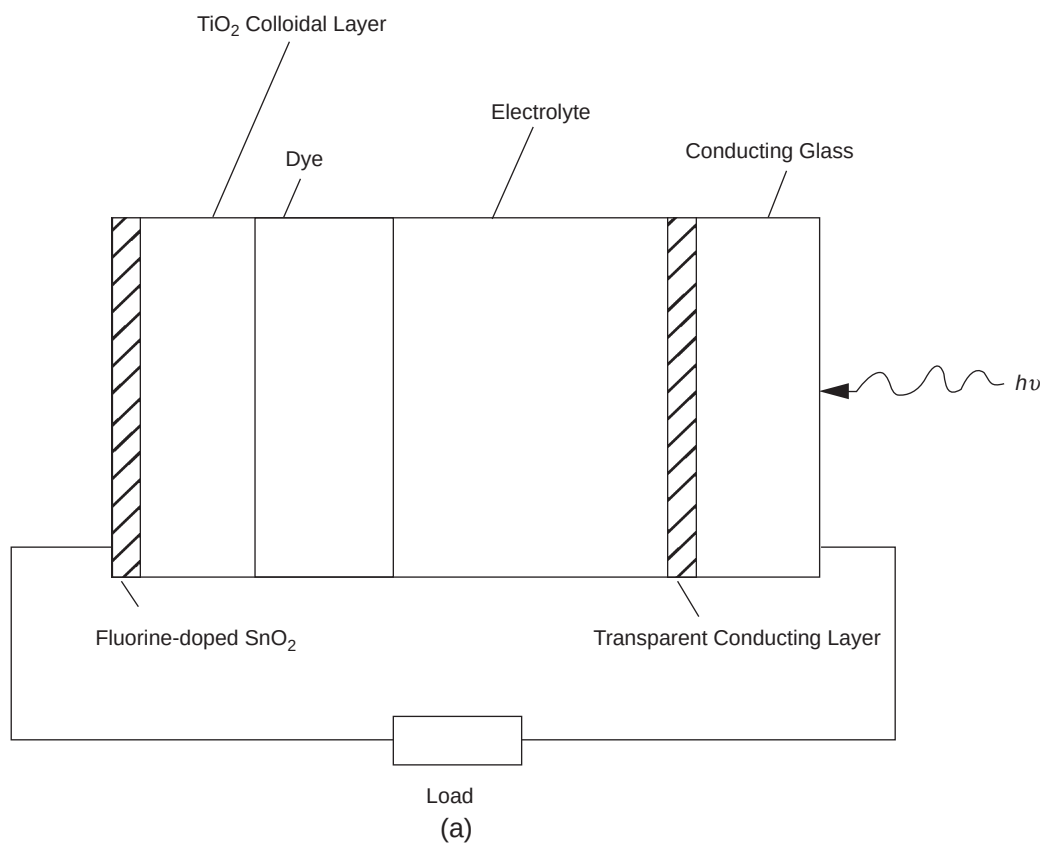


Figure 3-68 Schematic diagrams showing (a) the basic structure of a photosynthesis photovoltaic device and (b) the photosynthetic process for producing photovoltaic voltages.

simultaneously, the iodide is oxidized back to iodine. The cycling of the redox species enables the conversion of light into electrical current. The photovoltage V_{ph} generated by this process is the difference between the Fermi level of the TiO_2 under light illumination and the Nernst potential of the redox couple in the electrolyte.

At this time, the conversion efficiency of this type of photovoltaic solar cell has reached about 10%, and the cost is expected to be much lower than that of conventional silicon solar cells. However, more research and development are required to make this new idea feasible for practical applications. For more details about this type of photovoltaic device, see references 237–242.

3.6.4 Anomalous Photovoltaic Effects

Some semiconductors in thin film form exhibit a high photovoltage when exposed to intense light, and in some cases the photovoltages may reach a value as high as 5000 V, for example, in III–V compound semiconductor films.²⁴³ Such an anomalous photovoltaic phenomenon has been observed in many semiconductor films, such as germanium,^{244,245} silicon,²⁴⁶ and many compound semiconductors.⁹⁹

The semiconductor films exhibiting such anomalous behavior are usually fabricated using an oblique vapor deposition technique. This means that when a film is grown, the insulating substrate must be inclined with respect to the plane normal to the direction of evaporating vapor for deposition. In general, the angle of inclination between the direction of evaporation and the direction normal to the substrate plane must be in the range of 30 to 60 degrees. This phenomenon is sensitive to film preparation processes. Apart from the inclination of evaporation direction, this phenomenon depends on the substrate temperature during oblique deposition, which should be in the range of 50–100°C, and also on the pressure and the composition of the gases in the deposition chamber in which the film is fabricated. It may also depend on the substrate material and its surface treatment, film thickness, adsorption of oxygen, etc.⁹⁹ Open-circuit photovoltages as

functions of light intensity for some semiconductors are shown in Figure 3-69. The experimental results are from Adirovich, Rubinov, and Yuabov.²⁴⁷ Photovoltage generally increases with decreasing temperature.

Several models have been put forward to explain such anomalous phenomena. However, none of them can account for all observed phenomena. It is generally believed that the large photovoltage along the length of the film is due to the addition of photovoltages generated in microelements connected in a series arrangement. Such microelements could be micro p–n junctions, the Demer effect in microcrystals separated by grain boundaries, or both. It is also believed that surface recombination, which may lead to anomalous photogenerated carrier distribution, is responsible for the photovoltage polarity and the spectral sensitivity of the films.^{248,249} Obviously, anomalous photovoltaic effects are interesting and may be important for device applications. However, further work, both experimental and theoretical, is needed to clarify the ambiguities.

Obliquely deposited semiconductor films generally exhibit the same polarity of photovoltages regardless of the angle of the incident

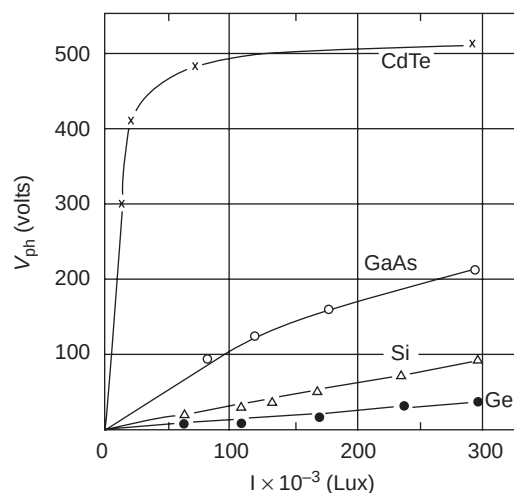


Figure 3-69 The photovoltage of V_{ph} as a function of incident light intensity I for various semiconductor thin films fabricated with an oblique vapor deposition technique.

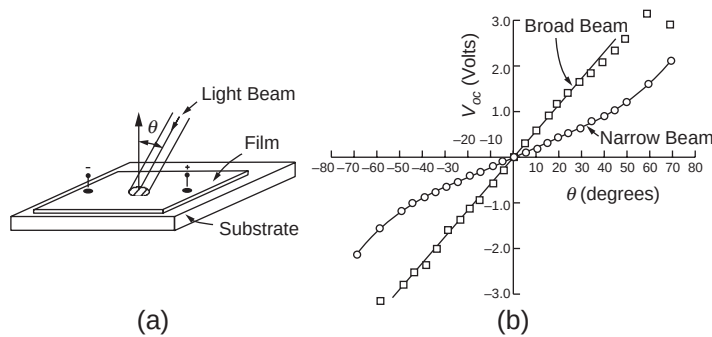


Figure 3-70 (a) The experimental arrangement for the observation of the anomalous photoangular effect and (b) the variation of the photovoltaic open-circuit voltage (V_{oc}) with the angle (θ) of the incident light beam.

light direction with respect to the film, even when the film is illuminated through its substrate. However, there is another anomalous photovoltaic effect. In this case, the films are relatively thicker ($\geq 2\mu\text{m}$) and polycrystalline. The film is usually deposited by a vapor deposition technique, with the vapor traveling across a small gap between the vapor source and the substrate. The deposition is normal to the substrate, so that the structure of such films should be quite different from obliquely deposited films. When such a film is illuminated normally with a light beam, no photovoltage appears, but as the light beam is tilted with respect to the normal direction, as shown in Figure 3-70(a), a photovoltage will be generated and increases with increasing tilting angle, as shown in Figure 3-70(b). The polarity of the photovoltage along the film depends on the direction of the light beam. The photovoltage appears only when the light beam is illuminating the film surface directly. No photovoltage appears if the film is illuminated indirectly through the substrate.²⁵⁰ This angularly anomalous photovoltaic effect may be associated with the existence of an array of angular ridges on the film surface, and the development of the photovoltage due to nonuniform concentration of photogenerated carriers in the ridges. This phenomenon may be similar to the Dember effect occurring in an angularly shaped crystal, observed by Dember in 1932.²⁰⁷ A more detailed account of anomalous photovoltaic effects can be found in reference 99.

References

1. H.W. Leverenz, *An Introduction to Luminescence of Solids*, (Dover, New York, 1968).
2. R.A. Wheadon, *Principles of Light and Optics*, (Longmans, London, 1980).
3. R.W. Ditchburn, *Light*, Vols. I and II, (Academic Press, New York, 1976).
4. M. Born and E. Wolf, *Principles of Optics*, 5th ed. (Pergamon, Oxford, 1975).
5. J. Wilson and J.F.B. Hawkes, *Optoelectronics*, (Prentice Hall, New York, 1989).
6. A.A. Michelson, *Amer. J. Science*, 22, 120 (1881); *Phil. Mag.*, 13, 256 (1882) and 31, 338 (1891), and 34, 280 (1892).
7. Lord Rayleigh, *Phil Mag.*, 34, 407 (1892).
8. W.E. Williams, *Applications of Interferometry*, (Methuen, London, 1957).
9. H. Fizeau, *Ann. Chim. Phys.*, 66, 429 (1862).
10. F.H. Rolt, *Engineering*, 144, 162 (1937).
11. C.F. Meyer, *The Diffraction of Light, X-rays, and Material Particles*, (University of Chicago, Chicago, 1934).
12. G.B. Airy, *Phil. Mag.*, 2, 20 (1833); also *Ann. Phys. und Chem.*, 41, 512 (1837).
13. M. Spencer, *Fundamentals of Light Microscopy*, (Cambridge University Press, Cambridge, 1982).
14. E. Malus, *Optique Diotrique*, J. Ecole Polytechn., 7, pp. 1-44, 84-129 (1805).
15. D. Brewster, *Phil. Trans.*, 125 (1815).
16. A.R. Von Hippel, *Dielectrics and Waves*, (Wiley, New York, 1954).
17. O.S. Heavens, *Optical Properties of Thin Films*, (Dover, New York, 1965).
18. F. Goos and H. Hanchen, *Ann. Phys.*, 6, 1 (1947).

19. F. Pockels, *Lehrbuch de Kristalloptik*, (Taubner, Leipzig, 1906); also Ann. d. Physik, 87, 158, (1889).
20. F.A. Jenkins and H.E. White, *Fundamentals of Optics*, (McGraw-Hill, New York, 1976), Chapter 32.
21. A. J. Moulson and J.M. Herbert, *Electroceramics*, (Chapman and Hall, London, 1990).
22. A. Yariv and P. Yeh, *Optical Waves in Crystals*, (Wiley, New York, 1984).
23. P. Lorrain and D. Corson, *Electromagnetic Fields and Waves*, 3rd Edition, (Freeman, New York, 1988), Chapter 9.
24. J. Kerr, Phil. Mag., 4 (50), 337 (1875).
25. R. Coelho, *Physics of Dielectrics*, (Elsevier, Amsterdam, 1979).
26. C.J.F. Bottcher and P. Bordewijk, *Theory of Electric Polarization: Dielectrics in Time Dependent Fields*, (Elsevier, Amsterdam, 1978).
27. C.G. Le Fevre and R.J. Le Fevre, "The Kerr Effect" in *Techniques of Chemistry*, Vol. 1, edited by A. Weissberger and B.W. Rossiter (Wiley, New York, 1972), p. 399.
28. S. Kielich, "Electro-Optic Properties of Dielectrics" in *Dielectric and Related Molecular Processes*, Vol. 1 (The Chem. Soc., London, 1972), p. 278.
29. M. Zahn, IEEE Trans. Dielectrics and Elect. Insul., DEI-6, 627 (1998).
30. A.D. Buckingham, Proc. Phys. Soc., B69, 344 (1956).
31. G. Mayer and F. Gires, Compt. Rend., B258, 2039 (1964).
32. C.C. Wang, Phys. Rev., 152, 149 (1966).
33. G. Haertling, Am. Ceram. Soc. Bull., 43, 875 (1964).
34. C. Land and P. Thatcher, Proc. IEEE, 57, 751 (1969).
35. P. Thatcher and C. Land, IEEE Trans. Electron. Devices, ED-16, 515 (1969).
36. R.C. Buchanan (Ed.), *Ceramic Materials for Electronics*, (Marcel Dekker, New York, 1986).
37. L.M. Levinson (Ed.), *Electronic Ceramics*, (Marcel Dekker, New York, 1988).
38. J.M. Herbert, *Ferroelectric Transducers and Sensors*, (Gordon and Breach London, 1982).
39. G. Haertling, Am. Ceram. Soc. Bull., 49, 564 (1970).
40. G. Haertling and C. Land, Ferroelectrics, 3, 269 (1972).
41. T.S. Narasimhamurty, *Photoelectric and Electro-Optic Properties of Crystals*, (Plenum, New York, 1981).
42. M.E. Lines and A.M. Glass, *Principles and Applications of Ferroelectrics and Related Materials*, (Clarendon, Oxford, 1977).
43. J.C. Purfoot and G.W. Taylor, *Polar Dielectrics and Applications*, (Macmillan, London, 1979).
44. A. Ashkin, G.D. Boyd, J.M. Dziedzic, R.G. Smith, A.A. Ballman, and K. Nassau, Appl. Phys. Lett., 9, 72 (1966).
45. F.S. Chen, J. Appl. Phys., 38, 3418 (1967).
46. F.S. Chen, J. Appl. Phys., 40, 3389 (1969).
47. F.S. Chen, J.T. LaMcChia, and D.B. Fraser, Appl. Phys. Lett., 13, 723 (1968).
48. A.M. Glass, D. von der Linde, and T.J. Negran, Appl. Phys. Lett., 25, 233 (1974).
49. P. Gunter, Physics Report, 93, (No. 4), 199 (1982).
50. P.S. Brody, U. Efron, J. Feinberg, A.M. Glass, R.W. Hellwarth, R.R. Neurgaonkar, G. Rakuljic, G.C. Valley, and C. Woods, "Photorefractive and Liquid Crystal Materials" in "Research on Non-Linear Optical Materials: An Assessment," Applied Optics, 26, 211 (1987).
51. E.H. Turner, Appl. Phys. Lett., 8, 303 (1966).
52. S.H. Wemple, M. DiDonenico, Jr., and I. Connlibel, Appl. Phys. Lett., 12, 209 (1968).
53. G. Chanussot and A.M. Glass, Phys. Lett., 50A, 405 (1976).
54. G. Chanussot, Ferroelectrics, 20, 37 (1978).
55. H. Kogelnik, Bell Syst. Tech. J., 48, 2909 (1969).
56. N.V. Kukhtarev, V.B. Markov, S.G. Odulov, M.S. Soskin, and V.L. Vinetskii, Ferroelectrics, 22, 949 and 961 (1979).
57. M.D. Ewlanck, R.R. Neurgaonkar, and W.K. Carg, J. Appl. Phys., 62, 374 (1987).
58. J. Rodriguez, A. Sich-n-akoun, and G. Salano, Appl. Optics, 26, 1732 (1987).
59. A.M. Glass, Optical Engineering, 17, 470 (1978).
60. D.I. Staebler and J.J. Amodei, J. Appl. Phys., 43, 1042 (1972).
61. J.P. Huignard, J.P. Herriau, and T. Valentin, Appl. Optics, 16, 2796 (1977).
62. J.P. Huignard and J.P. Herriau, Appl. Optics, 16, 180 (1977).
63. J.P. Huignard, J.P. Harriau, and F. Micheron, Appl. Phys. Lett., 26, 256 (1978).
64. J.P. Huignard, J.P. Herriau, and F. Micheron, Ferroelectrics, 11, 393 (1976).
65. V. Harkov, S. Odulov, and M. Soskin, in *Optics and Laser Technology*, (April 1979), p. 95.
66. B. Lax and S. Zwardling, Progress in Semiconductors, 5, 251 (1966).

67. L.M. Roth, *Phys. Rev.*, **133A**, 542 (1964).
68. W. Voigt, *Phys. Z.*, **8**, 612 and 815 (1905).
69. T.S. Moss, G.J. Burrell, and B. Ellis, *Semiconductor Opto-Electronics*, (Wiley, New York, 1973).
70. A. Yariv, *Quantum Electronics*, (Wiley, New York, 1975).
71. C.L. Chen, *Elements of Opto-Electronics and Fiber Optics*, (Erwin, Chicago, 1996).
72. R. Adler, *IEEE Spectrum*, **4**, 42 (1967).
73. I.C. Chang, *IEEE Trans. Sonic and Ultrasonics*, **SU-23**, 2 (1976).
74. A. Korpel, *Proc. IEEE*, **69**, 48 (1981).
75. G. Herzberg, *Atomic Spectra and Atomic Structure*, (Dover, New York, 1944).
76. G.N. Lewis, *Valence and the Structure of Atoms and Molecules*, (Dover, New York, 1966).
77. A.E. Ruark and H.C. Urey, *Atoms, Molecules and Quanta*, Vols. I and II, (Dover, New York, 1964).
78. A.H. Compton, *X-rays and Electrons*, (Van Nostrand, New York, 1926).
79. A.C.G. Michell and M.W. Zemansky, *Resonance Radiation and Excited Atoms*, (Cambridge Univ. Press, Cambridge, 1971).
80. E.U. Condon, *The Franck-Condon Principle and Related Topics*, *Am. J. Phys.*, **15**, 365 (1947).
81. J.B. Birks, *Photo-Physics of Aromatic Molecules*, (Wiley, New York, 1970).
82. C.H. Gooch, *Injection Electroluminescent Devices*, (Wiley, New York, 1973).
83. P.T. Landsberg, *Phys. Stat. Sol.*, **41**, 457 (1970).
84. C. Benoit-a-la-Guillauma and J. Cernogora, *J. Phys. Chem. Solids*, **24**, 383 (1963).
85. V. Kryukova, *Soviet Phys. Solid State*, **7**, 2060 (1966).
86. M. Lax, *J. Phys. Chem. Solids*, **8**, 66 (1959).
87. M. Lax, *Phys. Rev.*, **119**, 1502 (1960).
88. J.S. Blackemore, *Semiconductor Statistics*, (Pergamon, Oxford, 1962).
89. K.C. Kao, *J. Appl. Phys.*, **55**, 752 (1984).
90. D.L. Greenaway and G. Harbeke, *Optical Properties and Band Structure of Semiconductors*, (Pergamon, Oxford, 1968).
91. T.S. Moss, *Optical Properties of Semiconductors*, (Butterworths, London, 1959).
92. R.A. Smith, *Semiconductors*, (Cambridge Univ. Press, Cambridge, 1978).
93. P.T. Landsberg, *Lectures—Theor. Phys. Summer Institute*, (Univ. of Colorado Press, Boulder, 1966), **8A**, p. 313.
94. V.L. Bonch-Bruевич and E.G. Landsberg, *Phys. Stat. Sol.*, **29**, (1968) pp. 9–43.
95. L. Rosenfeld, *Theory of Electrons*, (North Holland, Amsterdam, 1951).
96. P. Norzieses and D. Pines, *Phys. Rev.*, **109**, 741 (1958).
97. R.W. Ditchburn, *Eye Movement and Visual Perception*, (Oxford Univ. Press, Oxford, 1973); also *Light*, (Blackie, London, 1952).
98. J.T. Houghton and S.D. Smith, *Infrared Physics*, (Clarendon, Oxford, 1966).
99. J.L. Pankove, *Optical Processes in Semiconductors*, (Dover, New York, 1975).
100. W. Franz, *Z. Naturforsch.*, **13A**, 484 (1958).
101. L.V. Keldysh, *Soviet Physics*, (JETP), **7**, 788 (1958).
102. L.V. Keldysh, *J. Exptl. Theoret. Phys. (U.S.S.R.)*, **47**, 1945 (1965); English translation: *Soviet Phys. JETP*, **20**, 1307 (1965).
103. S. Wang, *Solid State Electronics*, (McGraw-Hill, New York, 1966), p. 369.
104. T.S. Moss, *J. Appl. Phys.*, Supplement, **32**, 2136 (1966).
105. G.S. Paige and H.D. Rees, *Phys. Rev. Lett.*, **16**, 444 (1966).
106. A. Frova and P. Handler, *Phys. Rev.*, **137A**, 1857 (1965).
107. R. Williams, *Phys. Rev.*, **117**, 1487 (1960).
108. E.O. Kane, *J. Phys. Chem. Solids*, **12**, 181 (1959).
109. T.N. Morgan, *Phys. Rev.*, **148**, 890 (1966).
110. J. Frenkel, *Phys. Rev.*, **38**, 309 (1931).
111. R. Peierls, *Ann. Phys.*, **13**, 905 (1932).
112. G.H. Wannier, *Phys. Rev.*, **52**, 191 (1937).
113. N.F. Mott, *Trans. Faraday Soc.*, **34**, 500 (1938).
114. A.S. Davydov, *Theory of Molecular Excitons*, translated by M. Kasha and M. Oppenheimer, (McGraw-Hill, New York, 1962).
115. F.O. Rice and F. Teller, *The Structure of Matter*, (Wiley, New York, 1949).
116. P.W. Baumeister, *Phys. Rev.*, **121**, 359 (1961).
117. E.F. Gross and A. Kaplianski, *Zh. Tekh. Fiz.*, **25**, 2061 (1955).
118. H. Meier, *Organic Semiconductors*, (Verlag Chemie, Weinheim, Germany, 1974).
119. M. Pope and C.E. Swenberg, *Electronic Processes in Organic Crystals*, (Clarendon, Oxford, 1982).
120. D.L. Dexter and R.S. Knox, *Excitons*, (Wiley, New York, 1965).
121. R.S. Knox, *Theory of Excitons*, *Solid State Physics Supplement 5*, (Academic Press, New York, 1963).

122. M. Trlifaj, Czech, J. Phys., *8*, 510 (1958).
123. R.W. Munn and W. Siebrand, J. Chem Phys., *52*, 47 (1970).
124. R.C. Powell and Z.G. Soos, Phys. Rev., *B5*, 1547 (1972).
125. D.L. Dexter, J. Chem Phys., *21*, 836 (1953).
126. M. Pope, J. Chem. Phys., *47*, 2197 (1967).
127. G.R. Johnston and L.E. Lyon, Chem. Phys. Lett., *2*, 489 (1968).
128. C.L. Braun and G.M. Dobbs, J. Chem. Phys., *53*, 2718 (1970).
129. S. Singh, W.J. Jones, W. Siebrand, B.P. Stoicheff, and W.G. Schneider, J. Chem. Phys. *42*, 330 (1965).
130. W. Helfrich and F.R. Lipsett, J. Chem. Phys., *43*, 4368 (1965).
131. H.C. Wolf and K.W. Benz, Pure Appl. Chem., *27*, 439 (1971).
132. R.C. Powell and Z.G. Soos, J. Luminescence, *11*, 1 (1975).
133. N. Wakayama and D.F. Williams, J. Chem. Phys., *57*, 1770 (1972).
134. M. Schott and I. Berehar, Mol. Cryst. Liq. Cryst. *20*, 13 (1973).
135. M. Pope, J. Burgos, and N. Wotherspoon, Chem. Phys. Lett., *12*, 140 (1971).
136. R.R. Chance and A. Prock, Phys. Stat. Sol. (b), *57*, 597 (1973).
137. M.E. Michel-Beyerle, W. Harengel, and R. Haberkorn, Mol. Cryst. Liq. Cryst., *25*, 323 (1974).
138. S.A. Rice and J. Jortner, "Comments on the Theory of the Excited States in Molecular Crystals," in *Physics and Chemistry of Organic Solid State*, edited by D. Fox, M.M. Labes, and A. Weissberger, *3*, (1967) pp. 199–497.
139. M.G. Sceats and S.A. Rice, J. Chem. Phys., *62*, 1098 (1975).
140. J.R. Lakowicz, *Principles of Fluorescence Spectroscopy*, (Plenum, New York 1983).
141. L. Onsager, Phys. Rev., *84*, 554 (1938).
142. D. Curie, *Luminescence in Crystals*, (Methuen, London, 1963).
143. P. Avakian and R.E. Merrifield, Mol. Cryst., *5*, 37 (1968).
144. W. Seibrand, J. Chem. Phys., *42*, 3951 (1965).
145. M. Pope and R. Selsby, Chem. Phys. Lett., *14*, 226 (1972).
146. T.G. Dewey (Ed.), *Biophysical and Biochemical Aspects of Fluorescence Spectroscopy*, (Plenum, New York, 1991).
147. W.T. Mason, *Fluorescent and Luminescent Probe for Biological Activity*, (Academic Press, New York, 1993).
148. A.A. Bergh and P.J. Dean, *Light-Emitting Diodes*, (Oxford University Press, Oxford, 1976), Chapter 5.
149. K. Gillesen and W. Schaiver, *Light Emitting Diodes—Introduction*, (Prentice Hall International, London, 1987).
150. H.F. Ivey, "Electroluminescence and Related Effects," Advan. Electron. Phys., Supplement 1, 1963.
151. G. Destriau, J. Chem. Phys., *33*, 620 (1936).
152. A. Vecht, J. Vac. Sci. Technol., *10*, 789 (1973).
153. D. Kahng, Appl. Phys. Lett., *13*, 210 (1968).
154. J. Russ and D.I. Kennedy, J. Electrochem. Soc., *114*, 1066 (1967).
155. H.K. Henisch, *Electroluminescence*, (Pergamon, Oxford, 1962).
156. T. Inoguchi and S. Mito, in *Electroluminescence*, edited by J.I. Pankove, (Springer Verlag, Berlin, 1976).
157. W. Helfrich and W.G. Schneider, Phys. Rev. Lett., *14*, 29 (1965).
158. W. Helfrich and W.G. Schneider, J. Chem. Phys., *44*, 2902 (1966).
159. W. Hwang and K.C. Kao, J. Chem. Phys., *60*, 3845 (1974).
160. A. Sana, Phys. Rev., *B1*, 1716 (1970).
161. D.F. Williams and M. Schadt, J. Chem. Phys., *53*, 3480 (1970).
162. W. Mehl and B. Funk, Phys. Lett., *25A*, 364 (1967).
163. H.P. Schwob and I. Zschokke-Granacher, Mol. Cryst. Liq. Cryst., *13*, 115 0(1971).
164. M. Kawabe, K. Masuda, and S. Namba, Japan J. Appl. Phys., *10*, 527 (1971).
165. N. Wakayama and D.F. Williams, Chem. Phys. Lett., *9*, 45 (1971).
166. H.P. Kunkel and K.C. Kao, J. Phys. Chem. Solids, *37*, 863 (1976).
167. H.F. Ivey, "Electroluminescence and Semiconductor Lasers," IEEE J. Quantum Electron., *QE-2*, 713 (1966).
168. S.M. Sze, *Physics of Semiconductor Devices*, (Wiley Interscience, New York, 1981).
169. F. Gutmann and L.E. Lyons, *Organic Semiconductors*, (Wiley, New York, 1967).
170. M. Kochi, Y. Harada, I. Hirooka, and H. Inokuchi, Bull. Chem. Soc. Japan, *43*, 2690 (1970).
171. R. Williams, "Injection by Internal Photemission," in *Semiconductors and Semimetals*, edited by R.K. Willardson and A.C. Beer, Vol. 6, (1970) pp. 97–139.

172. G. Vaubel and H. Basessler, *Phys. Stat. Sol.*, 26, 599 (1968).
173. J.M. Caywood, *Mol. Cryst. and Liq. Cryst.*, 12, 1 (1970).
174. W. Ruppel, "The Photoconductor-Metal Contact," in *Semiconductors and Semimetals*, 6, (1970) pp. 315-345.
175. A. Rose, *Concepts in Photoconductivity and Allied Problems*, (Wiley, New York, 1963).
176. R.H. Fowler, *Phys. Rev.*, 38, 45 (1931).
177. A. L. Hughes and L.A. DuBridge, *Photoelectric Phenomena*, (McGraw-Hill, New York, 1932).
178. J. Kadlec and K. H. Gundlach, *Phys. Stat. Sol.*, (A) 37, 11 (1976).
179. F. Abeles (Ed.), *Optical Properties of Solids*, (North -Holland, Amsterdam and New York, 1972).
180. A. M. Goodman, *J. Appl. Phys.*, 35, 573 (1964).
181. C.A. Mead and W.G. Spitzer, *Phys. Rev.*, 134A, 713 (1964); also *Appl. Phys. Lett.*, 2, 74 (1963).
182. A.M. Cowley and H. Heffner, *J. Appl. Phys.*, 35, 255 (1964).
183. H.G. White and R.A. Logan, *J. Appl. Phys.*, 34, 1990 (1963).
184. J.J. Tietjen and J.A. Amick, *J. Electrochem. Soc.*, 113, 724 (1966).
185. J. Kadlec and K.H. Gundlach, *Solid State Commun.*, 16, 621 (1975).
186. K.H. Gundlach and J. Kadlec, *J. Appl. Phys.*, 46, 5286 (1975).
187. G. Lewicki, J. Maserjian, and C.A. Mead, *J. Appl. Phys.*, 43, 1764 (1972).
188. K.H. Gundlach and J. Kadlec, *Appl. Phys. Lett.*, 20, 445 (1972).
189. C.N. Berglund and R.J. Powell, *J. Appl. Phys.*, 42, 573 (1971).
190. J. Kadlec and K.H. Gundlach, *J. Appl. Phys.*, 47, 672 (1976).
191. R. Williams and J. Dresner, *J. Chem. Phys.*, 46, 2133 (1967).
192. J. Dresner, *Phys. Rev. Lett.*, 21, 356 (1968).
193. E.G. Ramberg, *Appl. Optics*, 6, 2163 (1967).
194. W.E. Spicer, *Phys. Rev.*, 154, 385 (1967).
195. E.O. Kane, *Phys. Rev.*, 127, 131 (1962).
196. G.W. Gobeli and F.G. Allen, "Photoelectric Threshold and Work Function," in *Semiconductors and Semimetals*, edited by R.K. Willardson and A.C. Beer, Vol. 2, (Academic Press, New York, 1966), pp. 263-300.
197. J. Sworakowski, *Phys. Stat. Sol. (a)* 13, 381 (1972).
198. J. Sworakowski, *Phys. Stat. Sol. (a)* 22, K73 (1974).
199. K. Seki, T. Hirooka, Y. Kamura, and H. Inokuchi, *Bull. Chem. Soc. Japan*, 49, 904 (1976).
200. M. Cardona and L. Ley (eds), *Photoemission in Solids*, Vol. 1 (1978) and Vol. 2 (1979), (Springer-Verlag, New York, 1978 and 1979).
201. S. Yoshimura, *J. Phys. Soc. Japan*, 28, 701 (1970).
202. B.H. Schechtma, S.F. Lin, and W.E. Spicer, *Phys. Rev. Lett.*, 34, 667 (1975); and also; B.H. Schechtman, Ph.D. thesis, Stanford University, 1968.
203. T. Hirooka, K. Tanaka, K. Kuchitsu, M. Fujihara, H. Inokuchi, and Y. Harada, *Chem. Phys. Lett.*, 18, 390 (1973).
204. W.E. Spicer and R.E. Simon, *J. Phys. Chem. Solids*, 23, 1817 (1962).
205. R.C. Eden, J.L. Moll, and W.E. Spicer, *Phys. Rev. Lett.*, 18, 597 (1967).
206. M. Pope, H. Kallmann, and J. Giachino, *J. Chem. Phys.*, 42, 2540 (1965).
207. H. Demmer, *Physik Zeits*, 32, 554 and 856 (1931); and 33, 209 (1932).
208. R. Robertson, T.T. Fox, and A.M. Martin, *Nature*, 129, 579 (1932).
209. T.S. Moss, L. Pincherle, and A.M. Woodward, *Proc. Phys. Soc. (London)*, 66B, 743 (1953).
210. T. Tauc, *Rev. Mod. Phys.*, 29, 308 (1959).
211. H. Kallmann, "Energy Transfer Processes," in *Comparative Effects of Radiation*, edited by M. Burton, J.S. Kirby Smith and J.L. Magee, (Wiley, New York, 1960).
212. H. Kallmann and M. Pope, *J. Chem. Phys.*, 30, 585 (1959).
213. I.K. Kikoin and M.M. Noskov, *Phys. Z. Sowjet*, 5, 586 (1934).
214. O. Garreta and J. Grosvalet, *Progress in Semiconductors*, Vol. 1, 165 (Heywood, London, 1956).
215. W. Van Roosbroek, *Phys. Rev.*, 101, 1713 (1956).
216. C.A. Mead, *Solid State Electron.*, 9, 1023 (1966).
217. C.R. Crowell, W.G. Spitzer, L.E. Howarth, and E.E. Labate, *Phys. Rev.*, 127, 2006 (1962).
218. H.J. Novel, "Solar Cells" in *Semiconductors and Semimetals*, Vol. 11, edited by R.F. Willardson and A.C. Beer (Academic Press, New York, 1975).
219. E.H. Rhoderick, *Metal-Semiconductor Contacts*, (Clarendon Press, Oxford, 1978).

220. H.C. Card and E.S. Yang, *Appl. Phys. Lett.*, **29**, 51 (1976).
221. R.J. Van Overstraeten and R.P. Mertens, *Physics, Technology and Use of Photovoltaics*, (Adam Hilger, Bristol and Boston, 1986).
222. D.A. Jenny, J.J. Loferski, and P. Rappaport, *Phys. Rev.*, **101**, 1208 (1956).
223. A.K. Sreedhar, B.L. Sharma, and R.K. Purohit, *IEEE Trans. Electron. Devices*, **ED-16**, 309 (1969).
224. H.J. Hovel and J.M. Woodall, *J. Electrochem. Soc.*, **120**, 1246 (1973).
225. D.A. Cusano, *Solid State Electron.*, **6**, 217 (1963).
226. D.E. Carlson, "Solar Cells" in *Semiconductors and Semimetals*, Vol. 21, Part D (Academic Press, New York, 1984).
227. G. Nakamura, R. Sato, and Y. Yukimoto, *Japan J. Appl. Phys.*, **21**, Suppl. 297 (1982).
228. S. Nakano, H. Tarui, T. Takahama, M. Isomura, T. Matsuyama, H. Haku, M. Nishikuni, S. Tsuda, M. Ohnishi, Y. Kishi, and Y. Kuwano, *Proc. 7th EC Photovoltaic Solar Energy Conference*, Seville, Spain, 1986, p. 27; also *Jpn. J. Appl. Phys.*, **25**, 1936 (1986).
229. A. Nakano, S. Ikegama, H. Matsumoto, H. Uda, and Y. Komatsu, *Solar Cells*, **17**, 233 (1986).
230. K. Takahashi and M. Konsgai, *Amorphous Silicon Solar Cells*, (Wiley, New York, 1986).
231. M. Wolf, *Proc. IRE*, **48**, 1246 (1960).
232. M.A. Green, *High Efficiency Silicon Solar Cells* (Trans. Tech. Publications, Brookfield, 1987).
233. Y. Hamakawa, "Amorphous Silicon Solar Cells," in *Current Topics in Photovoltaics* edited by T.J. Coutts and J. Meakin (Academic Press, New York, 1985), pp. 111–165.
234. C.E. Backus, ed., *Solar Cells*, (IEEE Press, New York, 1976).
235. M.A. Green, *Solar Cells: Operating Principles, Technology and System Application*, (Prentice Hall, New Jersey, 1982).
236. W.C. Sinke, ed., "The Photovoltaic Challenge," in *MRS Bulletin* of October 1993 (Materials Research Society, Pittsburgh, 1993), pp. 18–66.
237. B. O'Regan and M. Gratzel, *Nature*, **353**, 737 (1991).
238. M. Gratzel, "Nanocrystalline Thin Film PV Cells," in *MRS Bulletin* of October 1993 (Materials Research Society, Pittsburgh, 1993), p. 61.
239. B. O'Regan, J. Moser, M. Anderson, and M. Gratzel, *J. Phys. Chem.*, **94**, 8720 (1990).
240. M.K. Nazeeruddin, A. Kay, J. Rodicio, R. Humphrey-Baker, E. Muller, P. Liska, N. Vlachopoulos, and M. Gratzel, *J. Am. Chem. Soc.*, **115**, 6382 (1993).
241. A. Kay and M. Gratzel, *J. Phys. Chem.*, **97**, 6272 (1993).
242. G. Hodes, I.O.F. Howell, and L.M. Peter, *J. Electrochem. Soc.*, **130**, 3136 (1992).
243. E.I. Adirovich, T. Mirzamakhmudov, V.M. Rubinov, and Yu. M. Yuabov, *Sov. Phys.—Solid State*, **7**, 2946 (1966).
244. H. Kallmann, G.M. Spruch, and S. Trester, *J. Appl. Phys.*, **43**, 469 (1972).
245. W. Ma, R.M. Anderson, and S.J. Hruska, *J. Appl. Phys.*, **46**, 2650 (1975).
246. H. Kallmann, B. Kramer, E. Maidemanakis, W.J. McAleer, H. Barke-meyer, and P.I. Pollak, *J. Electrochem. Soc.*, **108**, 247 (1961).
247. E.I. Adirovich, Y.M. Rubinov, and Yu. M. Yuabov, *Sov. Phys.—Solid State*, **6**, 2540 (1965).
248. M.I. Korsunskii and M.M. Sominskii, *Soviet Phys.—Semicon.*, **7**, 342 (1973).
249. H. Onishi, S. Kurokawa, and K. Leyansu, *J. Appl. Phys.*, **45**, 3205 (1974).
250. F.H. Nicoll, *J. Electrochem. Soc.*, **110**, 1165 (1963).

4 Ferroelectrics, Piezoelectrics, and Pyroelectrics

The whole of science is nothing more than a refinement of everyday thinking. Science is the attempt to make the chaotic diversity of our sense experience correspond to a logically uniform system of thought.

Albert Einstein

4.1 Introductory Remarks

The term *ferroelectrics* arose by analogy with ferromagnetics, mainly because they have similar characteristics: under electric fields for ferroelectric phenomena and under magnetic fields for ferromagnetic phenomena. The prefix *ferro-* derived from *ferum*, which means iron in Latin. The term is perfect for ferromagnetics, since all ferromagnetic phenomena are associated with the special type of spin arrangement of the iron atoms. But in ferroelectrics there are no iron atoms, so the prefix does not mean iron. Rather, it implies the similarity in characteristics to ferromagnetics. Like ferromagnetics, ferroelectrics exhibit a spontaneous electric polarization below the Curie temperature, a hysteresis loop, and an associated mechanical strain. However, ferroelectrics differ from ferromagnetics in their fundamental mechanisms and also in some of their applications.

In Europe, ferroelectrics are sometimes called *Seignette electrics*. This term is somewhat misleading, because Seignette did not discover the ferroelectric phenomena. Instead, in 17th-century Rochelle, France, he discovered Rochelle salt (potassium-sodium tartrate-tetrahydrate, $\text{KNaC}_4\text{H}_4\text{O}_6 \cdot 4 \text{H}_2\text{O}$), a colorless crystalline compound with an orthorhombic structure. At that time, the material was used as a laxative. More than 200 years later, in 1921, Valasek discovered the ferroelectric phenomena in the same material.¹⁻³ Ferroelectrics were discovered much later than ferromagnetics. However, there are now more than 1,000 solid materials possessing ferroelectric properties.

The prefix *piezo-* in the word *piezoelectrics* is derived from a Greek word, *piezein*, meaning *pressure*. Piezoelectrics are materials in which electricity can be generated by an applied mechanical stress or a mechanical stress can be produced by an applied electric field. This interconvertible behavior was first discovered by Pierre and Jacques Curie in 1880 in certain crystals, such as quartz, zinc blends, tourmaline, and Rochelle salt.^{4,5} The term *piezoelectricity* has been used by scientists since 1881 to distinguish the piezoelectric phenomena from electrostriction. The piezoelectric phenomena occurs in both the ferroelectric and the nonferroelectric states.^{2,3}

The prefix *pyro-* in the term *pyroelectrics* means *heat* in Greek. Thus, pyroelectricity means heat-generated electricity. This effect is also convertible. This implies that heat can be generated by electricity resulting from the change of the state of electric polarization, such as electrothermal and electrocaloric effects. Pyroelectric phenomena were in fact discovered much earlier. Theophrast discovered the pyroelectric phenomena in tourmaline in 314 BC, and the word *pyroelectricity* has been used since 1824, first by Brewster.⁶ The theory of the pyroelectric effect was first formulated by W. Thomson (Lord Kelvin) in 1878 and further developed by Voigt in 1897 and later by Born, based on his famous crystal lattice dynamics in 1921 and 1928.⁷

In general, all materials undergo a small change in dimension when subjected to an external force, such as an applied electric field, a mechanical stress, or a change in temperature

(heat). Depending on the material structure, such a small change in dimension may result in a change in electric polarization and hence give rise to the occurrence of the ferroelectric, piezoelectric, or pyroelectric effects. It can be imagined that materials exhibiting these effects must be polar and have an electrical order, implying that they must be crystals or polycrystalline materials composed of crystallites. A crystal or a crystallite must have a definite chemical composition, with the molecules made up of positive ions (atoms sharing part of their valence electrons with others) and negative ions (atoms receiving part of electrons from others) occupying lattice sites to constitute a crystal structure lattice. The smallest repeating unit of the lattice is called the *unit cell*, and the specific symmetry of the unit cell determines whether the crystal exhibits ferroelectric, piezoelectric, pyroelectric, or electro-optic effects.

On the basis of the symmetry elements of translational position and orientation, there are 230 space groups. Ignoring translational repetition, these 230 groups break down into 32 classes, known as the 32 point groups. Point groups are based on orientation only.^{8,9} Any point may be defined by coordinates x , y , and z , with respect to the origin of symmetry. A centrosymmetric crystal is a crystal in which the movement of each point at x , y , z , to a new point at $-x$, $-y$, $-z$, does not cause a recognizable difference. This implies that centrosymmetric crystals are nonpolar and thus do not possess a finite polarization or dipole moment. Of the 32 classes (or point groups), 11 classes are centrosymmetric and 21 classes are noncentrosymmetric, possessing no center of symmetry. The latter is the necessary requirement for the occurrence of piezoelectricity.

However, one of the 21 classes, though classified as the noncentrosymmetric class, possesses other combined symmetry elements, thus rendering no piezoelectricity. So, only 20 classes of noncentrosymmetric crystals would exhibit piezoelectric effects. In 10 of these 20 classes, polarization can be induced by a mechanical stress, while the other 10 classes possess spontaneous polarization, so they are permanently polar and thus can have piezo-

electric as well as pyroelectric effects. There is a subgroup within these 10 classes that possesses spontaneous polarization and reversible polarization; this subgroup can exhibit all three effects—ferroelectric, piezoelectric, and pyroelectric. In fact, the ferroelectric effect is an empirical phenomenon distinct from piezoelectric and pyroelectric effects in that it exists with a reversible polarization. The relationship between polarization behavior and crystal structure is shown in Figure 4-1.

The 32 point groups are subdivisions of seven basic crystal systems, which are based on the degree of symmetry. In the order of ascending symmetry, these seven basic crystal systems are triclinic, monoclinic, orthorhombic, tetragonal, trigonal (rhombohedral), hexagonal, and cubic. For a more detailed description of these seven systems, see references 8 and 9. Based on the optical properties, these systems can be classified into three major optical groups, as shown in Table 4-1. The optical biaxial group refers to the crystals in which the molecules are not spherical, with the long axis longer than the axes perpendicular to it. Depending on the structure of the crystal, if the index of refraction in the direction along the long axis is different from that along the direction of the short axis, the crystals are optically biaxial and exhibit birefringence. If the indices of refraction in the two axes are identical, then the crystals are optically uniaxial. However, some

Table 4-1 The seven basic crystal systems.

Crystal System	Optical Group	Order of Symmetry
Triclinic	Biaxial	Very Nonsymmetric
Monoclinic	Biaxial	Increasing 
Orthorhombic	Biaxial	
Tetragonal	Uniaxial	
Trogonal (Rhombohedral)	Uniaxial	
Hexagonal	Uniaxial	
Cubic	Optically Isotropic	Very Symmetric

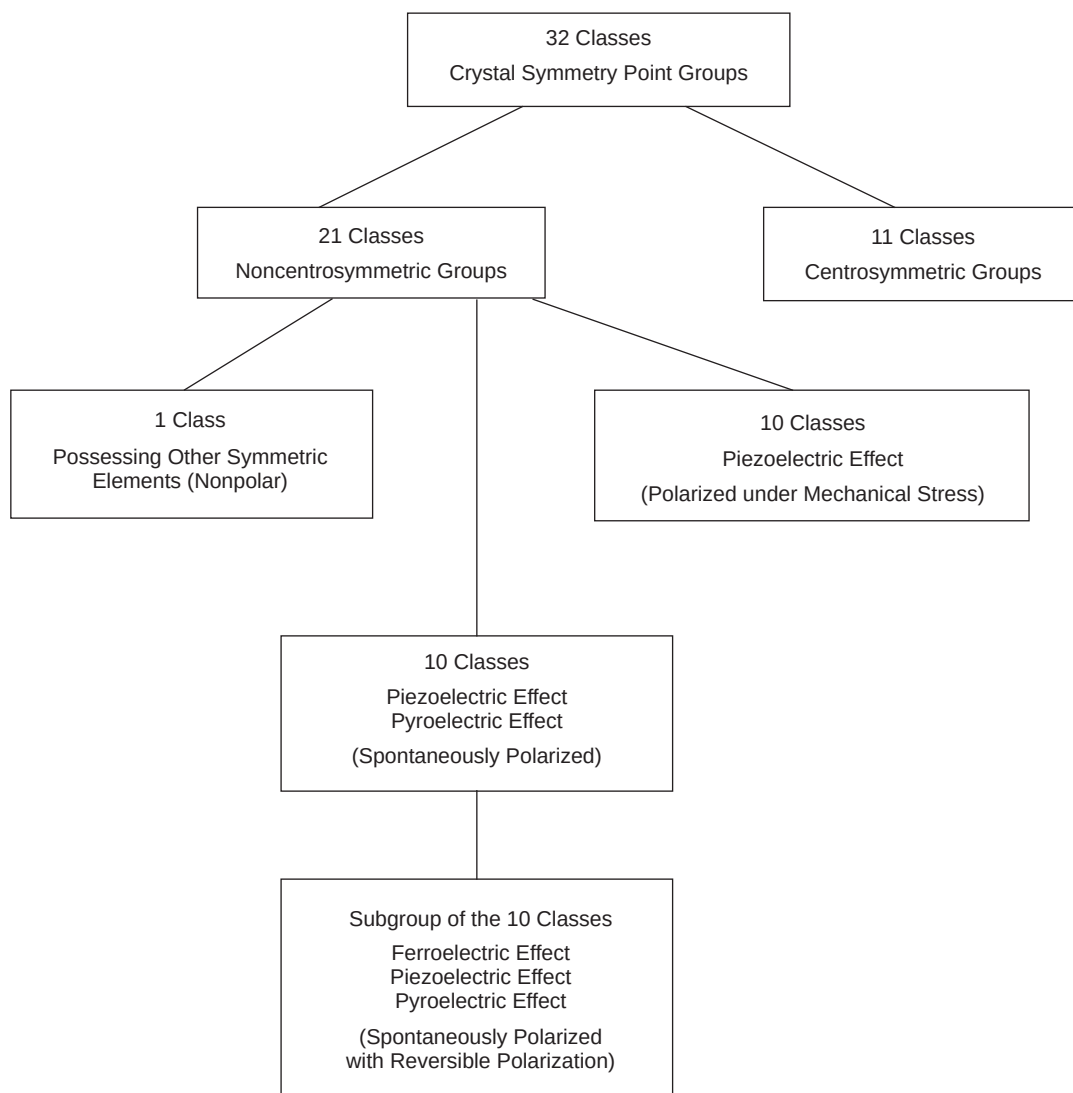


Figure 4-1 Classification of crystals showing the classes with piezoelectric, pyroelectric, and ferroelectric effects.

crystals with their constituent molecules being close to spherical particles in shape; then in such cases the index of refraction is identical in all directions and the crystals are optically isotropic.

Apart from the crystals with the structures just described, some synthetic and biological polymers with different structures also exhibit piezoelectric, pyroelectric, and ferroelectric phenomena.¹⁰⁻¹⁵ In view of the complex molecular, crystalline, and morphological structures

of polymers, there may exist a possibility that the combined collective and coordinate effects of the structure would satisfy the requirements for the occurrence of piezoelectric, pyroelectric, and even ferroelectric phenomena. Polyvinylidene fluoride (PVDF) polymer is a good example that complies with such requirements under certain conditions. However, polymers are potential materials suited particularly for many practical applications because of their easy processability into

thin, tough, and flexible films. We shall discuss polymers having these properties later. In the present chapter, we shall deal only with ferroelectric, piezoelectric, and pyroelectric effects. Electro-optic effects were discussed in Chapter 3.

4.2 Ferroelectric Phenomena

Ferroelectricity is one of the most fascinating properties of dielectric solids. Materials exhibiting ferroelectric properties must be either single crystals or polycrystalline solids composed of crystallites; they must also possess reversible spontaneous polarization. In this section, we shall discuss the various features of ferroelectrics, the mechanisms responsible for the appearance of these features, and ferroelectric materials and their applications.

4.2.1 General Features

The polarization induced by an externally applied field in normal dielectric materials is very small, with the dielectric constant usually less than 100, and its effects on other physical

properties are also very small. However, there are a number of crystals with a nonsymmetrical structure (see Figure 4-1) that exhibit a large polarization, with the dielectric constant up to 10^5 , under certain conditions. Obviously, such a large magnitude of polarization has attracted many researchers to study it theoretically and to develop various practical applications.

A ferroelectric crystal shows a reversible spontaneous electric polarization and a hysteresis loop that can be observed in certain temperature regions, delimited by a transition point called the Curie temperature, T_c . At temperatures above T_c , the crystal is no longer ferroelectric and exhibits normal dielectric behavior. Ferroelectric materials usually, but not always, exist in a nonpolar state at temperatures above T_c , and have anomalously high dielectric constants, especially near the Curie temperature. Typical dielectric constant–temperature and polarization–temperature characteristics are shown in Figure 4-2. The dielectric constant increases very rapidly to a very high peak value at T_c . The anomalously high value of ϵ_r in the neighborhood of T_c is generally referred to as the *anomalous value*. At $T > T_c$, anomalous

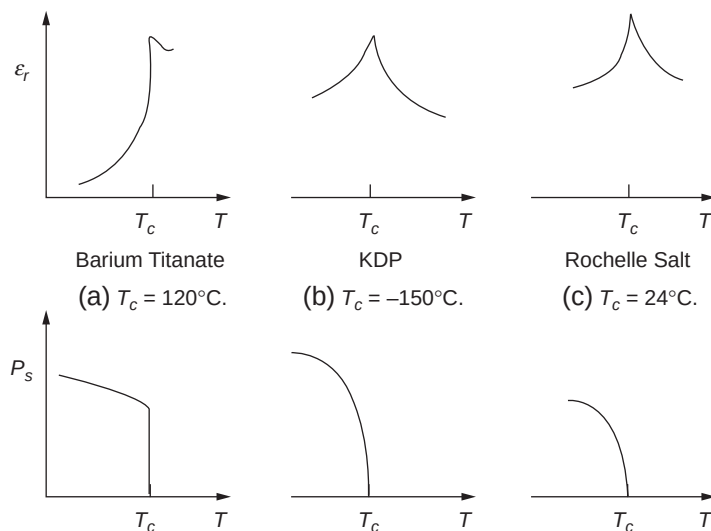


Figure 4-2 Schematic illustration of the variation of the dielectric constant ϵ_r and the spontaneous polarization of P_s with temperature for three typical ferroelectric crystals: (a) Barium titanate (BaTiO_3) with $T_c = 120^\circ\text{C}$, (b) Potassium dihydrogen phosphate (KDP, KH_2PO_4) with $T_c = -150^\circ\text{C}$, and (c) Potassium Sodium tartrate-tetrahydrate (Rochelle Salt, $\text{KNaC}_4\text{H}_4\text{O}_6 \cdot 4\text{H}_2\text{O}$) with $T_c = 24^\circ\text{C}$.

behavior follows closely the Curie–Weiss relation

$$\epsilon_r = \frac{C}{T - T_c} \quad (4-1)$$

where C is known as the Curie constant. In fact, anomalous behavior always appears near any transition point between two different phases, even at T below T_c . At the transition points, there are anomalies not only in the dielectric constant and polarization, but also in piezoelectric and elastic constants and specific heat, because of the change in crystal structure.

Ferroelectrics have reversible spontaneous polarization. The word *spontaneous* may mean that the polarization has a nonzero value in the absence of an applied electric field. The word *reversible* refers to the direction of the spontaneous polarization that can be reversed by an applied field in opposite direction. The spontaneous polarization P_s usually increases rapidly on crossing the transition point and then gradually reaches a saturation value at lower temperatures. The most prominent features of ferroelectric properties are hysteresis and nonlinearity in the relation between the polarization P and the applied electric field F . The simplest method for measuring spontaneous polarization is the Sawyer and Tower method,¹⁶ as shown in Figure 4-3, in which C is the capacitance of the ferroelectric specimen and C_o is the standard capacitor. The voltage across C should be sufficiently large to render a saturation in polarization, so V_o should be proportional to the polarization charge, $V_o = AP/C_o$,

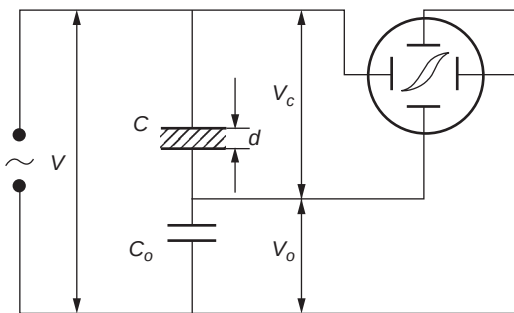


Figure 4-3 The Sawyer–Tower method for the measurement of the polarization–electric field (P–F) characteristics.

where A is the area of the specimen. V is the applied voltage, which is usually an AC signal voltage of low frequencies. Thus, the applied field across the specimen is $F = V_c/d = (V - V_o)/d$.

A typical hysteresis loop is shown schematically in Figure 4-4. When the field is small, the polarization increases linearly with the field. This is due mainly to field-induced polarization, because the field is not large enough to cause orientation of the domains (portion OA). At fields higher than the low-field range, polarization increases nonlinearly with increasing field, because all domains start to orient toward the direction of the field (portion AB). At high fields, polarization will reach a state of saturation corresponding to portion BC , in which most domains are aligned toward the direction of the poling field. Now, if the field is gradually decreased to zero, the polarization will decrease, following the path CBD . By extrapolating the linear portion CB to the polarization axis (or zero-field axis) at E , OE represents the spontaneous polarization P_s and OD represents the remanent polarization P_r . The linear increase in polarization from P_s to P_p is due mainly to the normal field-induced dielectric polarization. P_r is smaller than P_s because when the field is reduced to zero, some domains may return to their original positions due to the

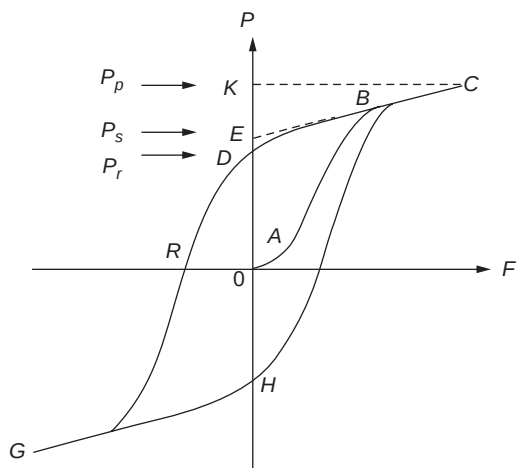


Figure 4-4 Schematic diagram of a typical ferroelectric hysteresis loop.

strain situation, thus reducing these domains' contribution to the net polarization.

For most ferroelectric materials, the component due to the normal field-induced dielectric polarization is very small compared to the spontaneous polarization; therefore, for most applications, this component can be ignored. The magnitude of the difference between P_p and P_s in Figure 4-4 is exaggerated for the purpose of clear illustration. The field required to bring the polarization to zero is called the coercive field F_c (portion OR on zero polarization axis). F_c depends not only on temperature, but also on the measuring frequency and the waveform of the applied field. When the field in the opposite direction decreases to zero, the polarization is reversed, indicating that domains (see Section 4.2.4) have already been formed before poling and that the motion of the domain walls results in the change of direction of polarization. The hysteresis arises from the energy needed to reverse the metastable dipoles during each cycle of the applied field. The area of the loop represents the energy dissipated inside the specimen as heat during each cycle. In general, the hysteresis loop is measured with AC fields at low frequencies, 60 Hz or lower, to avoid heating the specimen.

In general, ferroelectricity is harder to demonstrate in polycrystalline materials composed of crystallites, such as ceramics, than in a single crystal because of the random orientation of crystallites. This is why in some single crystals the polarization reverses quite abruptly

to form a square loop, as shown in Figure 4-5(a), while in most ceramics the loop is rounded, as shown in Figure 4-5(b), because of the more sluggish reversal, which is due partly to the axes of the unit cells in the randomized arrangement of the nonuniform crystallites.

Ferroelectric materials exhibit ferroelectric properties only at temperatures below T_c because they are polar; at temperatures above it, they are not polar. Obviously, the shape of the hysteresis loop depends on temperature. Figure 4-6 shows the shape of the hysteresis loop for Rochelle salt at two different temperatures. The loop becomes gradually diminished at $T > T_c$, eventually degenerating to a straight line at T much larger than T_c , when the ferroelectric behavior disappears completely. However, some ferroelectric materials can be driven from the paraelectric state to a ferroelectric state at $T > T_c$ by an applied field larger than a certain critical value F_t . In other words, the applied electric field larger than F_t tends to shift the Curie point to a higher temperature. The higher the temperature above T_c , the higher the field required to induce ferroelectricity. If a ferroelectric material is subjected to a large AC field at a temperature slightly higher than T_c , two small hysteresis loops may appear, as shown schematically in Figure 4-7. A double hysteresis loop has been observed in BaTiO_3 at a temperature a few degrees above T_c .^{17,18} For the transition from the tetragonal structure to the orthorhombic structure at the transition temperature of 5°C for BaTiO_3 at an AC field

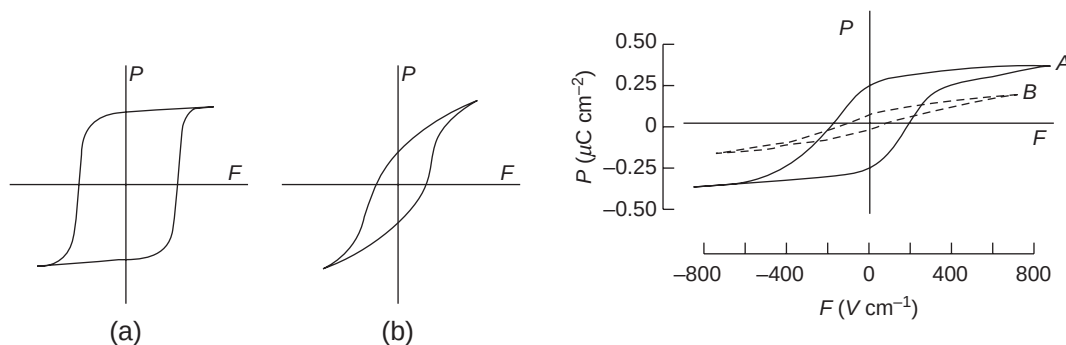


Figure 4-5 Schematic diagrams of (a) the “square” and (b) the “rounded” hysteresis loops.

Figure 4-6 Temperature dependence of the hysteresis loop for Rochelle salt with $T_c = 24^\circ\text{C}$. (A) $T = 19^\circ\text{C} < T_c$, and (B) $T = 42^\circ\text{C} > T_c$.

higher than F_t , a triple hysteresis loop, consisting of one main loop and two small loops, has also been observed.¹⁹

Using the circuit shown in Figure 4-3, the voltage across C is proportional to the polar-

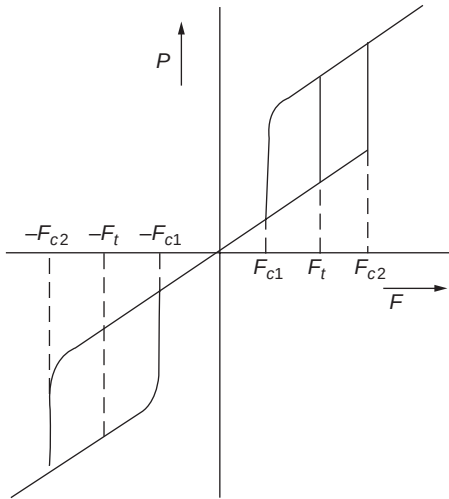


Figure 4-7 Schematic diagram showing a typical double hysteresis loop at T slightly higher than T_c and at F larger than a critical value F_t .

ization P , which is the dipole moments per unit volume and is also equal to the surface charge density on the electrode. During polarization, a current $j = dP/dt$ will flow through the specimen per unit area. Suppose that the wave shape of the switching signal voltage is as shown in Figure 4-8(a) for two cases: one with the voltage (or field) increasing linearly with time and the other with the field decreasing linearly with time. In this case, the resulting current flowing through the specimen, for a normal linear dielectric, is constant, as shown in Figure 4-8(b). But if the specimen is a ferroelectric material, the current will be a short pulse, as shown in Figure 4-8(c). This is one of the simplest ways to distinguish ferroelectric from nonferroelectric materials.

The size of the unit cell and the force of ions in the lattice of a crystal are temperature dependent. As the temperature changes, there exists a critical temperature (i.e., the transition temperature), at which a particular structure becomes unstable and tends to transform to a more stable one. Although the transition involves only small ionic movements, they cause a marked change in properties. A change in dimensions of a crystal or a crystallite will

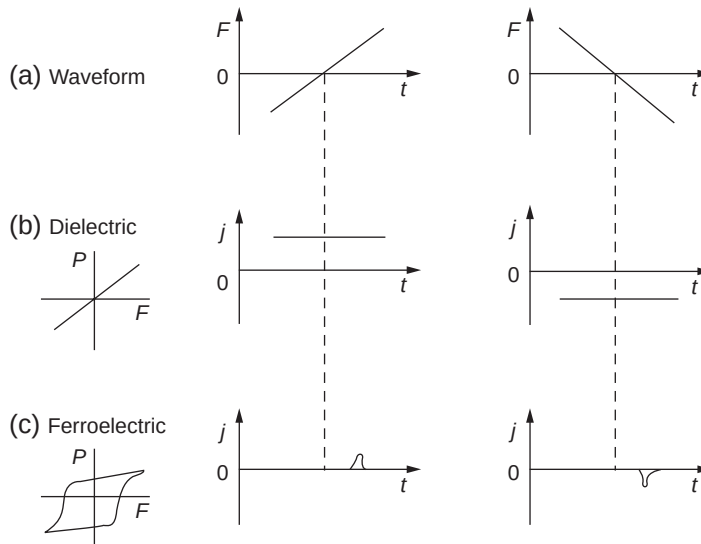


Figure 4-8 Schematic diagrams showing (a) the waveforms of the linearly time-varying electric field (b) the corresponding current j flowing through the loss-free dielectric material, and (c) the corresponding current j flowing through the ferroelectric material.

create internal stresses, particularly at the boundaries between crystallites in polycrystalline materials, such as in ceramics. Under certain conditions, the magnitude of such stresses may be large enough to cause internal cracks.

Transition from one crystal structure to another usually involves a change in entropy and volume. In general, there are two different orders of transition. When the spontaneous polarization goes from zero to a finite value, or from one value to another, the change in polarization may be continuous or discontinuous. If there is discontinuity in the change of the polarization, the transition is referred to as a *first-order transition*, as in BaTiO_3 and KNbO_3 . See Figure 4-9(a) and (c). In this case, the entropy changes at a constant temperature (e.g., $T = T_c$), and consequently the latent heat also changes. If the change of the polarization is continuous, the transition is referred to as a *second-order transition*, as in KH_2PO_4 and Rochelle salt. See Figure 4-9(b) and (d). In this case, there is no change in entropy and latent heat at the transition temperature. It is interesting to note that at the Curie temperature, the transition in ferro-magnetic materials does not involve changes in crystal structure but only a small change in the

coupling forces between the outer electrons of neighboring magnetic ions.

Since the dielectric constant of ferroelectric materials is extremely high near the transition temperature, the polarization induced in the paraelectric (nonpolar) region at $T > T_c$ by an applied electric field along the ferroelectric axis goes gradually over into the spontaneous polarization region upon cooling below T_c . The effect of this field tends to shift T_c to a higher temperature, as shown in Figure 4-9. The polarization-temperature curve shifts to a higher temperature as the applied field is increased. However, the concept of the Curie temperature cannot be applied to ferroelectric materials under the electrically biased condition.^{18,20} In general, if the direction of the applied field is parallel to the polar direction of any phase, it favors the existence of that phase and extends its temperature range. At T_c , the ferroelectric phase is favored by an applied field; therefore, the Curie temperature is raised.

A change in crystal structure and crystal volume caused by any external forces, such as electric fields, mechanical stresses, or hydrostatic pressures, will affect the phase transition. A shear stress is essentially equivalent to an electric field along the ferroelectric axis and

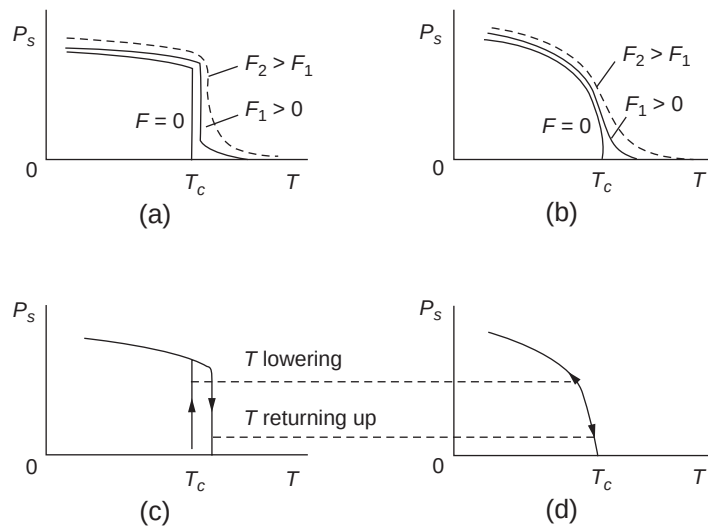


Figure 4-9 Schematic diagrams illustrating the effect of applied electric field on (a) first order transition and (b) second order transition, and the shift of the transition point when the temperature is returning up (increasing), (c) T_c shifts to a higher temperature for the first order transition, and (d) there is no shift for the second order transition.

therefore would have a similar effect on the polarization–temperature curve, as shown in Figure 4-9. It should be noted that in Equation 4-1 the Curie temperature T_c coincides with the actual transition temperature at which spontaneous polarization begins to rise if the transition is of the second order. For first-order transitions, T_c is generally slightly lower than the actual transition temperature by several degrees (typically ~ 10 K). Because the difference is very small, and for general uses, T_c is usually considered the actual transition temperature.

When the temperature decreases through T_c , the polarization rises from zero continuously over several degrees of temperature if the transition is of the second order. However, when the polarization rises abruptly from zero at T_c following the first-order transition, then the rise involves a change of the latent heat, implying that the process is equivalent to supercooling. When the temperature is reversed—that is, increases through the transition point—the process becomes superheating, and a collective or cooperative action among the dipoles and surrounding ions tends to hold the spontaneous polarization until a slightly higher temperature is reached, as shown in Figure 4-9(c). A similar phenomenon occurs at any transition point if the transition is of the first order. (See Figures 4-11 and 4-12.) But this phenomenon does not occur if the transition is of the second order, as shown in Figure 4-9(d).

4.2.2 Phenomenological Properties and Mechanisms

This section discusses the phenomenological behaviors and their mechanisms of six typical ferroelectric materials:

- Barium titanate (BaTiO_3)—type ferroelectrics
- Potassium dihydrogen phosphate (abbreviated KDP, KH_2PO_4)—type ferroelectrics
- Potassium-sodium tartrate tetrahydrate (Rochelle salt, $\text{KNaC}_4\text{H}_4\text{O}_6 \cdot 4\text{H}_2\text{O}$)—type ferroelectrics
- Triglycine sulphate (abbreviated TGS, $(\text{NH}_2\text{CH}_2\text{COOH})_3\text{H}_2\text{SO}_4$)—type ferroelectrics

- Alloys of lead, zirconium, and titanium oxide (abbreviated PZT, alloys of PbO , ZrO_2 , and TiO_2)—type ferroelectrics
- Polyvinylidene fluoride (abbreviated PVDF, $-(\text{CH}_2-\text{CF}_2)_n$ —type ferroelectrics.

BaTiO₃-Type Ferroelectrics

BaTiO_3 belongs to the family of ABO_3 Perovskite mineral (CaTiO_3) structures, in which A and B are metals. The total charge of the A and B positive ions must be $+6$, and A and B must be of quite different sizes; the smaller ion, with a larger charge, must be a transition metal. For BaTiO_3 , Ti is a 3d transition element and has the d orbital for electrons to form covalent bonds with its neighbors. The radius of Ti^{4+} ion is about $0.68 \text{ \AA}$, and that of Ba^{2+} is about $1.35 \text{ \AA}$. These ions form nice octahedral cages, with the O^{2-} ions held apart. At temperatures higher than the Curie temperature ($>120^\circ\text{C}$), Ti^{4+} stays in the cage, rattling around it to make the unit cell maintain a symmetrical cubic structure, as shown in Figure 4-10(a).

However, the structure of the unit cell is temperature dependent. At a certain transition temperature, the particular structure of the unit cell becomes unstable and must transform to a more stable one. So, at the Curie temperature T_c , the octahedral cages distort and the positive ions move to off-center positions. The crystal takes a tetragonal form, resulting from the stretching of the cubic unit cells along one edge, as shown in Figure 4-10(b). In fact, the Ba^{2+} ions shift $0.05 \text{ \AA}$ upward from their original position in the cubic structure; Ti^{4+} ions shift upward by $0.1 \text{ \AA}$, and the O^{2-} ions downward by $0.04 \text{ \AA}$ to form the tetragonal structure. As a result of the ion shifts, the centroid of the positive charges no longer coincides with the centroid of the negative charges; therefore, the unit cells become permanently polarized and behave as permanent dipoles, leading to spontaneous polarization.

The direction of the displacement can be reversed by a sufficiently high electric field of opposite polarity. This possibility of dipole reversal distinguishes ferroelectric materials from nonferroelectric ones. At temperatures

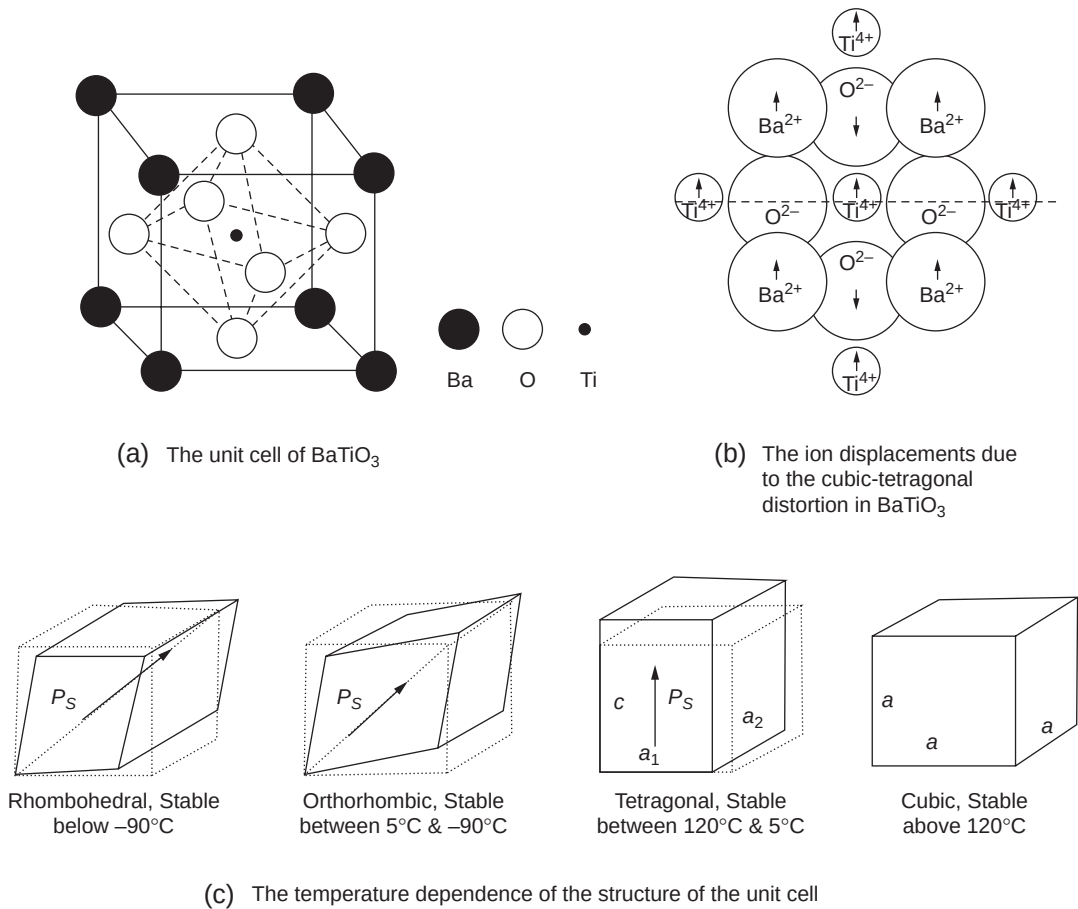


Figure 4-10 Schematic diagrams showing (a) the unit cell of BaTiO_3 , (b) the ion displacement due to the cubic-tetragonal distortion in BaTiO_3 , and (c) the temperature dependence of the structure of the unit cell.

below T_c and between 120°C and 5°C , the structure is tetragonal and the polar axis (i.e., the direction of the spontaneous polarization) is along the c -axis of the unit cells, in which $c > a$. At about 5°C , the tetragonal unit cells undergo a transition for a higher stability to the orthorhombic structure, which is formed by stretching the cell along the face-diagonal direction with the polar axis also along the same direction. This structure will remain stable for temperatures between 5°C and -90°C . Similarly, at around -90°C , a rhombohedral structure, formed by stretching the unit cell along the body-diagonal direction, becomes preferred, as shown in Figure 4-10(c). The direction of spontaneous polarization is

always along the direction of the unit cell's elongation, that is, the stretching direction. This is also referred to as the *ferroelectric polar axis*.

BaTiO_3 -type ferroelectrics are nonpiezoelectric in the unpolarized state. The cubic symmetry implies that, depending on the temperature, spontaneous polarization may occur along several axes. For example, there is a set of three (100) axes, a set of six (110) axes, and a set of four (111) axes. The three sets are non-equivalent.

Dielectric Constants

Dielectric constants are usually measured with small AC signals. At low frequencies,

piezoelectric deformation can follow the change of the applied electric field, and the crystal can be considered mechanically unconstrained. In this case, the dielectric constant measured is under the *free* crystal condition. However, at frequencies higher than the piezoelectric resonance frequency of the crystal, piezoelectric deformation cannot follow the change of the applied AC field due to inertia effects, and the dielectric constant measured is under the *clamped* crystal condition. In nonpiezoelectric crystals, such as BaTiO_3 (which is nonpiezoelectric in the unpolarized state), the dielectric constant is practically independent of frequency at temperatures near or higher than T_c . But in piezoelectric crystals, such as KDP and Rochelle salt, the dielectric constant is strongly frequency-dependent and may differ by several orders of magnitude, depending on the frequency used for the measurement.

Obviously, the frequency dependence of ϵ_r of BaTiO_3 is expected to be different in the various phases. In the tetragonal phase, ϵ_r , measured at low frequencies under the free crystal condition, is expected to be higher than that measured at frequencies higher than the piezoelectric resonance frequency under the clamped crystal condition. At still higher frequencies, an additional drop in ϵ_r may occur, possibly due to dielectric relaxation. In general, the frequency range used for measurements of ϵ_r is from zero to microwave frequencies in which the crystal can still be considered a free crystal. The dielectric constant–temperature (ϵ_r – T) characteristics for BaTiO_3 under the free crystal condition are shown in Figure 4-11. The data are from Merz.^{17,18} The special features of this ϵ_r – T relationship are discussed as follows.

For temperatures above the Curie temperature T_c , the crystal assumes a cubic structure and is in the paraelectric unpolarized state. The dielectric constant decreases with increasing temperature, following closely the Curie–Weiss relation given by Equation 4-1. X-ray analysis has indicated that the dipole moment is associated with the distortion of the molecular system and the cooperative alignment of the dipoles, resulting in spontaneous polarization. BaTiO_3 is

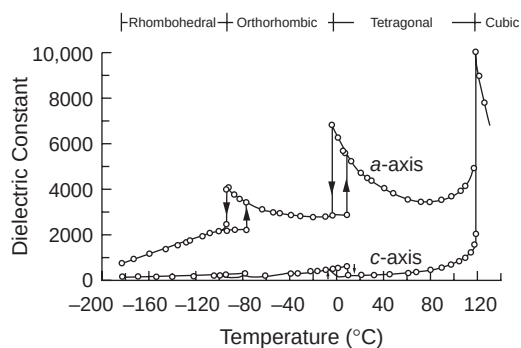


Figure 4-11 Dielectric constant of BaTiO_3 as a function of temperature measured along the *c*-axis and along the *a*-axis (perpendicular to the *c*-axis).

a multiaxial ferroelectric, and its basic molecular structure, as shown in Figure 4-10(a), is an octahedral arrangement of oxygen ions around a central Ti^{4+} ion. Above the Curie point, the Ti^{4+} ion is rattling around from its symmetrical position. Each shift will create a dipole moment along that shifting direction, but this ion can also shift in the opposite direction, due mainly to thermal agitation, which is randomized.

However, when a small AC signal field is applied to the crystal along one axis for the measurement of ϵ_r , the ions will be inclined to shift along the direction of the applied field, thus producing net dipole moments, provided that the driving force from the field is larger than the thermal agitation force. This can happen only at a temperature close to, but above, the Curie point. This shift obviously depends on the local field and the temperature the ions experience. The local field, derived on the basis of Lorentz's internal field, is for nonpolar materials (see Internal Fields in Chapter 2). At $T > T_c$, however, BaTiO_3 would become polar, due to the shifting of the ions under an applied field. This mechanism is different from atomic polarization in ionic crystals and also from the orientational polarization in dipolar materials in which permanent dipoles are present. So far, there is no satisfactory method available to calculate the local field and the average dipole moment in ferroelectric materials for $T > T_c$. By assuming that the ionic shifting in producing dipole moments is equivalent

to the dipole orientation (although in fact, they are quite different), we can borrow Debye's formula originally derived for dipolar materials. Then, the average dipole moment per molecule in the direction of the applied field can be written as

$$\langle \mu_F \rangle = \frac{\mu_o^2}{3kT} F_{\text{local}} \quad (4-2)$$

where μ_o is the dipole moment of the permanent dipole in the material and F_{local} is the local field. Again, we borrow the formula for local field based on Lorentz's internal field, which is given by

$$F_{\text{local}} = \frac{\epsilon_r + 2}{3} F \quad (4-3)$$

Since the polarization is given by

$$P = (\epsilon_r - 1)\epsilon_o F = N\langle \mu_F \rangle \quad (4-4)$$

where N is the number of molecules per unit volume, then from Equations 4-2 through 4-4 we obtain the famous Clausius–Mossotti equation:

$$\frac{\epsilon_r - 1}{\epsilon_r + 2} = \frac{N\mu_o^2}{9\epsilon_o kT} \quad (4-5)$$

Rearrangement of Equation 4-5 yields the dielectric susceptibility

$$\chi = \epsilon_r - 1 = \frac{3N(\mu_o)^2/9\epsilon_o k}{T - N(\mu_o)^2/9\epsilon_o k} = \frac{3T_d}{T - T_d} \approx \epsilon_r \quad (4-6)$$

because $\epsilon_r \gg 1$. This equation is similar to Equation 4-1. This indicates that ϵ_r diverges as T approaches T_c , and that the system becomes unstable and must make a transition to a new phase. The divergence of ϵ_r at $T = T_c$, based on Equation 4-6, is generally referred to as the *Clausius–Mossotti* or *polarization catastrophe*. It is important to note that $T_d = N(\mu_o)^2/9\epsilon_o k$ is definitely not T_c and cannot be used to evaluate T_c , nor can $3T_d$ be used to estimate the Curie constant C for ferroelectric materials, simply because for $T > T_c$, there are no permanent dipoles as such in BaTiO_3 . It appears that research in the derivation of the Curie–Weiss equation based on the ion-shifting mechanism (which is similar to ionic polarization but not quite the same and has a different form of local

field) could be very rewarding. A thermodynamic approach to this phenomenon will be discussed in Section 4.2.3.

At T_c , the dielectric constant for BaTiO_3 is typically of the order of 10,000, and the ions are moving into a difference phase, corresponding to spontaneous polarization. Consequently, an applied electric field will be able to produce relatively large shifts. This action results in a big change in dipole moment, corresponding to a high dielectric constant. When T is lowered just below T_c , the structure of the unit cell changes from the cubic form to a tetragonal one, as shown in Figure 4-10(b) and (c). Once the structure begins to change or spontaneous polarization to arise, the net dipole moment in the direction of the applied AC field, and hence the dielectric constant, will start to decrease, as shown in Figures 4-11 and 4-12.

At $T < T_c$, the dielectric constant is strongly anisotropic in the ferroelectric phase for all crystals. For multiaxial ferroelectrics such as BaTiO_3 , the dielectric behavior, which is isotropic in the cubic nonpolarized phase, becomes strongly anisotropic in the polarized phase. The dielectric constant measured along the principal ferroelectric axis (c -axis in the tetragonal structure) is lower than that normal to it (i.e., a -axes). This big difference can be observed only in single-domain crystals.^{17,18} This is readily understandable because the spontaneous polarization along the c -axis is saturated, while the polarization along the a -axes is not. For uniaxial ferroelectrics, such as KDP

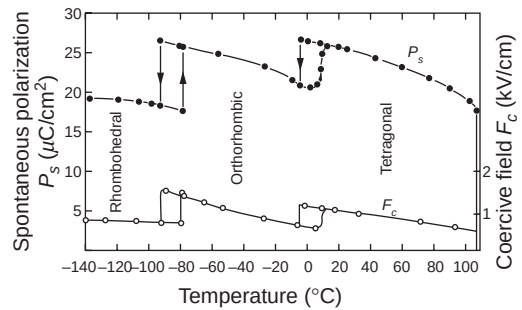


Figure 4-12 Spontaneous polarization and coercive field of a BaTiO_3 single crystal as functions of temperature measured along the tetragonal [001] direction (c -axis).

and Rochelle salt, the dielectric constant measured along the principal ferroelectric axis is much larger than that normal to it, simply because the ions shift only along one principal ferroelectric axis and do not shift in any direction other than the principal axis. For example, for KDP, the dielectric constant along the principal axis is about 1100, while that normal to it is only 12. For Rochelle salt, the corresponding dielectric constants along and normal to the principal axis are 1000 and 10, respectively.

In general, pure crystal specimens have a lower T_c than the impure ones. Also, the electrical conductivity of a material has a screening effect on the cooperative force that leads to the ferroelectric behavior. Other ferroelectric materials, which are similar to BaTiO_3 in their chemical structure (PbTiO_3 , KNbO_3 , KTaO_3 , etc.), have similar dielectric and structural properties to those of BaTiO_3 , except that their Curie temperatures and Curie constants are different.

Spontaneous Polarization and Coercive Field

The temperature dependence of the spontaneous polarization P_s and the coercive field F_c for BaTiO_3 single crystals is shown in Figure 4-12. The data are from Merz.^{17,18} The measurements were made along the c -axis of the tetragonal structure, that is, along the (001) direction.^{18,21} The P_s value along the (110) face-diagonal direction of the orthorhombic structure can be obtained by multiplying the corresponding value along the c -axis by $\sqrt{2}$. Similarly, the P_s value along the (111) body-diagonal direction of the rhombohedral structure can be obtained by multiplying the corresponding value along the c -axis by $\sqrt{3}$. The transition at T_c is of the first order for single, pure BaTiO_3 crystals. However, if the crystals are not pure, the expected first-order transition may become a second-order transition, and the value of P_s would also be lowered, due to the hindering of the motion of domain walls by impurities or other defects.

The coercive field F_c is temperature dependent, as shown in Figure 4-12. F_c depends also on the time required for the polarization to reverse its direction. According to Merz²² and

Landauer,²³ the rate of the reversal of the polarization P can be expressed as

$$\frac{dP}{dt} = f(P)\exp[-\alpha/F(t)] \quad (4-7)$$

where the function $f(P)$ is related to the switching rate. For most hysteresis loop measurements, the applied field varies sinusoidally with time (i.e., $F(t) = F_o \sin \omega t$), and α is a parameter depending on temperature. Equation 4-7 implies that the coercive field depends on the rise rate of the applied field. For slow rates, the crystal can have enough time to reverse its polarization so that the reversal can be completed before the peak field is reached. For fast rates, however, the crystal may not have sufficient time to complete the reversal process. In this case, a larger F_c is needed to bring P_s to zero. In other words, F_c is not a constant physical parameter; it depends on both the value of F_o and the frequency ω of the applied field.^{24,25} The coercive field depends also on the thickness of the ferroelectric specimens. The thicker the specimen, the smaller the coercive field. The experimental results follow the thickness-dependence relation²⁶

$$F_c = F_{co} + G/d \quad (4-8)$$

where F_{co} is the coercive field for very thick specimens, d is the specimen thickness, and G is a constant. Based on the experimental results for BaTiO_3 , G turns out to be 2.2 V. Polarization reversal is a process of nucleation and growth of antiparallel domains, which is thickness dependent.^{22,27} The coercive field can be considered the field at which the reversal process can proceed easily.

Before ending this section, it is worth mentioning that upon repeated switching of the polarization of the crystals, notably BaTiO_3 , spontaneous polarization becomes clamped or decreases, and the coercive field increases. This phenomenon is referred to as the *fatigue* or *decay* of the crystal,^{28,29} and it has been attributed to the build-up of the space charge near the crystal surfaces and the interaction of this space charge with domain walls. The degree of the fatigue depends on several factors including the ambient atmosphere, the elec-

trode material, and the waveform of the applied switching field. The decay is more rapid with pulsed fields than with low-frequency AC fields.^{28–30} Ferroelectric ceramics exhibit also fatigue behavior. Apart from the factors previously mentioned, pores, impurities, and micro-cracking may also play important roles in this phenomenon.

The displacement current observed during the reversal of spontaneous polarization is often accompanied by transient pulses. Such noise-like pulses were first observed in ferromagnetic materials by Barkhausen in 1919, and the analog of this phenomenon in ferroelectric materials still bears his name. The processes of domain nucleating or coalescing produce sharp discontinuities in the change of polarization. These are picked up from the electrode as Barkhausen pulses. This phenomenon has been studied by many investigators.^{31–33} In BaTiO₃, there are typically 10⁵ to 10⁶ pulses during a complete polarization reversal, and the average pulse height is of the order of 10^{–12} coulomb. The total number of pulses of all sizes does not vary with the applied field, but pulse heights increase with increasing field. Similar features of this phenomenon appear in other ferroelectric materials, including TGS, Rochelle salt, etc. Experiments have revealed that Barkhausen pulses are associated with the nucleation and growth of domains. However, this phenomenon provides an additional avenue for the study of polarization reversal processes under various conditions.

Electrical Conductivity and Breakdown Strength

In oxide crystals, bonds are formed by the complete transfer of the electrons from the metal atoms to the nonmetal atoms, such as oxygen atoms. Each atom then becomes an ion with an inert gas atomic structure. These ions are so stable that it is very difficult to remove electrons from them or create holes in them. This implies that their energy band gaps are large. The energy band gap for the Perovskite oxides, such as BaTiO₃, is around 3 to 4 eV. Even assuming electron mobility of 100 cm²V^{–1}s^{–1},

the intrinsic conductivity at room temperature is thousands of millions of times smaller than the real electrical conductivity of BaTiO₃, which is of the order of 10^{–14} (ohm-cm)^{–1} and of the n-type. So, the contribution from the intrinsic condition is negligible and the electronic conductivity must be extrinsic (see Electronic Conduction in Chapter 7).

For this finite n-type electronic conductivity, we must find the source of the electrons. The mechanism of electrical conduction for BaTiO₃ is similar to that for titanium oxide TiO₂ (rutile) or other oxides, such as ZnO, etc. It is almost impossible to obtain stoichiometric BaTiO₃. Crystal specimen fabrication processes are never perfect. There is always a possibility for the four-valent Ti⁴⁺ ions to reduce to the three-valent Ti³⁺ or even to the two-valent Ti²⁺ ions. Each titanium atom can give four electrons to oxygen atoms, and each oxygen atom can take only two electrons. If the oxide is oxygen-deficient (or metal-rich), then not all the titanium atoms have shed four electrons. Some must be left in a lower valency state of Ta³⁺ or Ta²⁺ in order to maintain an overall neutrality of the system. This situation results in a few Ta³⁺ or Ta²⁺ in a sea of Ta⁴⁺, and a few of oxygen vacancies in the lattice of O^{2–} ions. The Ta³⁺ or Ta²⁺ ions act like electron donors, ready to provide electrons for electronic conduction, while O^{2–} ions may also diffuse via the vacancies for ionic conduction. So, the electrical conductivity consists of electronic conductivity σ_n and ion conductivity σ_i . Thus,

$$\sigma = \sigma_n + \sigma_i = q\mu_n n + q\mu_i n_i \quad (4-9)$$

where u_n and n are, respectively, the mobility and the concentration of electrons; and u_i and n_i are, respectively, the mobility and the concentration of vacancies.

The electrons on Ta³⁺ or Ta²⁺ ions may hop to Ta⁴⁺ ions, contributing to an extrinsic n-type conduction. For pure BaTiO₃ single crystals, electronic conduction is dominant for temperatures up to 500°C, but ionic conduction becomes significant at about 100°C. It is interesting to note that the activation energy for electronic conduction in BaTiO₃ is considerably smaller in the polarized state than in the

paraelectric state, possibly due to the big difference in dielectric constant, polarization, and lattice parameters between these two states.

It should be noted that there is also a possibility of the oxide being oxygen-rich (or metal-deficient). In this case, Ta^{4+} can give four electrons to oxygen atoms, but not all oxygen atoms can have two electrons each; some oxygen atoms can get only one each and become O^- ions. The O^- ions act as acceptors, creating holes in the crystals. These holes can migrate from O^- to O^{2-} (equivalent to electrons hopping from O^{2-} to O^-). Cations may also diffuse via vacancies, but the activation energy for cation diffusion in oxide crystals is always high. Thus, the extrinsic p-type conduction is dominant. However, both electronic conduction (n-type or p-type) and ionic conduction occur in $BaTiO_3$; the ratio of $\frac{\sigma_n}{\sigma_i}$ depends strongly on the purity and the fabrication process of the material, as well as the temperature.

The breakdown strength of $BaTiO_3$ single crystals is of the order of 0.5 MV cm^{-1} . For $BaTiO_3$ ceramics, it is about 0.1 MV cm^{-1} . Generally, ceramics have a lower breakdown strength, partly because of their porosity. It should be noted that a high poling field may sometimes cause electrical breakdown, because the poling time is usually quite long (e.g., one hour). High field stressing coupled with long stressing time may be sufficient to overcome the interdomain stresses in $BaTiO_3$. If impurities, pores, and other defects are present in specimens, as in ceramics, the breakdown strength will be strongly dependent on specimen thickness, electrode area, and temperature. For more details about the breakdown processes, see Electrical Breakdown in Chapter 8.

KH_2PO_4 -Type Ferroelectrics

Potassium di-hydrogen phosphate (KH_2PO_4) is generally referred to as KDP crystals, with a tetragonal structure at room temperature. Ferroelectrics of the KH_2PO_4 type exhibit a much simpler behavior than other ferroelectrics because they belong to the uniaxial ferroelectric family. These materials polarize along only one

axis (up or down along the c -axis) and undergo only one phase transition from a paraelectric into a spontaneously polarized state. The transition does not show a thermal hysteresis, implying that the transition is of the second order. In the paraelectric unpolarized state, the crystal assumes a tetragonal structure at T above T_c , with the ferroelectric axis along the c -axis. The crystal is piezoelectric because it belongs to the noncentrosymmetric point group. At T below T_c , an orthorhombic phase exists, which is ferroelectric with spontaneous polarization due mainly to the displacement of K, P, and O ions in the direction of the polar c -axis.³⁴ The structure of KDP is quite complicated. Basically, it consists of two interpenetrating body-centered lattices of PO_4 tetrahedra and two interpenetrating body-centered K lattices, with the PO_4 and K lattices separated along the c -axis. Every PO_4 is linked to four other PO_4 groups by hydrogen bonds, which lie perpendicular to the c -axis. This arrangement gives rise to an ionic configuration of $K^+(H_2PO_4)^-$. The distribution of proton density shifts from the center of the bond at T above T_c to a localized asymmetric position below T_c . The direction of this shift reverses when the spontaneous polarization is reversed.^{34,35}

The dielectric constant–temperature characteristics are shown in Figure 4-13, the data being from Busch,³⁶ and the spontaneous polarization–temperature characteristics in Figure

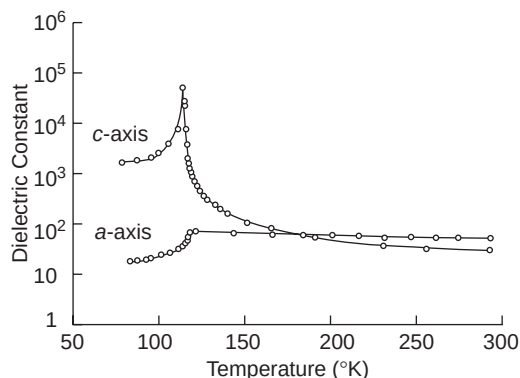


Figure 4-13 Dielectric constants of KH_2PO_4 as a function of temperature measured along the c -axis and along the a -axis.

4-14, the data being from Von Arx and Bantle.³⁷ The Curie temperature T_c is $+150^\circ\text{C}$ (123 K). At $T > T_c$, the dielectric constant decreases with temperature, following closely the Curie–Weiss law with the Curie constant C being 3250°C .^{36,37} At $T < T_c$, the dielectric constant along the c -axis drops sharply (but not abruptly) due to the dielectric saturation effect, which results in a decrease of dielectric constant with increasing spontaneous polarization. The dielectric constant along the a -axes also shows an anomaly at T_c , possibly due to the effect of dielectric cooperative action among the ions.³⁸ The dielectric constant along the c -axis is dependent on the applied DC biasing field. The bias tends to lower the dielectric constant and also to shift the peak of the dielectric constant toward higher temperatures.³⁹

Figure 4-14 shows that the spontaneous polarization begins to rise at Curie point continuously, so the transition is of the second order. The coercive field F_c is temperature dependent. It is interesting to note that at a temperature of about -213°C , (60 K) there is a sharp break. At $120\text{ K} > T > 60\text{ K}$, F_c is practically independent of temperature. But at $T < 60\text{ K}$, F_c increases rapidly with decreasing temperature.⁴⁰ This phenomenon is probably caused by the hindered motion of domain walls at low temperatures.

There are a number of crystals isomorphous with KH_2PO_4 that have a similar dielectric behavior, including KH_2AsO_4 , RbH_2PO_4 , CsH_2PO_4 , etc. In this KDP family, hydrogen bonds play a vital role in the ferroelectric behavior.

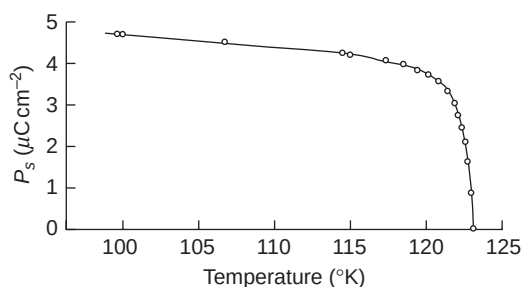


Figure 4-14 Spontaneous polarization of KH_2PO_4 as a function of temperature.

$\text{KNaC}_4\text{H}_4\text{O}_6 \cdot 4\text{H}_2\text{O}$ (Rochelle Salt)–Type Ferroelectrics

Rochelle salt is a potassium-sodium tartrate-tetrahydrate ($\text{KNaC}_4\text{H}_4\text{O}_6 \cdot 4\text{H}_2\text{O}$) and belongs to the family of uniaxial ferroelectrics. This was the first material in which Valasek discovered the ferroelectric phenomenon in 1921, and from which he also introduced the term *Curie temperature* or *Curie point* for ferroelectrics.

It is interesting to describe briefly the ferroelectric properties of this material. Rochelle salt has a rather poor mechanical strength and low disintegration temperature; it is also apt to absorb water. The crystal decomposes at a low temperature of about 55°C . However, the most interesting property of Rochelle salt is that it exhibits two Curie points, one at $+24^\circ\text{C}$, called the *upper Curie temperature*, and the other at -18°C , called the *lower Curie temperature*. The nonpolar paraelectric phases for $T > +24^\circ\text{C}$ and for $T < -18^\circ\text{C}$ are orthorhombic. The polar ferroelectric phase, occurring at the temperature range between -18°C and $+24^\circ\text{C}$, is monoclinic. The dielectric constant, measured along the a -axis with a small signal field of 1 kHz as a function of temperature, is shown in Figure 4-15.

There is no thermal hysteresis at the transition points, indicating that the transition is of the second order. For temperatures from $+24^\circ\text{C}$

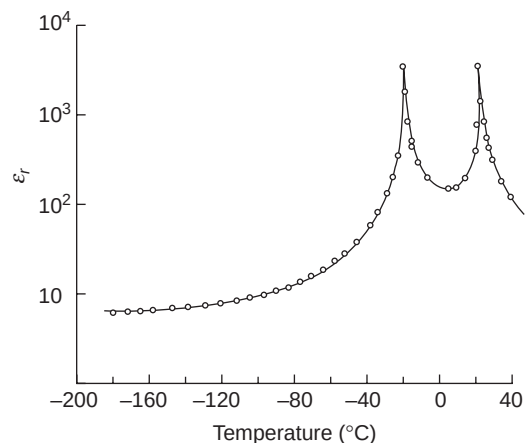


Figure 4-15 Dielectric constant of Rochelle salt as a function of temperature.

to about $+50^{\circ}\text{C}$, the dielectric constant obeys the Curie–Weiss law, with the Curie constant C 2240°C . For the temperature range from -18°C to -28°C , the dielectric constant again follows the Curie–Weiss law, with a Curie constant of 1180°C . The decrease of the dielectric constant in the ferroelectric region is due to the effect of spontaneous polarization saturation, similar to BaTiO_3 and KH_2PO_4 . The temperature dependence of spontaneous polarization is shown in Figure 4-16. The coercive field for Rochelle salt depends on the specimen thickness and the magnitude and the frequency of the applied field, similarly to other ferroelectric materials. The data given in Figures 4-15 and 4-16 are from Hablutzel.⁴¹

Rochelle salt is like KDP; the hydrogen bond plays an important role in the ferroelectric behavior. This idea has led to the study of a deuterated Rochelle salt, prepared by dissolving ordinary Rochelle salt in highly concentrated D_2O to make it thoroughly desiccated at higher temperatures, in order to replace hydrogen with deuterium, the isotope of hydrogen (heavy hydrogen consisting of a neutron and a proton in the nucleus). After this process, the Rochelle salt becomes a deuterated compound with a chemical formula of $\text{KNaC}_4\text{H}_2\text{D}_2\text{O}_6 \cdot 4\text{D}_2\text{O}$. This compound has the Curie points at -22°C and $+35^{\circ}\text{C}$, thus widening the region of

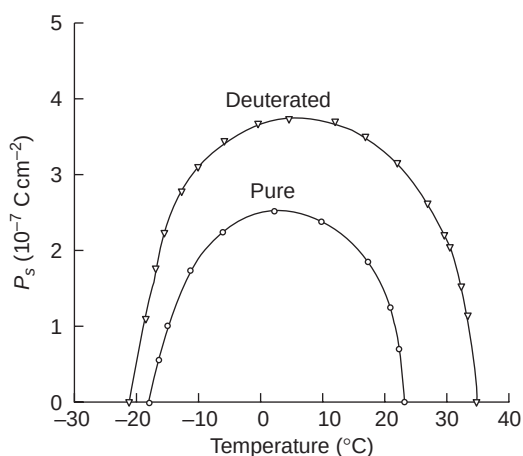


Figure 4-16 Spontaneous polarization of Rochelle salt and deuterated Rochelle salt.

the stable ferroelectric phase and giving a higher spontaneous polarization, as shown in Figure 4-16.^{41,42}

$(\text{NH}_2\text{CH}_2\text{COOH})_3 \cdot \text{H}_2\text{SO}_4$ Type Ferroelectrics

Triglycine sulphate (TGS) is a uniaxial ferroelectric material, and its chemical formula is $(\text{NH}_2\text{CH}_2\text{COOH})_3 \cdot \text{H}_2\text{SO}_4$. This material exhibits ferroelectric properties at room temperature. The Curie temperature is 49°C . At $T > T_c$, the crystal structure is centrosymmetric and in the nonpolar paraelectric state with a monoclinic symmetry. At $T < T_c$, the crystal becomes polar and belongs to the polar group of the monoclinic system.³⁹ Ferroelectric direction is along the b -axis and the transition is of the second order. The dielectric constant–temperature characteristics measured at 1 kHz and 1 V cm^{-1} are shown in Figure 4-17. The dielectric constant along the polar b -axis shows a prominent anomaly in the vicinity of T_c . The dielectric constants along the a - and c -axes are practically independent of temperature, simply because this is a uniaxial polarized material. For $T > T_c$, ϵ_r follows the Curie–Weiss law with a Curie constant C of 3200°C . The temperature dependence of the spontaneous polarization is shown in Figure 4-18. The experimental data are from Hoshino et al.⁴³ Some crystals are isomorphous with TGS, such as triglycine selenate $(\text{NH}_2\text{CH}_2\text{COOH})_3 \cdot \text{H}_2\text{SeO}_4$

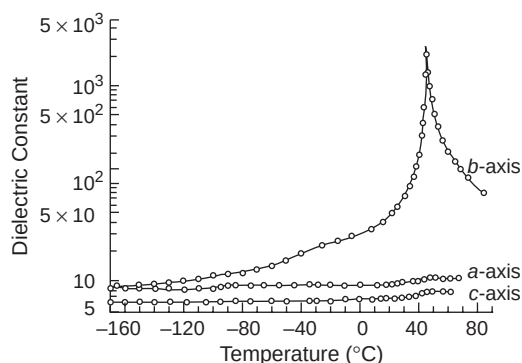


Figure 4-17 Dielectric constant of triglycine sulfate as a function of temperature measured along a -axis, b -axis, and c -axis.

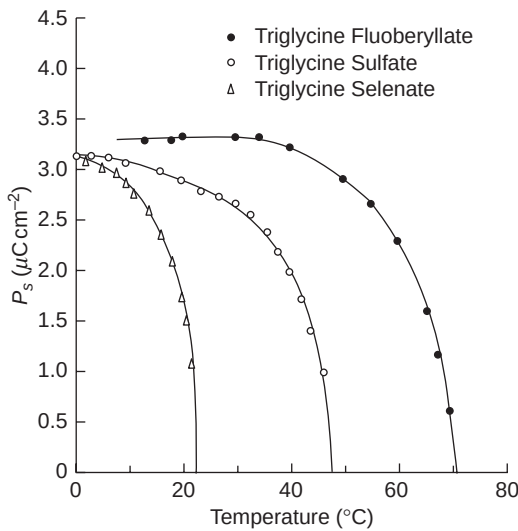


Figure 4-18 Spontaneous polarization of triglycine sulfate, triglycine selenate, and triglycine fluoberyllate as a function of temperature.

(abbreviated TGSe) and triglycine fluoberyllate ($(\text{NH}_2\text{CH}_2\text{COOH})_3\text{H}_2\text{BeF}_4$ abbreviated TGFB), have similar properties to TGS with different Curie temperatures, as shown in Figure 4-18.

Alloys of PbO , ZrO_2 , and TiO_2 (PZT alloys)-Type Ferroelectric Ceramics

PZT alloys belong to the family of ceramics, which are alloys of lead oxide, zirconium oxide, and titanium oxide (PbO , ZrO_2 , and TiO_2). The familiar Perovskite structure of BaTiO_3 also appears in PZT alloys because there are equal numbers of divalent (Pb^{2+}) and tetravalent (Zr^{4+} and Ti^{4+}) cations. PZT alloys are in fact the alloys of two components: PbZrO_3 and PbTiO_3 . This is why PZT alloys are similar to BaTiO_3 in their ferroelectric properties. Thus, by controlling the composition and microstructure of the PZT alloys, it is possible to tailor their properties to suit particular applications. The general chemical formula is $\text{PbZr}_{1-x}\text{Ti}_x\text{O}_3$, with x as a fraction from 0 to 1. The phase diagram of ferroelectric ceramics is normally plotted in phase transition temperature as a function of the composition. A typical equilibrium phase diagram for the PZT system

is shown in Figure 4-19.⁴⁴ The Curie temperature T_c depends on the composition of the system.

In this phase diagram, there are six major phase boundaries:

1. The cubic paraelectric–tetragonal ferroelectric (P – F_T) phase boundary
2. The cubic paraelectric–rhombohedral ferroelectric (HT) (P – $F_{R(HT)}$) phase boundary
3. The tetragonal ferroelectric–rhombohedral ferroelectric (HT) (F_T – $F_{R(HT)}$) phase boundary
4. The rhombohedral ferroelectric (HT)–rhombohedral ferroelectric (LT) ($F_{R(HT)}$ – $F_{R(LT)}$) phase boundary
5. The rhombohedral ferroelectric (HT)–orthorhombic antiferroelectric ($F_{R(HT)}$ – AF_o) phase boundary
6. The rhombohedral ferroelectric (LT)–orthorhombic antiferroelectric ($F_{R(LT)}$ – AF_o) phase boundary

The phase is cubic paraelectric, implying that an applied field can distort the system and change it from unpolarized to a polarized state, but as soon as the field is removed, the system will quickly relax back to its original unpolarized state due to thermal agitation. Figure 4-19 has some prominent features worth mentioning.

- The Curie temperature T_c for all compositions is higher than that for BaTiO_3 . The main reason is that the Pb^{2+} ion has two outer electrons beyond its last full shell, but the Ba^{2+} ion does not. These outer electrons can contribute to covalent bonding with neighboring oxygen ions. With this extra bonding, a greater thermal energy or a higher temperature is required to convert the polarized state into an unpolarized state. This results in a higher T_c for the PZT system.
- The morphological boundary between the tetragonal and the rhombohedral phases occurs at x around 0.5. Since the PZT system is used mostly in piezoelectric devices, the PZT system with the composition near the morphological boundary has an advantageous crystal morphology. The efficiency

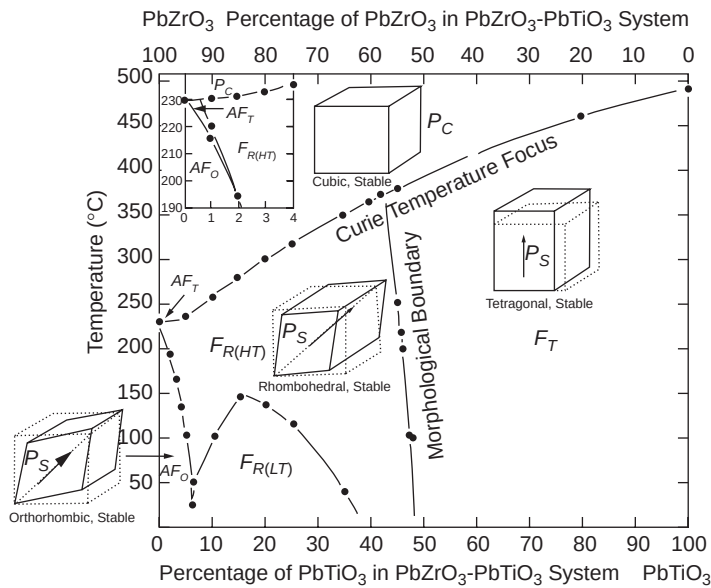


Figure 4-19 The equilibrium phase diagram of the PZT system and microstructure of the unit cell for different phases.

of the conversion between electrical and mechanical energy depends strongly on crystal morphology. The piezoelectric coupling coefficient climbs to a maximum value at x around 0.5 because the shear transformation between the two types of structures is much easier. This is because there are six easy directions of polarization for the tetragonal unit cell and eight easy directions of polarization for the rhombohedral unit cell, which provide much easier mutual conformation.

- For most PZT alloys, the rich Zr region with x below 0.1 is attractive because they have superior properties. In this region, the Curie temperature T_c is lower. At $T < 200^\circ\text{C}$, there exists a boundary between the ferroelectric phase and the antiferroelectric phase.
- The rhombohedral ferroelectric phase is subdivided into high temperature $F_{R(HT)}$ and low temperature $F_{R(LT)}$ phases. Simple rhombohedral unit cells appear in both high-temperature and low-temperature regions. The basic difference between $F_{R(HT)}$ and $F_{R(LT)}$ is possibly due to the difference in the directions of the spontaneous polarization.^{45,46}

- The Curie temperature in pure PbZrO_3 is 230°C . Below this temperature, the crystal is in an antiferroelectric state, with the structure changing to orthorhombic one. But at or near the Curie temperature, the antiferroelectric phase assumes a tetragonal structure. The mechanism that causes this difference in structure at or near T_c is not certain; possibly it is due to the presence of foreign impurities.⁴⁷

It is interesting to review the ferroelectric properties of lead zirconate PbZrO_3 in the PZT system with $x = 0$. The dielectric constant-temperature relation for PbZrO_3 ceramic is shown in Figure 4-20. The dielectric constant reaches a peak at the Curie temperature 230°C . Above T_c , the structure is cubic and ϵ_r follows the Curie-Weiss law, with the Curie constant C equal to 1.6×10^5 degrees. The results are from Roberts.⁴⁸ Although the dielectric anomaly is similar to that in BaTiO_3 , no hysteresis loop has been observed for $T < T_c$, implying that there is no net spontaneous polarization. This indicates that the phase for $T < T_c$ is antiferroelectric.⁴⁹ In other words, all adjacent domains are oppositely polarized, so there is electrical ordering

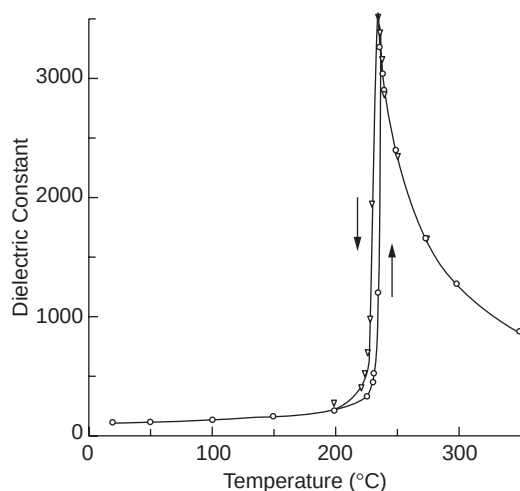


Figure 4-20 Dielectric constant of ceramic PbZrO_3 as a function of temperature.

but no net polarization (see Sections 4.2.3 and 4.2.5).

However, in the vicinity of the transition temperature, a strong electric field can induce a ferroelectric phase. This means that the field can shift the phase boundary between AF_o and F_R , with the tendency of extending the ferroelectric region. A double hysteresis loop for pure PbZrO_3 has been observed at temperatures below T_c .⁵⁰ The ferroelectricity is induced in the antiferroelectric phase when the applied electric field reaches a critical value. A similar double hysteresis loop occurs in BaTiO_3 , but in that case the ferroelectricity is induced in the cubic paraelectric unpolarized region by a sufficiently high electric field and at a temperature slightly above T_c . Obviously, the origins of this double-loop behavior for these two cases are different. In PbZrO_3 , the ferroelectric phase induced by the field has rhombohedral symmetry. The critical field is temperature dependent, as shown in Figure 4-21. The data are from Sawaguchi and Kittaka.⁵⁰

It is difficult, if not impossible, to produce a PZT alloy $\text{PbZr}_{1-x}\text{Ti}_x\text{O}_3$ with $0 < x < 1$ in single crystal form. The PZT alloys are ceramics, which are a solid solution. Generally, the three oxides— PbO , ZrO_2 , and TiO_2 —of selected purity and proportion, are ball-milled together

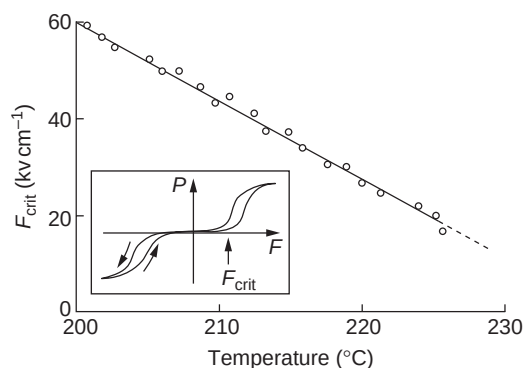


Figure 4-21 The critical electric field F_{crit} required to induce ferroelectricity in PbZrO_3 as a function of temperature. Insert: the occurrence of a double hysteresis loop at the critical field.

to produce a powdered mixture of very fine particle size. The chemical substance is then formed by calcining (fiercely heating) this mixture. This material is then ground up and artifacts made by compacting and sintering.

Ceramics are quite different from single crystals. Ceramics are formed by many randomly arranged crystallites (grains) with unavoidable microvoids and impurities trapped in between. In fact, they are imperfect polycrystalline materials. So a polycrystalline ceramic that has not been subjected to static field stressing would behave like a nonpolar material, even though the crystallites composing it are polar. However, ferroelectric ceramics can be transformed into polar materials by an externally applied static electric field. This process is referred to as the *poling process*, which is reversible. To pole a specimen, the specimen must be brought to the saturation of the spontaneous polarization by a strong field and then allowed to relax back to the remanent condition. Poling is usually carried out at temperatures as hot as the Curie threshold temperature allows, and this poling condition is held for some hours. However, the Curie temperature is much lower than the ceramic sintering temperature, at which decent rates of transport between grains can occur, so it is inevitable that poled ceramics always carry some residual internal stress. A poled ceramic can be depoled

by the application of electric fields or mechanical stresses.

A ceramic system consists of randomly originated grains (crystallites) intimately bonded together with grain sizes generally ranging from $1\mu\text{m}$ to $20\mu\text{m}$. There are always pores (microvoids) and foreign impurities unavoidably present in the material during the fabrication process. It is obvious that both porosity and grain size affect dielectric and ferroelectric properties. The important factor leading to the effect is the variation in the dimensions of the grains and hence in the internal stresses during the change of temperature through the Curie point. The random directions of the crystallographic axes of the crystallites limit the extent to which spontaneous polarization can be developed. Due to the inhibition of the motion of domain walls by internal strains and other defects (such as impurities), the saturation value of the spontaneous polarization of BaTiO_3 in ceramic structure is about half of the value for BaTiO_3 single crystals. The hysteresis loop of BaTiO_3 single crystals with a single domain is a square loop, like the one shown in Figure 4-5(a), having a spontaneous polarization P_s of about $26\mu\text{C cm}^{-2}$ and a coercive field F_c of about 1 kV cm^{-1} . The hysteresis loop for BaTiO_3 ceramics is like a round loop, like the one shown in Figure 4-5(b), having P_s of about $7\mu\text{C cm}^{-2}$ and F_c of about 4 kV cm^{-1} . Of course, the latter figures depend to a great extent on the preparation of the ceramic specimens.

One unusual feature of ferroelectrics is the aging phenomenon. A real crystal is never an ideal dielectric. It always contains imperfections, including vacancies, dislocations, impurities, etc., which disturb the uniformity of the polarization and depolarization. A ceramic material contains even more imperfections. The properties of ceramics change with time, even in the absence of external electric fields, mechanical stresses, or temperature changes. This aging phenomenon is associated with the continuous changing of the internal structure, which with time tends to reach a more stable arrangement—in other words, to lower the internal free energy. In fact, any poled ferroelectric ceramic is in a state of incessant depo-

larization with time. However, the magnitude of the poling field and the poling temperature are important factors in determining the extent of the dipole alignment and hence the resulting ferroelectric properties. In actuality, the alignment is never complete. Depending on the type of the crystal structure and the poling condition, the thoroughness of poling can be quite high, approaching 83% (for the tetragonal structure) and 91% (for the orthorhombic structure) of that for single crystals.⁵¹ The aging process may be accelerated by exposing the poled ceramics to

- Strong depoling fields
- High mechanical stresses
- Temperatures close to the Curie point

A common rule of thumb to reduce the rate of depolarization (aging) is that the poled ferroelectric materials should not be exposed to temperatures higher than one half of the Curie temperature, since heat is the major factor that causes depolarization. However, aging is a complicated process that may be caused by several different mechanisms, such as

- Gradual equilibration of domains toward the minimization of elastic and dielectric free energy
- Segregation of impurities and vacancies on domain walls and grain boundaries
- Ordering of impurities and vacancies inside the ferroelectric domains with respect to the polar axis⁵²

Note that PZT ceramics are usually used with a dopant, modifier, or other composition in solid solution in order to improve or modify the properties of the basic PZT system for specific applications. A detailed discussion of many different compositional systems is beyond the scope of this chapter. However, a few significant examples will demonstrate the importance of additives in extending the variety of the applications of the PZT system.

For the case of oxygen-rich PZT, the conduction is mainly p-type extrinsic due to the holes produced by O^- ions. To reduce the p-type conduction, we can use donor dopants, such as

Nb^{5+} or La^{3+} , replacing Pb^{2+} in order to provide electrons required to convert O^- ions to O^{2-} , thus reducing the conductivity. It is important to reduce the conductivity because high conductivity makes it difficult to apply a high field for poling for a long period of time. The additives are usually compensated by *A*-site vacancies. (The Perovskite structure is expressed in the general formula ABO_3 . For BaTiO_3 , the vacancies created by missing Ba^{2+} are referred to as the *A-site vacancies*. Similarly, the vacancies created by missing Ti^{4+} are *B-site vacancies*.) In this case, these additives will enhance domain reorientation and hence increase spontaneous polarization, dielectric constant, and piezoelectric constant; they will also lower the coercivity and reduce the aging effect.⁵³

One important compositional system, generally referred to as the *PLZT system*, is the PZT system modified by the lanthanum. This system can be formed by the complete miscibility of PbO , ZrO_2 , TiO_2 , and La_2O_3 in a solid solution; the amount of *La* can be adjusted on the basis of the general chemical formula $\text{Pb}_{1-y}\text{La}_y(\text{Zr}_{1-x}\text{Ti}_x)_{1-y/4}\text{O}_3$ with the fraction *y* ranging from 0 to 1. Usually, the PLZT system is formulated so that La^{3+} ions replace mainly the *A*-site vacancies created by missing Pb^{2+} ions. In this case, the chemical formula can be expressed as $\text{Pb}_{1-3y/2}\text{La}_y(\text{Zr}_{1-x}\text{Ti}_x)\text{O}_3$. Selected compositions within this system offer high electromechanical coupling coefficients for piezoelectric transducers.⁵⁴⁻⁵⁶ The PLZT system has also excellent electro-optic properties.^{57,59} Depending on its composition, the PLZT system exhibits one of three major types of electro-optic characteristics: *memory*, *linear*, or *quadratic*. Each of these three types can be used for different electro-optic devices.⁶⁰

Another example is the Nb^{5+} -doped PZT system. It is well known that for the $\text{Pb}(\text{Zr}_{1-x}\text{Ti}_x)\text{O}_3$ system, a *Zr*-rich system (i.e., small *x*)—for example with *x* = 0.05—exhibits an extensive phase boundary between the ferroelectric F_r and the antiferroelectric AF_o phases, as shown in Figure 4-19. However, for *x* < 0.05, it becomes difficult to sinter to obtain a high-density ceramic, compared to common PZT with higher *x* values. It is found that Nb^{5+}

dopants can improve this situation.^{61,62} For the $\text{PbZr}_{0.975}\text{Ti}_{0.025}\text{O}_3$ system doped with 1 wt% Nb_2O_5 , the phase boundary at *x* = 0.025 occurs at room temperature. In general, the external electric field tends to broaden the region of the ferroelectric state, and the mechanical compression extends the antiferroelectric region. Thus, the *AF*–*F* boundary can be made to shift either by an applied field or by a mechanical compression. This implies that the stored energy due to spontaneous polarization in the ferroelectric phase can be released by a shock compression wave to convert the ferroelectric phase to an antiferroelectric phase. This principle can be applied for the generation of high-energy electric pulses.⁶³

PVDF [(C H₂ – C F₂)_n]-Type Ferroelectric Polymers

The chemical formula for polyvinylidene fluoride (PVDF) is $(\text{CH}_2\text{--CF}_2)_n$. PVDF crystallizes from the melt below 150°C into spherulitic structures.^{64,65} In the melt, the polymer chains have randomly coiled shapes. Configurational defects can crystallize into regular conformations when cooled from the melt. This is accomplished by the torsional bond arrangements, in order to minimize the potential energy of the chains. The favorable arrangements are those having substituents at 180° between each other (called *trans* or *t*) or having substituents at ±60° (called *gauche*[±] or *g*[±]), the actual torsional angles commonly deviating from these values. There are three known conformations of PVDF: all-*trans*, *tg*⁺*tg*[−], and *tttg*⁺*tttg*[−]. The first two are by far the most common and most important ones,⁶⁵⁻⁶⁷ and they are illustrated schematically in Figure 4-22. In the all-*trans* conformation, all dipoles are aligned in the same direction normal to the chain axis. This yields the most highly polar conformation. The *tg*⁺*tg*[−] conformation is also polar but has components of the dipole moments parallel and perpendicular to the chain axis.

Depending on the temperature and the condition at which the material crystallizes, for most common polymorph PVDF produced by

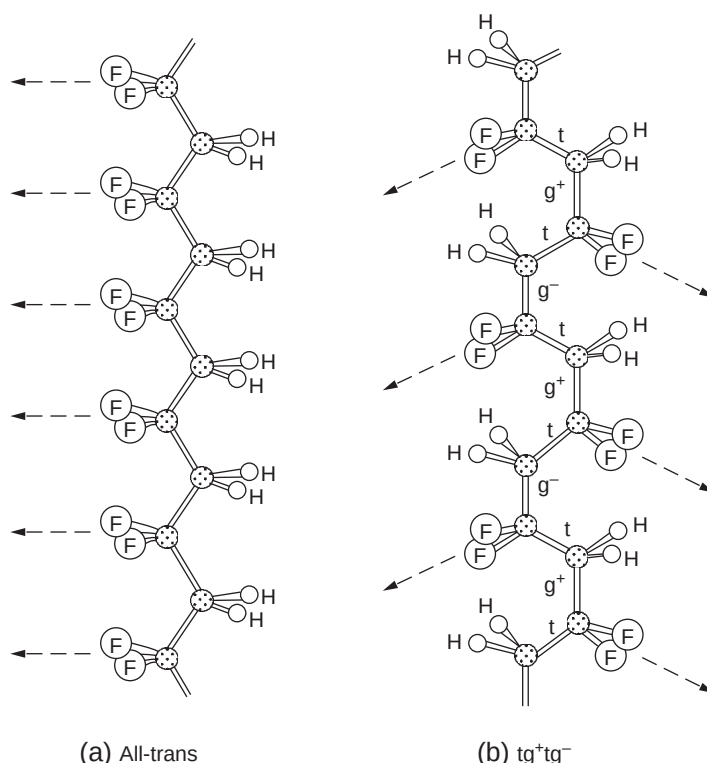


Figure 4-22 Schematic illustration of the two most common crystalline chain conformations in PVDF (a) all-trans conformation and (b) tg^+tg^- conformation. The arrows indicate the projection of the CF_2 dipole directions on planes defined by the carbon backbone.

melt–solidification, the unit cell consists of two chains in the tg^+tg^- configuration and possesses dipole moments with components parallel and normal to the chain axis. So, the dipole moments of the two chains in the unit cell are antiparallel and the crystal does not exhibit spontaneous polarization. In this case, the unpolar phase of PVDF is referred to as the α -phase. However, if the PVDF crystallizes at a temperature below 130°C and at a high pressure⁶⁸ or in conjunction with a special epitaxial technique,⁶⁹ the morphology of PVDF is changed to a structure oriented to the c -axis, and the unit cell consists of two all-trans chains packed with their dipoles pointing in the same direction. This crystal structure renders PVDF the most highly polar phase, which is generally referred to as the β -phase. The chains take on a planar zigzag ($tttt$) conformation. The chain

has a large dipole moment ($u_o = 7.06 \times 10^{-30}$ C·m per monomeric unit).⁶⁵ The degree of crystallinity for both α -phase and β -phase PVDF is approximately 50%. In the as-drawn β -PVDF film specimens, the dipole moment is preferentially oriented along the stretched $\pm y$ direction in the film plane, since mechanical stretching (drawing) causes a breakdown of the original spherulite structure into an array of crystallites whose molecules are oriented in the direction of the stretching force. If electrodes are then deposited on the film surfaces and a strong poling field of about 0.5 MV cm^{-1} is applied across the film specimen, the dipoles will reorient to align predominantly along the poling field in z direction perpendicular to the film plane.

Ferroelectricity of PVDC has been established on the basis of the measured dielectric

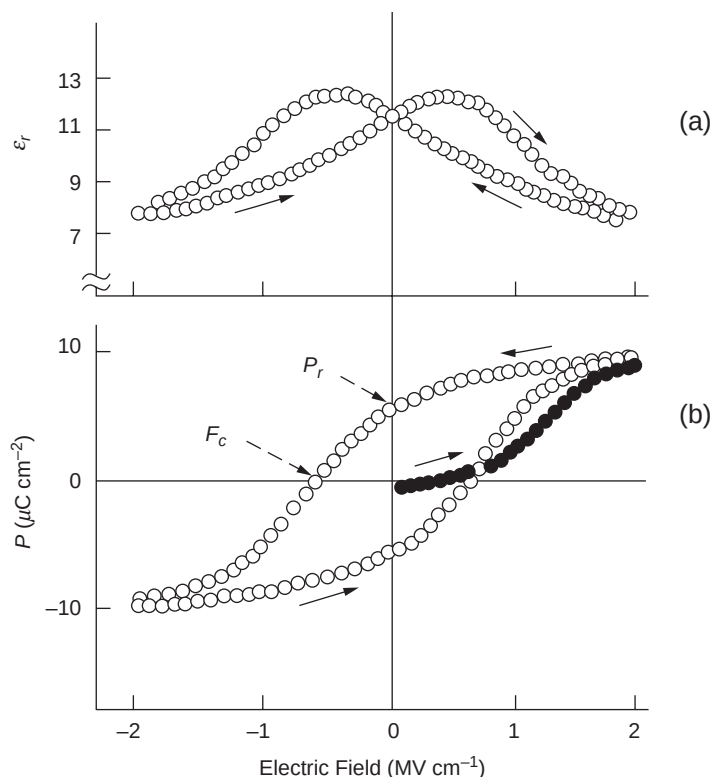


Figure 4-23 The variation of (a) dielectric constant and (b) polarization of β -phase PVDF films with applied electric field. ● Initial poling, ○ subsequent poling cycles.

constant of CF_2 and the corresponding polarization as functions of temperature, as shown in Figure 4-23. The experimental data are from Takahashi, Date, and Fukada.⁷⁰ A hysteresislike behavior of PVDF is good evidence that PVDF is a ferroelectric material. The Curie transition involves mainly intramolecular changes of polar orientation through the introduction of $g^\pm$ bonds that alter the polar all-trans conformation to a somewhat $tg^\pm$ and tt arrangement. As a result of this change, the Curie temperature for PVDF for the transition from the ferroelectric phase to a paraelectric phase would be around 205°C —about 20°C above its melting point.⁷¹

Although PVDF is ferroelectric, its dielectric constant is relatively small, around 10 to 15. But it exhibits strong piezoelectric and pyroelectric effects compared to other polymers, as first discovered by Kawai¹⁵ and Bergman et al.⁷²

This material has been widely used for piezoelectric transducers and pyroelectric devices.

4.2.3 Thermodynamic Theory

Classical thermodynamics deals mainly with the macroscopic behavior of materials. There are laws of thermodynamics that define the relationships between properties. To gain an understanding of the phase transitions and related effects, we need to understand the concept of free energy. One of the well known thermodynamic functions is the Helmholtz free-energy function A , which is given by

$$A = U - TS \quad (4-10)$$

where U is the internal energy, S is the entropy, and T is the temperature. There is also another well-known function, called the Gibbs free-energy function, which is given by

$$G = U - TS + pV \quad (4-11)$$

including a pV term. Most experiments in solids are performed at atmospheric pressure or pressures not far from it, which is so small that it makes the pV term negligibly small compared to other terms. Thus, by ignoring the pV term, the Gibbs function becomes

$$G_o = U - TS = A \quad (4-12)$$

The basic principle is the concept of the minimum free energy for the most stable system. If a system is allowed to have several alternative states, it will choose the one with the lowest free energy. To clarify the meaning of this principle, we apply it to a simple phenomenon. Based on Equation 4-12, a system can have the free energy at a lowest possible value if the internal energy U can be made as small as possible and the entropy S as large as possible. The internal energy U in a solid is primarily the potential energy, which is usually a negative quantity, implying that for a minimum potential energy all atoms (or ions) must rest at their lattice sites, that is, each atom (or ion) is resting at the bottom of its potential well. But this is a well ordered arrangement with very low entropy. Such an arrangement does not lead to a minimum free energy. Not all atoms (or ions) are at the bottom of the potential wells due to the thermal agitation, so the degree of order varies accordingly. At a given temperature T , the value of U and the degree of order S always tend to balance each other in order to reach a stable equilibrium state in which the free energy is the lowest. This is also why all atoms (or ions) in a solid are oscillating (vibrating) around their lattice sites.

Statistically, entropy is a measure of disorder. The higher the disorder (or the lower the degree of order), the larger is the entropy. Based on thermodynamics, the entropy can be defined as

$$dS = dQ/T = C_p dT/T \quad (4-13)$$

where dQ is the amount of heat absorbed by the system in a reversible process, and C_p is the specific heat at constant pressure. Based on

statistical mechanics, the entropy of a system is defined as

$$S = k \log q \quad (4-14)$$

where q is the number of microstates of the system. *Disorder* can be interpreted as the number of ways in which the particles inside a system can be arranged microscopically, so each different way can be considered a microstate or a thermodynamic state. Suppose that an alloy consists of two components: A and B . The total number of these two components is N per unit volume, of which n are of A type and the remaining $N-n$ are of B type. Since the A components can be interchanged themselves in $n!$ different ways and the B components in $(N-n)!$ different ways, the number of different arrangements for N components (i.e., the number of microstates in the system) is

$$q = \frac{N!}{n!(N-n)!} \quad (4-15)$$

Heat is one form of energy and energy cannot be created, it can only be converted from one form into other, different forms. The first law of thermodynamics states that the increase in internal energy dU is equal to the sum of the heat put in the system dQ and the work put in the system. When a solid body is subjected to an external electric field and mechanical stress, the internal energy will change accordingly. We can write

$$dU = TdS + Xdh + FdP \quad (4-16)$$

where X and h are, respectively, the mechanical stress and the strain; and F and P are, respectively, the electric field and the polarization. F and P are vectors, and they have three components in x , y , and z directions. X and h are also vectors and have six components in x , y , and z , as well as shear directions on the plane normal to each of x , y and z , so the energy state of a crystal is specified by the values of nine variables plus the temperature.

Thermodynamics can be used to describe the macroscopic behavior of the ferroelectric system by including external work done on the system. If the work put in the system is due to external mechanical stresses and electric fields, then the Gibbs function can be written as

$$G = G_o + \sum_i X_i h_i + \sum_j F_j P_j \quad (4-17)$$

The independent variables in the measurements of normal ferroelectric properties are the temperature T , the mechanical stress X , and the polarization P . Thus, it is convenient to discuss the G function as $G(T, X, P)$. In the absence of mechanical stresses, $X = 0$ and at a constant T , Gibbs function can be simplified to

$$G(T, P) = G_o(T) + \sum_j \frac{P_j^2}{2\epsilon_{ij}\epsilon_o} \quad (4-18)$$

where $G_o(T)$ is the Gibbs function under the unstressed and unpolarized conditions. The thermodynamic approach can be used to explain many ferroelectric properties. The following sections offer two transition processes as examples.

Ferroelectric Transition

It should be noted that in Equation 4-18 both ϵ_{ij} and P_j are temperature dependent and their behavior is anomalous and nonlinear near the transition point. Therefore, the behavior of the ferroelectric system can best be studied by expanding the energy in powers of the polarization. The Gibbs function for Perovskite-type ferroelectrics, formulated on the basis of the form of a power series by Devonshire⁷³ and further developed by Huibregtse and Young,⁷⁴ is given by

$$\begin{aligned} G(T, P) = G_o(T) &+ \frac{1}{2} f_2 (P_x^2 + P_y^2 + P_z^2) \\ &+ \frac{1}{4} f_4 (P_x^4 + P_y^4 + P_z^4) \\ &+ \frac{1}{2} f_{12} (P_y^2 P_z^2 + P_z^2 P_x^2 + P_x^2 P_y^2) \\ &+ \frac{1}{6} f_{111} (P_x^6 + P_y^6 + P_z^6) \\ &+ \frac{1}{2} f_{112} [P_x^2 (P_y^2 + P_z^4) + P_y^2 (P_z^4 + P_x^4) \\ &\quad + P_z^2 (P_x^4 + P_y^4)] \\ &+ \frac{1}{6} f_{123} P_x^2 P_y^2 P_z^2 + \dots \end{aligned} \quad (4-19)$$

where $f_2, f_4, f_{12}, f_{111}, f_{112}, f_{123}$, etc., are temperature-dependent coefficients. The series

does not contain terms of odd powers of P because crystals such as BaTiO₃ in the unpolarized phase have a center of inversion symmetry. Supposing that the applied poling field is in the z direction (c -axis of the cubic-tetragonal structure), then the polarization is also along the same z direction. By setting $P_y = P_x = 0$ and letting $P_z = P$ for simplicity, Equation 4-19 can be simplified to

$$G(T, P) = G_o(T) + \frac{1}{2} g_2 P^2 + \frac{1}{4} g_4 P^4 + \frac{1}{6} g_6 P^6 \quad (4-20)$$

where g_2, g_4, g_6 , etc., are new temperature-dependent coefficients. For the isothermal case, the electric field acting on the ferroelectric material, expressed in terms of P , can be obtained by

$$F = \frac{dG}{dP} = g_2 P + g_4 P^3 + g_6 P^5 \dots \quad (4-21)$$

At and near the Curie point, g_2 can be approximated with a linear function of temperature.⁷³ Thus, g_2 may be written as

$$g_2 = \beta(T - T_c) \quad (4-22)$$

where β is taken as a positive constant and T_c may be equal to or close to the Curie transition temperature. The nature of the transition depends on the signs of the coefficients: g_2, g_4, g_6 , etc. The simplest case is that all coefficients, except g_2 , are positive, and P is continuous at the transition point, implying that the transition is of the second order.

To find the values of P for which G is at the lowest value (minimum free energy), we differentiate Equation 4-20 with respect to P and set it equal to zero, that is, the values of P at $F = 0$.

$$g_2 P + g_4 P^3 + g_6 P^5 = 0 \quad (4-23)$$

For simplicity, we ignore terms with the powers of P equal to and higher than 5. Then we have the values of P for the lowest G value as

$$P = 0$$

and

$$P = \pm (g_2/g_4)^{1/2} \quad (4-24)$$

The plot of $G - G_0$ as a function of P for three cases— $g_2 > 0$, $g_2 = 0$ and $g_2 < 0$ —is shown in Figure 4-24(b). It can be seen that for $g_2 = 0$ (i.e., when $T = T_c$), P changes gradually from zero at $T = T_c$ to a finite value at $T < T_c$. The transition in BaTiO_3 is of the first order, but the energy of the Ti^{4+} ions in terms of their positions along the c -axis takes the form of two potential wells. An applied electric field in the

direction opposite to the polarization may enable the Ti^{4+} ions to pass over the potential barrier between the two wells and reverse the direction of the polarity.

If some coefficients other than g_2 are negative, the situation becomes more complicated. For simplicity, assume that only g_4 is negative and ignore all terms with powers of P higher than 6, then Equation 4-20 can be simplified to

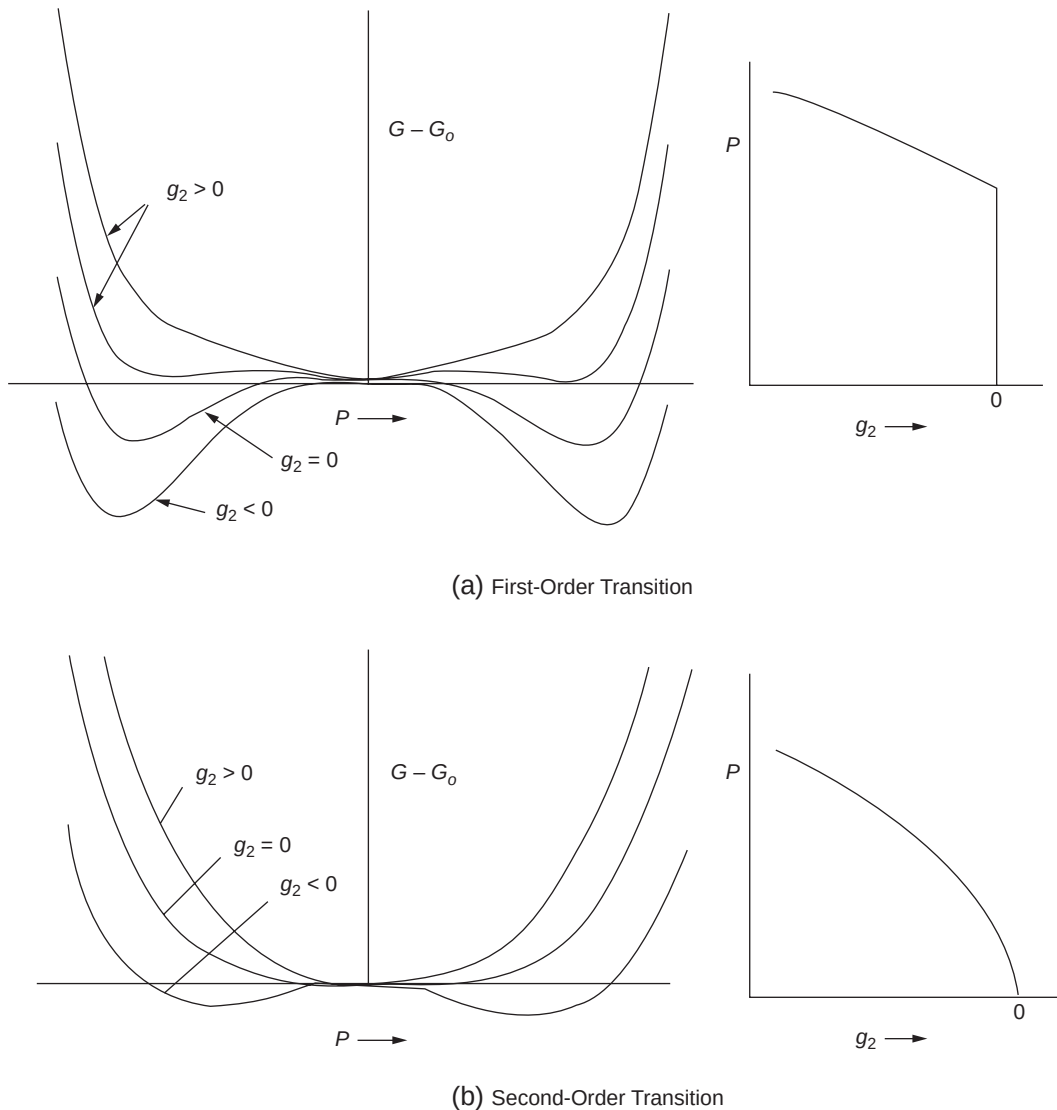


Figure 4-24 $G - G_0$ as a function of polarization at transition temperature and at temperatures just above and just below T_c for (a) first-order transition and (b) second-order transition.

$$G = G_o + \frac{1}{2}g_2P^2 - \frac{1}{4}|g_4|P^4 + \frac{1}{6}g_6P^6 \quad (4-25)$$

For the minimum values of G , we have

$$\frac{dG}{dP} = g_2P - |g_4|P^3 + g_6P^5 = 0 \quad (4-26)$$

The solution of Equation 4-26 yields

$$P = 0$$

and

$$P = \pm \left\{ \frac{1}{2g_6} \left[|g_4| \pm (|g_4|^2 - 4g_6g_2)^{1/2} \right] \right\}^{1/2} \quad (4-27)$$

Figure 4-24(a) shows the value of $G - G_o$ as a function of P for $g_2 > 0$, $g_2 = 0$ and $g_2 < 0$. For $g_2 = 0$ (i.e., $T = T_c$), we have $P = \pm(|g_4|/g_6)^{1/2}$. This implies that P changes discontinuously from $P = -(|g_4|/g_6)^{1/2}$ or $P = +(|g_4|/g_6)^{1/2}$ to $P = 0$ abruptly at $T = T_c$, and that the transition is of the first order.

If the poling field is not high, we can use first-order approximation by ignoring all terms with the powers of P higher than 2 in either Equation 4-20 or Equation 4-25. Then we have

$$F = \frac{dG}{dP} = g_2P = \beta(T - T_c)P \quad (4-28)$$

and P for $T > T_c$ can be expressed as

$$P = (\epsilon_r - 1)\epsilon_o F = \chi \epsilon_o F \quad (4-29)$$

Thus, from Equations 4-28 and 4-29, we obtain the Curie-Weiss relation

$$\chi = \frac{1}{\beta \epsilon_o (T - T_c)} = \frac{C}{T - T_c} \quad (4-30)$$

where $C = (\beta \epsilon_o)^{-1}$. This equation is similar to Equation 4-1.

This thermodynamic approach can also be used to explain many ferroelectric phenomena related to the first-order transition. Some of the common phenomena follow.

The Difference between the Curie Temperature and the Actual Transition Temperature

The actual transition temperature T_1 and the corresponding value of the spontaneous polarization can be calculated by imposing the con-

dition that the free energy of the polar and the nonpolar phases are equal. This leads to

$$\frac{1}{2}\beta(T_1 - T_c)P^2 - \frac{1}{4}|g_4|P^4 + \frac{1}{6}g_6P^6 = 0 \quad (4-31)$$

At the same time, the applied field is set equal to zero. This gives

$$F = \frac{dG}{dP} = (T_1 - T_c)P - |g_4|P^3 + g_6P^5 = 0 \quad (4-32)$$

From Equations 4-31 and 4-32, we obtain

$$P = \pm \left[\frac{3}{4} \left(\frac{|g_4|}{g_6} \right) \right]^{1/2} \quad (4-33)$$

$$T_1 - T_c = \frac{3}{16} \left[\frac{|g_4|^2}{\beta g_6} \right] \quad (4-34)$$

This indicates that the Curie temperature is lower than the actual transition temperature by $\frac{3}{16} [|g_4|^2 / \beta g_6]$. For BaTiO₃, $T_1 - T_c = 7.7^\circ\text{C}$.³⁰

Thermal Hysteresis

On heating the crystal through the Curie point, the polarization changes discontinuously from $P = 0$ in the ferroelectric phase to $P = 0$ in the paraelectric nonpolar phase. In the first-order transition, the crystal suddenly loses all the energy associated with the polarization at the Curie point, indicating a big change in its latent heat.

We can calculate the actual transition temperature T_2 , above which the polar phase does not exist in the absence of electric field. In Equation 4-27, for the minimum free-energy condition, there are several distinct values of P , for which G is a minimum. Above T_2 there is only one value of P , corresponding to the condition

$$|g_4|^2 - 4g_6\beta(T_2 - T_c) = 0 \quad (4-35)$$

Thus,

$$T_2 = T_c + |g_4|^2 / 4\beta g_6 \quad (4-36)$$

In this case the Curie temperature is lower than the actual transition temperature by $\frac{1}{4} [|g_4|^2 / \beta g_6]$. For BaTiO₃, $T_2 - T_c = 10^\circ\text{C}$,³⁰

indicating $T_2 > T_1$. In ferroelectrics with a first-order transition, such as BaTiO_3 , the transition temperature is always at a slightly lower value when the crystal is cooling from the high temperature than when it is heating from the low temperature. This phenomenon is generally referred to as *thermal hysteresis*.

Electric-Field Dependence of the Curie Temperature

Based on Equation 4-17, the differential form of the Gibbs free-energy function can be written as

$$dG = -SdT + \sum_i h_i dX_i + \sum_j P_j dF_j \quad (4-37)$$

Here we choose T, X, F instead of T, X, P as the independent variables. Under a constant stress condition with the applied field along one direction, Equation 4-37 can be simplified to

$$dG = -SdT + PdF \quad (4-38)$$

At the Curie point, the free energy in the ferroelectric phase and that in the paraelectric nonpolar phase are equal. So for $dG = 0$, we have

$$\frac{\partial T}{\partial F} = \frac{P}{S} \quad (4-39)$$

In this case, p and S are, respectively, the spontaneous polarization and the entropy at $T = T_c$. The discontinuous change in polarization at the Curie point causes a discontinuous change in entropy and hence the latent heat typical of a first-order transition. The entropy can be expressed as

$$S = -\left(\frac{\partial G}{\partial T}\right)_{X,P} \quad (4-40)$$

By differentiating Equation 4-25 with respect to T , and, for simplicity, ignoring all the terms with powers of P equal to or higher than 4 (because the coefficients of these high-power terms do not vary much with temperature), we obtain

$$S = \left(\frac{\partial G}{\partial T}\right)_{X,P} = -\frac{1}{2} \frac{\partial g_2}{\partial T} P^2 = \frac{1}{2} \beta P^2 \quad (4-41)$$

Substituting Equation 4-33 for P into Equation 4-41, we obtain

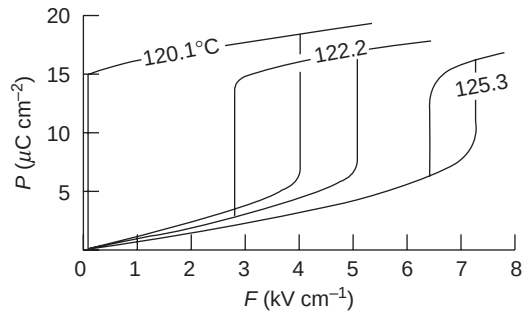


Figure 4-25 Hysteresis of BaTiO_3 above the Curie point at three different temperatures.

$$\frac{\partial T}{\partial F} = \frac{P}{S} = \frac{2}{\beta P} = \frac{4}{\sqrt{3}\beta} \left(\frac{|g_4|}{g_6}\right)^{1/2} \quad (4-42)$$

For BaTiO_3 , $\partial T/\partial F = 1.72 \times 10^{-3} \text{ } ^\circ\text{C cm V}^{-1}$. This value agrees well with the experimental value, which is $1.43 \times 10^{-3} \text{ } ^\circ\text{C cm V}^{-1}$.³⁰

The above argument indicates mainly that at temperature $T > T_c$, the crystal can be forced by an applied field to make a transition into a ferroelectric state and then return to the paraelectric nonpolar state when the field is reduced, producing a double hysteresis loop, as shown in Figure 4-25 for BaTiO_3 . The data for Figure 4-25 are from Meyerhofer.⁷⁵

Antiferroelectric Transition

A crystal is antiferroelectric if it exhibits paraelectric behavior, (i.e., the dielectric constant increases with decreasing temperature above the transition point) but does not show a net spontaneous polarization and the dielectric constant is considerably low below the transition point. A typical example is the PbZrO_3 crystal (see Figure 4-20). This indicates that the transition involves a different mechanism for dielectric polarization. X-ray analysis has revealed the existence of a superstructure below the transition point. The superstructure can be described in terms of two superlattices having equal but opposite polarization; above the transition temperature, the two superlattices are crystallographically equivalent and are unpolarized.

In the previous section, we analyzed ferroelectric phenomena based on an appropriate

thermodynamic free energy function G , which is expressed in powers of the polarization P . According to Kittel,⁷⁶ the same approach can be applied to an antiferroelectric crystal if G is expressed in powers of the polarization P_a and P_b of the two superlattices. It should be noted that, unlike P for ferroelectrics (which can be detected), P_a and P_b for antiferroelectrics are not macroscopically observable quantities. However, following the same procedure for Equation 4-20, we can write

$$G(P_a, P_b, T) = G_o(T) + f(P_a^2 + P_b^2) + gP_aP_b + h(P_a^4 + P_b^4) + k(P_a^6 + P_b^6) + \dots \quad (4-43)$$

Whether the ferroelectric state ($P_a = P_b$), the antiferroelectric state ($P_a = -P_b$), or the paraelectric state ($P_a + P_b = 0$) has the lowest free energy depends on the values of the coefficients.

To deal with the antiferroelectric transition, it is convenient to transform P_a and P_b into the following parameters:

$$\begin{aligned} P_+ &= P_a + P_b \\ P_- &= P_a - P_b \end{aligned} \quad (4-44)$$

We can consider P_+ as the total polarization and P_- as a measure of the antiferroelectric state. Substitution of Equation 4-44 into Equation 4-43 gives

$$\begin{aligned} G(P_+, P_-, T) &= G_o(T) + \frac{1}{2}g_2P_+^2 + \frac{1}{4}g_4P_+^4 \\ &\quad + \frac{1}{6}g_6P_+^6 + \dots \\ &\quad + \frac{1}{2}g'_2P_-^2 + \frac{1}{4}g'_4P_-^4 \\ &\quad + \frac{1}{6}g'_6P_-^6 + \dots \\ &\quad + \frac{1}{2}\lambda P_+^2P_-^2 + \dots \end{aligned} \quad (4-45)$$

where the coefficients $g_2, g_4, g_6, g'_2, g'_4, g'_6, \dots$ and λ , etc., are related to the coefficients f, g, h, k , etc.

Detailed analysis of antiferroelectric behavior is beyond the scope of this chapter. In general, an antiferroelectric state exists if G has a minimum for a finite value of P_- . Obviously, this condition depends strongly on the sign and

the value of the coefficients. Normally, the coefficient g'_2 decreases with decreasing temperature, and the antiferroelectric state is stable at low temperatures. If the term $\frac{1}{2}\lambda P_+^2P_-^2$ is positive and sufficiently large, then one kind of the state (e.g., antiferroelectric state) will prevent the formation of the other (e.g., ferroelectric state).

The sharp decrease of the dielectric constant below the transition point is due to the saturation effect of P_a and P_b . Antiferroelectrics can be piezoelectric. Some antiferroelectric materials may also become ferroelectric under the application of a suitable electric field.⁷⁷ It is not so obvious that domains would be formed in antiferroelectrics. However, optical observation has revealed the domain structure of PbZrO_3 and NaNbO_3 antiferroelectrics. The domains are nonpolar; therefore, their patterns should not be affected by external electric fields. For more details about antiferroelectric phenomena, see references.^{38,78}

A two-dimensional picture illustrating the basic difference among the ferroelectric, antiferroelectric, and ferrielectric states is given in Figure 4.26. In the ferroelectric state, there is polarization in the vertical direction and no polarization in the horizontal direction; in the antiferroelectric state, there is no net polarization in either the vertical or the horizontal direction; in the ferrielectric state, the crystal could be antiferroelectric in the vertical direction and ferroelectric in the horizontal direction.

However, ferrielectric behavior is not really similar to its ferrimagnetic counterpart. It is a very strange behavior. An example of ferrielectric behavior is thiourea $[\text{SC}(\text{NH})_2]$. The reversal of the polarization in this crystal is related to relative displacements of entire molecules, rather than to the motion of single atoms or ions in the lattice, and its polarization phenomenon can be considered ferrielectric.⁷⁹

4.2.4 Formation and Dynamics of Domains

A single crystal or a system that is ferroelectric will undergo spontaneous polarization below the Curie temperature. If all the dipoles of the

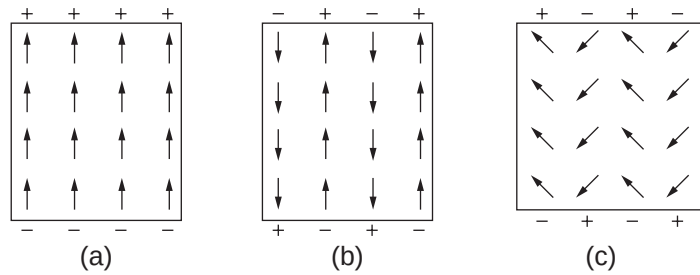


Figure 4-26 Schematic illustration of the polarization in (a) ferroelectric structure, (b) antiferroelectric structure, and (c) ferrielectric structure.

polarization are pointing in one direction, the electrostatic energy of the system is extremely large and the system becomes very unstable. Naturally, a system always tends to minimize its potential energy by forming the so-called domains, which are actually regions in the crystal, each containing a large number of dipoles all aligned in the same direction. Domains are arranged in such a way that the polarization of different domains will compensate for each other and the net polarization of the whole crystal along any direction will vanish. Domains are separated by domain walls, which are the loci of the dipole orientation from one direction of the domain to another of the neighboring one. A single crystal may contain many domains with zero net polarization, but it can be changed to a single crystal with a single domain by a strong poling electric field, in which the dipoles of all domains must point in the same direction. In other words, all domains will join to form a large domain. For polycrystalline materials such as ceramics, single domain cannot be achieved because the crystal axes of the grains (crystallites) in the material are randomly arranged, which cannot be altered by the electric field.

When a phase transformation begins, the domains will be nucleated at several places within the crystal, and the nuclei of domains will grow along the ferroelectric axes until the transformation of the new phase in the whole volume is completed. Ferroelectric domains are closely analogous to ferromagnetic domains. In uniaxial ferroelectrics such as TGS crystals, the domains can have only two possible dipole

directions, so the polarization will be in opposite directions in adjacent domains by twinning. In this case, the domain walls are called the 180° walls, as shown in Figure 4-27(a).

For multiaxial ferroelectrics such as BaTiO_3 , domains can have more than two dipole directions. For example, in the tetragonal phase, polarization is along the c -axis, that is, the elongated axis of the cubic structure. But in the ferroelectric phase, there are six easy directions deriving from the cubic structure, that is, the c -axis can be in the $\pm x$, $\pm y$, and $\pm z$ directions. This implies that there are three pairs of antiparallel directions along the cube edges and that domains can have six directions of polarization. This gives rise to different types of domain walls: Those separating antiparallel dipoles are 180° walls, and those separating dipoles at right angles to each other are 90° walls, as shown in Figure 4-27(b) and (c). Similarly, for the ferroelectric phase with the unit cell in rhombohedral structure, the polarization is along the body-diagonal directions. In this case, there are eight easy directions for spontaneous polarization. The angle between adjacent domains becomes 70.5° and 110° , so the domain walls are called the 70.5° and 110° walls. Some typical domain patterns in crystals are illustrated schematically in Figure 4-27.

The electrostatic energy decreases with an increase in the number of domains formed due to the decrease in the depolarizing field. However, the domain formation process does not continue indefinitely, because a certain amount of energy is stored in the walls between different domains. For ideal crystals, when the

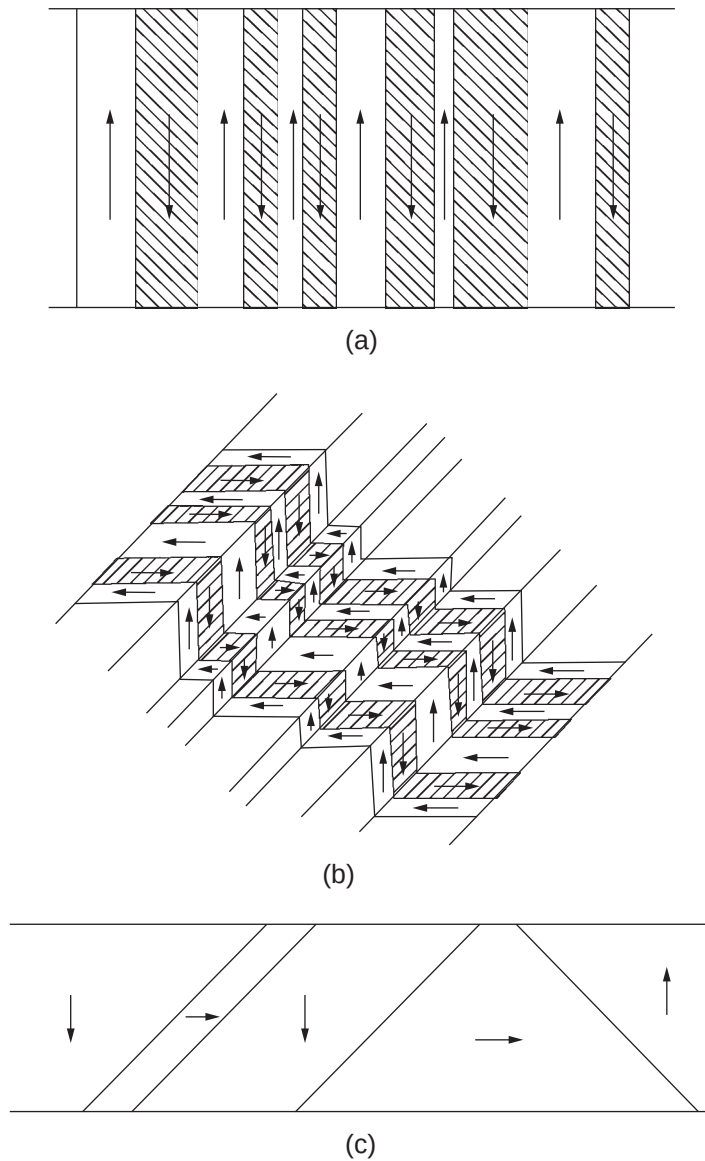


Figure 4-27 Schematic illustration of some domain patterns: (a) antiparallel domains with 180° walls, (b) domains with 180° and 90° walls, and (c) mixture of domains in c and a directions (a is perpendicular to c).

domain wall energy has increased to a certain critical value, at which it just balances the decrease in energy of the depolarizing field, the total potential energy of the crystal becomes minimal. Under this condition, the domain configuration has reached an equilibrium and stable state at the temperature considered. But for real crystals that are not completely

nonconductive and contain various defects, the domain configuration is hardly stable at any temperature because the charges induced by spontaneous polarization are partially compensated by the conducting carriers in the materials, and the uniformity of the polarization and the depolarizing field is disturbed by various defects. Because of such defects and the

conductivity of real crystals, the domain patterns do not always correspond to the absolute minimum of the free energy. This implies that the domain configuration is only metastable, resulting in the aging effect.

The calculation of domain wall energy is very complicated, but the width of the domain walls is directly related to the total energy of the domain walls. The total wall energy is the sum of the energy due to dipole interaction and the anisotropy energy.^{22,39} Thus, the total wall energy U_{wall} can be written as

$$U_{\text{wall}} = \frac{A}{\delta} + B\delta \quad (4-46)$$

where δ is the wall width and A and B are functions of the lattice constant, the spontaneous strain and the elastic constant. By minimizing the wall energy

$$\frac{dU_{\text{wall}}}{d\delta} = 0 \quad (4-47)$$

This leads to

$$\delta = (A/B)^{1/2} \quad (4-48)$$

Merz²² has estimated the wall width to be of the order of several lattice constants and wall energy of the order of 10^{-6} joule cm^{-2} , in contrast to the ferromagnetic domain walls whose width is of the order of several hundred lattice constants. In a ferromagnetic domain wall, the magnetization vectors must turn over gradually from one direction to the opposite one, which involves a change in exchange energy and anisotropy energy. This is why the ferromagnetic domain walls are much wider than the ferroelectric domain walls. The difference between the ferroelectric and the ferromagnetic domain walls is shown in Figure 4-28.

In Equation 4-46, the first term A/δ is the energy from dipole interaction due to the fact that the polarization changes from $+P$ to $-P$ in the wall, and the second term $B\delta$ is the energy from the stress causing the strain changes from $+$ strain to $-$ strain in the wall. In ferroelectric materials, the first term (due to sideways interaction of adjacent dipoles) is small, whereas the elastic energy due to the changes of the strain is large. This is why the domain-wall width is

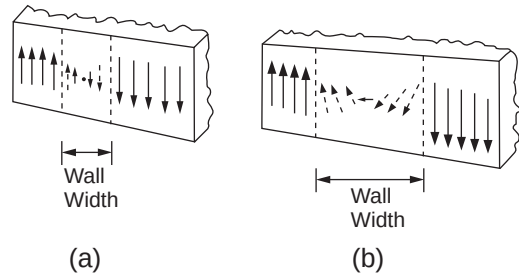


Figure 4-28 Schematic illustration of the difference between (a) the ferroelectric and (b) the ferromagnetic 180° domain walls. The width of the domain walls is about several lattice constants for ferroelectrics and about several hundred lattice constants for ferromagnetics.

only several lattice constants. For ferromagnetic materials, the domain walls are a hundred times wider because the interaction energy of magnetic dipoles is large and the elastic energy is small.

In general, strain energy is the dominant component in the ferroelectric domain wall energy budget, and the 180° and 90° walls are thus distinctly different in behavior. It is easy for the 180° walls to move under the influence of an applied electric field. The motion of the 180° wall merely requires the cations to move cooperatively toward the other end of their cages without involving the rotation of the polarization vectors or any change in the shape of the crystal. This implies that the domains with the polarization in the direction of the applied field expand sideways by moving the walls toward the domains with the polarization opposite to the field direction, making them shrink and eventually vanish. In other words, the two antiparallel domains will join or coalesce to form a big domain, with all dipoles aligned in the field direction.

For the reversal of the polarization, an applied field should be normally higher than the coercive field F_c , and the time required for the complete reversal depends on the magnitude of the applied field following the relation

$$t_s = kF^{-n} \quad (4-49)$$

where k and n are constants.³⁰ It should be noted that when the applied field is small, the reversal may still occur, but a much longer switch-

ing time is required. This implies that the so-called coercive field F_c does not have a clear physical meaning, as fields much smaller than F_c still switch on the reversal, if the applied field is held for a sufficiently long time. For low fields, t_s follows an exponential law $i_s \propto \exp(\alpha/F)$, rather than a power law, given in Equation 4-49.

The switching time is directly related to the velocity of the motion of domain walls. At high fields ($>1 \text{ kV cm}^{-1}$), the sideways wall velocity follows the relation⁸⁰

$$v_s = hF^m \quad (4-50)$$

where h and m are positive constants. It can be seen that v_s is approximately proportional to the reciprocal of the switching time t_s . However, when two 180° walls are sufficiently close, they do not move toward each other and there appears to be a repulsion between them. This repulsion results in the formation of a very narrow metastable region which suddenly disappears when an electric field is applied, creating Barkhausen pulses.

The 90° walls are relatively difficult to move and usually require an applied electric field greater than a certain threshold value. In general, the motion involves the nucleation and growth of new domains with the polarization in the direction of the applied field. Thus, making the domains with polarization perpendicular to the applied field change to polarization along the field direction may sometimes cause the crystal to change shape. The process for a complete polarization in the direction of the applied field for a multiaxial ferroelectric could be quite complicated. Let us take a simple case as an example. For a single-domain crystal with the polarization along the c -axis in the tetragonal phase, if an electric field is applied along the a -axis (perpendicular to the c -axis), then a large number of needlelike new domains nucleate at the surface and are polarized in the direction of the applied field, growing in the form of wedges.⁸¹ The formation of new domains continues until the whole crystal is polarized in the same direction of the applied field. These newly formed domains originate at the surface of the crystal specimen (usually at the electrode

surface) and grow in the direction of the applied field with very little sideways motion of the walls. However, it can be imagined that the nucleation of new domains becomes more difficult because more nuclei are already present, due to more electro-mechanical interaction, particularly in multidomain crystals.

Sideways motion of 90° walls in an applied electric field has been observed.^{39,81} Since the 90° wall width is only about several lattice constants, the activation energy for the wall motion is only a small fraction of the total wall energy. The critical field for the 90° wall motion is smaller than that for the nucleation and the growth of new domains. However, the motion causes a change in the shape of the crystal. For a more detailed discussion of the formation, structure, and dynamic behavior of domains, see references 22, 30, 38, 39, 78, and 80–86.

4.2.5 Ferroelectric Materials

There are now many hundred materials that exhibit ferroelectric or antiferroelectric properties. In Section 4.2.2, the properties of some representative ferroelectrics were discussed. In this section, we will not list all the ferroelectric materials but summarize some important properties of some ferroelectrics with the potential for practical applications in Table 4-2.

The most commonly used ferroelectrics have the Perovskite structure, with the chemical formula ABO_3 . The PZT system is a good example, consisting mainly of the alloy of PbTiO_3 and PbZrO_3 . The general formula of the PZT system is $\text{PZ}_{1-x}\text{T}_x$, so the ferroelectric properties can be tailored by varying the composition of Zr and Ti, that is, by varying the value of x .

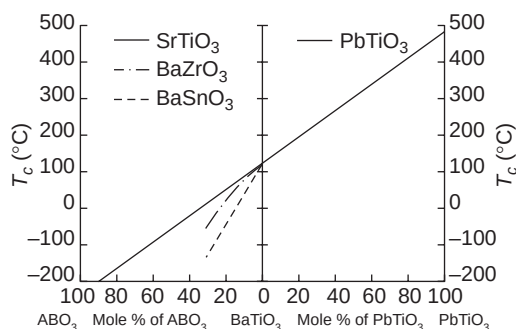
It is not always convenient to keep ferroelectric devices in an oven or a cryostat in order to maintain the required properties. Now, materials can be modified so that certain values of the required properties can be achieved at room temperature. The control of properties is possible by varying the composition in solid solution or by incorporating suitable additives.⁸⁷ One good example is a ceramic formed by a solid

Table 4-2 General transition properties of some ferroelectric crystals.

Name (Abbreviation)	Chemical Formula	Curie Temperature, T_c (°C)	Spontaneous Polarization P_s ($\mu\text{C m}^{-2}$) at $[T$ (°C)]	Crystal Structure	
				Above T_c	Below T_c
Barium Titanate	BaTiO_3	120	26.0 [23]	Cubic	Tetragonal
Lead Titanate	PbTiO_3	490	50.0 [23]	Cubic	Tetragonal
Potassium Niobate	KNbO_3	435	30.0 [250]	Cubic	Tetragonal
Potassium Dihydrogen Phosphate (KDP)	KH_2PO_4	-150	4.8 [-177]	Tetragonal	Orthorhombic
Triglycine Sulfate (TGS)	$(\text{NH}_2\text{CH}_2\text{COOH})_3 \cdot \text{H}_2\text{SO}_4$	49	2.8 [20]	Monoclinic (Centrosymm.)	Monoclinic (Noncentrosymm.)
Potassium-Sodium Tartrate-Tetrahydrate (Rochelle salt)	$\text{KNaC}_4\text{H}_4\text{O}_6 \cdot 4\text{H}_2\text{O}$	24	0.25 [5]	Orthorhombic (Centrosymm.)	Monoclinic (Noncentrosymm.)
Antimony Sulfo-iodide	SbSI	22	25.0 [0]	Orthorhombic (Centrosymm.)	Orthorhombic (Noncentrosymm.)
Guanidinium Aluminium Sulfate Hexahydrate (GASH)	$\text{C}(\text{NH}_2)_3\text{Al}(\text{SO}_4)_2 \cdot 6\text{H}_2\text{O}$	None	0.35 [23]	Trogonal	—

solution of PbTiO_3 in BaTiO_3 , in which the transition temperature depends on the concentration of Pb in BaTiO_3 . In this case, all concentrations are possible and they are ferroelectric. An increase in Pb concentration causes the transition temperature T_c to increase from 120°C to 490°C. Similarly, the substitution of zirconium for titanium in BaTiO_3 lowers the value of T_c , as shown in Figure 4-29. Iron, strontium, and calcium also tend to lower the value of T_c . For details about this subject, see references 39, 88, and 89.

Ferroelectrics in ceramic forms are commonly used for practical applications because they have the advantage of being a great deal easier to prepare than their single crystal counterparts. Also, in many cases their ferroelectric properties approach quite closely those of the single crystal. As mentioned above, a wide range of ceramic compositions and additives can be used to adjust the properties of the material for different applications. Ceramics are generally prepared by grinding the material,

**Figure 4-29** The effect of various additives in BaTiO_3 on the Curie temperature T_c .

together with any required additives, into powder. This reacted powder is usually ball-milled to increase homogeneity and particle reactivity. This may be processed with or without binder. Then the material is extruded or pressed into some shape and fired at a higher temperature (e.g., 1300°C), at which sintering occurs.⁴⁴ This fabrication procedure produces a

complicated polycrystalline material that has pores, some unavoidable impurities, and many crystallites of various sizes. Some properties of a ceramic may differ from those of the corresponding single crystal material because the crystallites in the ceramic are randomly oriented and some of them are under stresses even when no external force is applied; also, there exist grain boundaries and porosity.

The ferroelectric properties of ceramics depend greatly on the grain (crystallite) size. In a ceramic with small grain sizes, a significant portion of the material volume is affected by grain boundaries. The important difference between grain boundaries and free surfaces is that in high-density ceramics, the material close to the grain boundaries is elastically clamped by the neighboring grains at temperatures well below the sintering temperature. On cooling to the ferroelectric phase, large mechanical stresses may be developed by the anisotropic spontaneous strain, and these stresses may in turn affect the domain dynamics and configuration. If the grains are too small, the polarization may be completely clamped, preventing the action of domain reversal when the applied field is turned to the opposite direction. As the size of grains increases, domain reversal becomes possible, but the peak dielectric constant decreases, due mainly to the inhomogeneous distribution of the mechanical stresses and electric fields. If the polarization in the neighboring grains is not parallel, then the nonzero polarization at the boundaries creates depolarizing fields, tending to compensate free charges at the electrodes.

The grain sizes of ceramics can be controlled to a certain degree during the fabrication processing by adjusting the sintering temperature and time, the composition of the material, and the additives. Most ceramics consist of grains with a range of grain sizes, and their crystallographic axes are randomly oriented. It can be imagined that the poling of a ceramic under an external electric field involves microscopic processes much more complicated than those for single crystals. Because of the grain boundaries, neither dielectric nor elastic relaxations under an alternating field within the bulk of

ceramics are possible. This is why the hysteresis loops are generally rounded, the spontaneous and the remanent polarizations are lower, and the coercive fields are larger; this is also why the aging effect is enhanced. As knowledge of ferroelectric ceramics progresses, new effects and new fabrication techniques, as well as new materials, will certainly be discovered.

It is worth mentioning that ferroelectric thin films are very important for practical applications. For certain applications, ferroelectrics in thin-film form are preferred. In general, the dielectric and ferroelectric properties are thickness dependent if the thickness of the films is below a certain value. For example, BaTiO_3 films prepared by electron-beam evaporation or sputtering techniques exhibit normal dielectric and ferroelectric behavior if the film thicknesses are larger than $1\mu\text{m}$. Ferroelectric behavior becomes unstable for film thicknesses below $0.1\mu\text{m}$.⁹⁰ But some investigators⁹¹ have reported that vacuum-evaporated BaTiO_3 films show well defined dielectric anomalies near 120°C for films with thicknesses down to $0.023\mu\text{m}$. For BaTiO_3 films prepared by flash evaporation, the properties remain approximately similar to those of thick films down to a thickness of $0.1\mu\text{m}$.⁹² The discrepancy can be attributed to different techniques of film fabrication, which result in various structural and chemical defects. Defects aside, the interface between the film and the electrical contacts, the width of the depletion regions inside the film, and electronic conductivity also play important roles in the thickness dependence of the dielectric and ferroelectric properties of thin films.

Although several film fabrication techniques (chemical vapor deposition, single or multiple target sputtering, multiple target ion beam, pulsed deposition techniques, etc.) have been put forward,^{58,93} none can be considered very satisfactory. However, the required degree of perfection varies according to the specific application; thus, the technique for film fabrication should be chosen accordingly. For example, for ferroelectric films used in small capacitors, we would want the films to have a high dielectric constant, low dielectric loss, and a minimal variation of these parameters with

temperature, applied field, and frequency. But for films used in optical devices, we would want the films to have a high optical quality.

The trend of using ferroelectric films is ever increasing. The development of new techniques and the improvement of the existing techniques have to a great extent been motivated by various technical applications. It is certain that new experimental and theoretical results in this area will be available in the near future.

4.2.6 Applications of Ferroelectrics

Based on the classification of crystals given in Figure 4-1 the following statements can be made:

- All ferroelectrics possess piezoelectric and pyroelectric effects.
- All pyroelectrics possess piezoelectric effects, but not all are ferroelectric.
- Piezoelectrics may have only piezoelectric effects, both piezoelectric and pyroelectric effects, or all three effects: ferroelectric, piezoelectric, and pyroelectric.

Depending on the crystal structure, spontaneous polarization in a poled ferroelectric crystal can be varied to produce different effects by the application of an external force, such as a mechanical stress, an electric field, or a change in temperature. Figure 4-30 illustrates these effects. Each of these effects can be used

for practical applications. The most commonly used effects are the piezoelectric, pyroelectric, and electro-optic effects, which are discussed, respectively, in Piezoelectric Phenomena and Pyroelectric Phenomena in this chapter and Modulation of Light in Chapter 3. In this section, we shall briefly describe some applications related directly to the special features of ferroelectrics: mainly, spontaneous polarization, nonlinearity, and hysteresis behavior.

Capacitors

Capacitors are important elements in electrical and electronic circuits, performing various functions that include blocking, coupling and decoupling, AC-DC separation, filtering, power factor correction, energy storage, etc. For a capacitor with the dielectric material of dielectric constant ϵ_r and thickness d between two parallel metallic plates of area A , the capacitance is given by

$$C = \epsilon_r \epsilon_0 \frac{A}{d} \quad (4-51)$$

If the working voltage is V , the average working field F is V/d , which must be lower than the breakdown strength of the dielectric material F_b by factor k . The volume efficiency of the capacitor is defined as

$$\eta = \frac{C}{dA} = \frac{\epsilon_r \epsilon_0}{d^2} \quad (4-52)$$

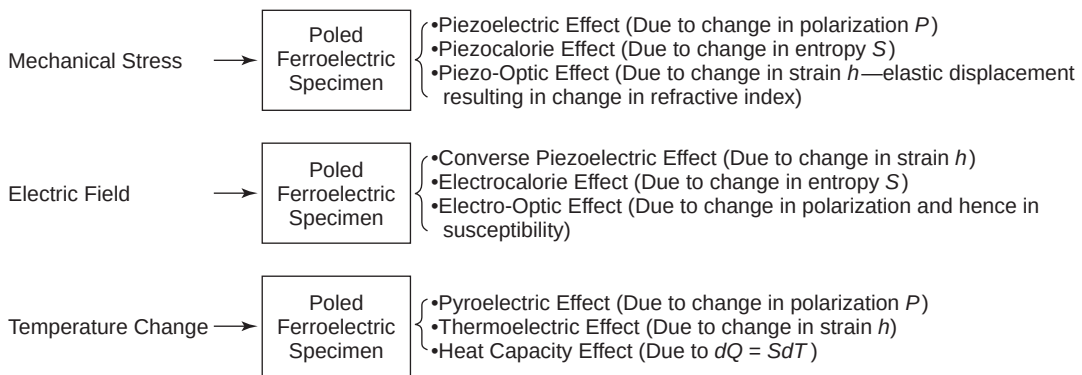


Figure 4-30 The effects of a mechanical stress, an electric field, or a temperature change on the behavior of a poled ferroelectric specimen.

The maximum permissible energy density stored in the capacitor is thus

$$\frac{CV^2}{2Ad} = \frac{\epsilon_r \epsilon_0 F_b^2}{2k^2} \quad (4-53)$$

The value of $\epsilon_r F_b^2$ is a figure of merit for dielectrics.

All dielectric materials have a finite value for DC resistivity. Thus, a charged capacitor will discharge gradually through its own resistance following

$$Q(t) = Q_0 \exp(-t/\tau) \quad (4-54)$$

where Q_0 and $Q(t)$ are, respectively, the initial charge and the charge remaining in the capacitor at time t ; τ is the dielectric relaxation time of the material, which is equal to $\epsilon_r \epsilon_0 \rho$; and ρ is the resistivity of the material. A capacitor dissipates its stored energy through DC leakage resistance and, more importantly, the dielectric losses under an AC electric field. These account for the dissipation factor or the loss tangent $\tan \delta$ (see section 2.2.3 in Chapter 2). It should also be noted that the leads and the electrodes always introduce an inductance L in the circuit. This is why a capacitor has a resonance frequency which may be very high, depending on the value of L . For general applications, it is important to keep the value of the unwanted stray inductance L as small as possible.

However, from Equations 4-52 and 4-53, both the volume efficiency and the maximum energy stored in a capacitor increase proportionally with increasing dielectric constant. Thus, ferroelectric materials with high dielectric constants are obviously ideal materials for use in miniature capacitors. But the capacitance of the capacitors made of ferroelectric materials is dependent on temperature, applied electric field, and frequency. Usually, it is not so difficult to tolerate a small change in electric field and frequency, but the range of temperature variation is large under certain environmental conditions, and also the dielectric constant is strongly dependent on temperature, particularly at temperatures near the transition point. So for high-permittivity capacitors, it is necessary to choose a material with a small temperature coefficient of the capacitance (TCC), which is

$$TCC = \frac{1}{C} \frac{dC}{dT} \quad (4-55)$$

From Equation 4-51 we obtain

$$\begin{aligned} TCC &= \frac{1}{\epsilon_r} \frac{\partial \epsilon_r}{\partial T} + \frac{1}{A} \frac{\partial A}{\partial T} - \frac{1}{d} \frac{\partial d}{\partial T} \\ &= TCD + \alpha_L \end{aligned} \quad (4-56)$$

where TCD is the temperature coefficient of the dielectric constant and α_L is the linear expansion coefficient. So the change in capacitance due to a change in temperature is caused by the change of the capacitor's dimension, which is usually not very significant, and by the change of the dielectric constant, which is significant, particularly at temperatures near the transition point.

Ferroelectric ceramics are commonly used for capacitors. The Electronic Industries Association (EIA) in the United States has grouped ceramic capacitors into three classes:

Class I: The materials have a low dielectric constant (ranging from 15 to 500), low dielectric losses ($\tan \delta < 0.003$), a low TCC , and a low rate of aging of the capacitance value. The working temperature range is -55°C to $+85^\circ\text{C}$.

Class II: The materials have medium to large dielectric constants (ranging from 500 to 20,000), with general properties primarily similar to BaTiO_3 ceramics. The dependence of temperature, electric field, and frequency is stronger than for Class I, and the TCC value is higher.

Class III: The materials consist of a conductive phase near the electrodes that reduces the effective thickness of the dielectric material. General properties are similar to those of Class II, but their working voltage is lower. The EIA has also introduced a scheme for specifying the variability of the capacitance with temperature in the range of practical interest for high-permittivity capacitors. Table 4-3 lists the EIA codes. For example, a capacitor of EIA code X7E implies that its capacitance may change by no more than $\pm 4.7\%$ in the range of temperatures from -55°C to $+125^\circ\text{C}$.

Table 4-3 The EIA codes for the specification of high-permittivity capacitors.

EIA Code	Range of Temperature Variation (°C)	EIA Code	Change of Capacitance (%)
X7	−55 to +125	D	−3.3 to +3.3
X5	−55 to +85	E	−4.7 to +4.7
Y5	−30 to +85	F	−7.5 to +7.5
Z5	+10 to +85	P	−10.0 to +10.0
		R	−15.0 to +15.0
		S	−22.0 to +22.0
		T	−33.0 to +22.0
		U	−56.0 to +22.0
		V	−82.0 to +22.0

Dielectric materials with dielectric constants larger than 1000 are mainly ferroelectric materials, and they are very sensitive to temperature, electric field, and frequency. Obviously, the development of techniques for modifying the materials to improve stability and to retain the desirable high dielectric constant feature is very important. In fact, this work has been carried out continuously for about 50 years.

In general, the dielectric constant is measured with a small AC signal. A bias of a large DC field will cause the value to decrease, but the field coefficient of the dielectric constant is normally small. Frequencies in the range of 0–100 MHz do not affect the dielectric constant to a great extent. The difference in dielectric constant between power and radio frequencies is only 5–10%, and for most ceramics based on BaTiO₃ and its isomorphs, the dielectric constant does not vary greatly with frequency until it reaches the gigahertz range. The major concern is the temperature dependence of the dielectric constant. To reduce the temperature coefficient of the dielectric constant (TCD) is thus the goal that many researchers have attempted to reach. High-permittivity ceramic capacitors are based mainly on ferroelectric BaTiO₃, its isomorphs, and their solid solutions. According to Lichtenecker's rule,⁹⁴ the TCD of a ceramic formed by the mixture of several components will be equal to the volume average of the TCDs of its constituents, if there is no chemical reaction between constituents leading to the formation of new compounds. Following this rule, the dielectric constant of the mixture can be predicted approximately by Lichtenecker's empirical formula

$$\log \epsilon_r = v_H \log \epsilon_{r1} + v_{f2} \log \epsilon_{r2} + v_{f3} \log \epsilon_{r3} + \dots$$

$$= \sum_{i=1}^n v_{fi} \epsilon_{ri} \quad (4-57)$$

where ϵ_{ri} and v_{fi} are, respectively, the dielectric constant and the fraction of the total volume of the mixture of the i th constituent.

In general, it is not possible to obtain a completely flattened $\epsilon_r - T$ curve in a homogeneous ceramic having BaTiO₃ as a major constituent. However, in the region below the peak, at temperatures below T_c , the value of the dielectric constant does not change much with temperature. This is due partly to the further change in crystal structure to rhombohedral and then to orthorhombic, and partly to the transition changing from first order to second order.⁹⁵ Some ceramics with different compositions, which would provide usefully high permittivity–temperature characteristics, are given here as examples. Many similar mixtures have already been developed.^{58,95,96}

BaTiO ₃ and Bi ₂ (SnO ₃) ₃	($\epsilon_r = 1000$ at 25°C)
BaTiO ₃ and MgSnO ₃	($\epsilon_r = 2000$ at 25°C)
BaTiO ₃ , SrTiO ₃ , and CaTiO ₃	($\epsilon_r = 2000$ at 25°C)
BaTiO ₃ and MgF ₂	($\epsilon_r = 3500$ at 25°C)
BaTiO ₃ , CaSnO ₃ , and CaO	($\epsilon_r = 3500$ at 25°C)
BaTiO ₃ and CaZrO ₃	($\epsilon_r = 5000$ at 25°C)
BaTiO ₃ and CaSnO ₃	($\epsilon_r = 8000$ at 25°C)

Furthermore, the locally inhomogeneous microstructure tends to broaden and flatten the $\epsilon_r - T$ curve. This tendency is enhanced if the growth of grain sizes during sintering can be sufficiently inhibited.

There is a large variety of ceramic capacitors, including thin-film capacitors, thick-film capacitors, and plate capacitors (based on the

thickness of the dielectric material), single-layer discrete capacitors in disk or wafer, or cylindrical tube form, or multilayer capacitors based on the construction shape. The design depends primarily on the requirements of the capacitors for certain specific applications. For more details about capacitors, see references 58, 59, and 95–98.

Thermo-Autostabilization Nonlinear Dielectric Elements (TANDEL)

When a ferroelectric specimen is switched on to the ferroelectric phase with an alternating field, heat will be produced in the material each cycle, part of which is due to normal dielectric losses and part to hysteresis losses. As a result, the power dissipation W_A in the material increases with temperature, reaches a peak at the transition temperature T_C , and then decreases rapidly for $T > T_C$, as shown in Figure 4-31. The value of W_A increases also with increasing applied AC field. There is heat W_B lost by the material through heat conduction and convection to the surrounding medium, which is proportional to temperature. Thus, at the intersection points (such as a , b , c), when the W_A - T curve intersects with the W_B - T curve, the heat produced in the material is equal to the heat lost from the material to its surrounding medium. So the points a , b , c

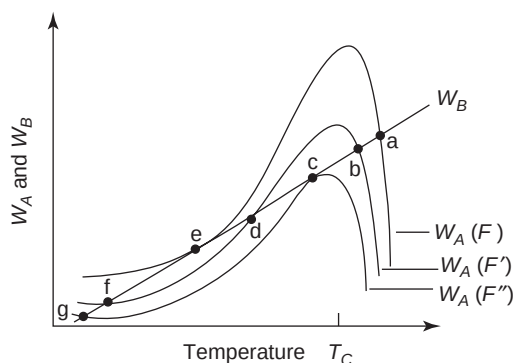


Figure 4-31 The variation of the power dissipated in the ferroelectric specimen W_A and the power lost by the specimen to its surroundings W_B with temperature at the applied AC field $F > F' > F''$.

represent the points of thermal equilibrium for ferroelectric materials. The material itself can stabilize its own temperature, such as point b near the transition temperature T_C where non-linearity is high.^{99–101} The elements based on the changes in the dielectric constant, and hence the capacitance with the change of the applied AC field, can be used to replace varactors as circuit elements in frequency modulation, thermostat control, etc.

High-Energy Electrical Pulse Generators

Ferroelectric ceramics possess a special feature of high spontaneous polarization and high dielectric constant, the latter being typically several hundred to several thousand, much larger than for organic solid insulators (usually below 10) and for inorganic solid insulators (below 20). Ferroelectric ceramics have already been used widely for energy storage capacitors.¹⁰² Ferroelectric ceramics also have a large piezoelectric constant, implying that the electric field produced per unit mechanical stress could be very large. For example, a compression stress of about 0.05 GPa would generate an open-circuit field of about $5\text{--}15\text{ kV cm}^{-1}$.⁴⁴ The direction of the applied stress for such a piezoelectrically generated high voltage is parallel to the poling direction, and the depoling process is based on the linear piezoelectric effect. However, exploring piezoelectrically generated electric power beyond the linear limit of the linear piezoelectric effect, Neilson¹⁰³ was the first to suggest that the remanent (or spontaneous) polarization of a poled ferroelectric material could be released by a shock wave of sufficiently high mechanical stress to cause the transition from the ferroelectric phase to an antiferroelectric one, thus resulting in the generation of high-energy voltage or current pulses. Since then, several investigators^{62,63,104,105} have studied further this method for the generation of high-energy power supplies actuated by a shock wave of mechanical compression through a poled ferroelectric ceramic material.

The basic principle is simple. It is important to choose a material with its ferroelectric phase

very close to the boundary between the ferroelectric phase and the antiferroelectric phase. The phase diagram of the PZT system shown in Figure 4-19 shows that the PZT 95/5 material with $x = 0.05$ in the composition $\text{PbZr}_{1-x}\text{Ti}_x\text{O}_3$ should be suitable for this application. This material is in the ferroelectric phase (F) with the rhombohedral structure between 50°C and 220°C . It becomes antiferroelectric (AF) with an orthorhombic structure at temperature below 50°C . There is a phase boundary between the F and AF phases that is temperature dependent. However, external electrical field tends to broaden the ferroelectric phase, implying that the F - AF phase boundary tends to move toward the left side. This means that external electric field may convert the original antiferroelectric phase to a ferroelectric phase. A mechanical (compression) stress tends to extend the antiferroelectric phase region, implying that the F - AF phase boundary tends to move to the right side. This means that a mechanical shock wave may convert the original ferroelectric phase to an antiferroelectric phase.

A poled specimen is a specimen that has been polarized by a sufficiently high electric field F until the spontaneous polarization reaches the saturation value P_p (see Figure 4-4). Usually, the polarizing field is kept across the specimen for a period of time to ensure that all of the spontaneously polarized domains are aligned in the direction of the field before the field is

reduced to zero and then removed from the specimen. Thus, the poled specimen has a remanent polarization P_r with a bound charge Q_b on the electrodes to balance the polarization charge in the specimen, thus keeping the whole system neutral, as shown in Figure 4-32(a). If a shock wave of compressive stress is now applied to the specimen in the direction perpendicular to the direction of the polarization, then the original ferroelectric phase will change to an antiferroelectric phase, in which $P_r \rightarrow 0$, and the bound charge Q_b on the electrodes will become free charge Q , ready to flow out of the system, as shown in Figure 4-32(b).

Based on this principle, we can generate high-energy electric pulses by a mechanical shock wave. Let us consider a simple mechanical-electrical energy conversion system, as shown in Figure 4-33(a), with three sides x_o , y_o , and z_o . The shock wave of the compressive stress travels in the z direction, perpendicular to the poling direction along the x -axis, with the velocity u . In this case, the operation is in the normal mode. It takes $\tau = z_o/u$ second for the wavefront to travel across the specimen of width z_o . So, when the wave front has traveled for time t , the portion of specimen up to the dashed line in Figure 4-33(a) has already been depoled. The equivalent circuit of the system is shown in Figure 4-33(b). The system can be considered a current generator with an internal capacitance C_o , resistance R_o , and inductance L_o . The value of L_o depends on the amount of

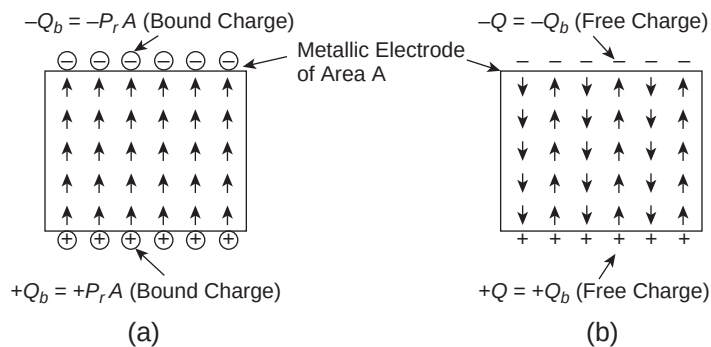


Figure 4-32 (a) Poled ferroelectric ceramic specimen with a bound charge Q_b on the metallic electrodes. (b) After the impact of a mechanical shock wave, the original ferroelectric phase changes to an antiferroelectric phase, and the bound charge on the electrodes is released and becomes a free charge.

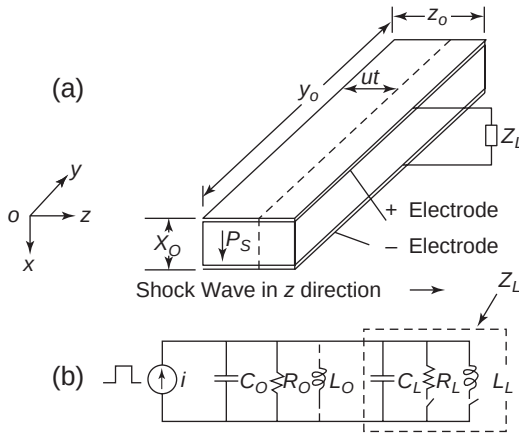


Figure 4-33 (a) Schematic diagram of the mechanical-electrical energy conversion system and (b) the equivalent circuit.

leakage current flowing through the specimen. The external load impedance may consist of capacitance C_L alone, C_L in parallel with resistance R_L , or C_L in parallel with inductance L_L . However, it is important to have C_L always connected to the circuit to protect the system in case breakdown occurs during the phase transition.

For a simple case, we can assume $R_0 = \infty$, $L_0 = \infty$, Z_L consists of C_L only and the permittivity of the specimen remains the same just after depoling. Then we have

$$C_0 = \epsilon y_0 z_0 / x_0 \quad 0 \leq t \leq \tau \quad (4-58)$$

The charge released for $t = 0$ to $t = \tau$ is

$$Q(t) = q(t) \text{ in the system} + q_L(t) \text{ in the load} \quad (4-59)$$

and the voltage between the electrodes is

$$\begin{aligned} V(t) &= q(t)/C_0 = q_L(t)/C_L \\ &= Q(t)/(C_0 + C_L) \end{aligned} \quad (4-60)$$

Since

$$\begin{aligned} Q(t) &= P_r A(t/\tau) = P_r y_0 z_0 (t/\tau) \quad 0 \leq t \leq \tau \\ &= P_r A = P_r y_0 z_0 \quad t > \tau \end{aligned} \quad (4-61)$$

so

$$\begin{aligned} V(t) &= P_r A(t/\tau)/(C_0 + C_L) \quad 0 \leq t \leq \tau \\ &= P_r A/(C_0 + C_L) \quad t > \tau \end{aligned} \quad (4-62)$$

The current $i(t)$ is thus

$$\begin{aligned} i(t) &= dQ(t)/dt = P_r A/\tau \quad 0 \leq t \leq \tau \\ &= 0 \quad t > \tau \end{aligned} \quad (4-63)$$

The variation of V and i with time is shown in Figure 4-34 (solid-line curves). However, the ideal case never occurs experimentally because there is always leakage current flowing through the specimen. So $Q(t)$ should include a loss factor, and thus $Q(t)$ should be expressed as

$$\begin{aligned} Q(t) &= P_r A(t/\tau)/(1 - \alpha t) \quad 0 \leq t \leq \tau \\ &= P_r A(1 - \alpha \tau) \quad t > \tau \end{aligned} \quad (4-64)$$

The variation of V and i will change accordingly. The dashed-line curves shown in Figure 4-34 represent V and i for a nonideal case.

We carried out a simple experiment to demonstrate how this type of high-energy pulse generator would perform. We used PZT with the composition of $\text{PbZr}_{0.975}\text{Ti}_{0.025}\text{O}_3$ plus 1 wt% Nb_2O_5 , because this material has the boundary between the F and AF phases at about room temperature. The phase diagram of this is shown in Figure 4-35(a). Its general properties are as follows⁶³:

- Density = 7.6 g cm^{-3}
- Dielectric constant = 1600 (high field)
- Loss factor $\tan \delta = 3 \times 10^{-2}$
- Curie temperature $T_c = 215^\circ\text{C}$

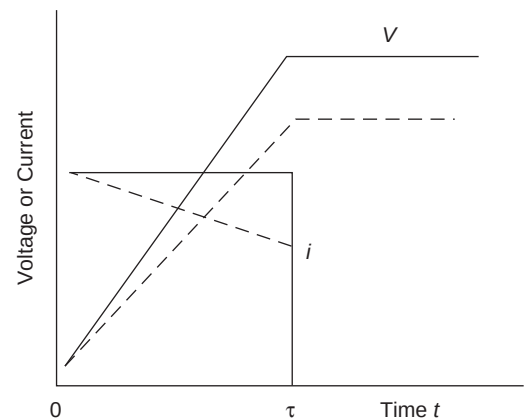


Figure 4-34 V and i as functions of time. Solid-line curves are for the ideal case and the dashed line curves for a nonideal case.

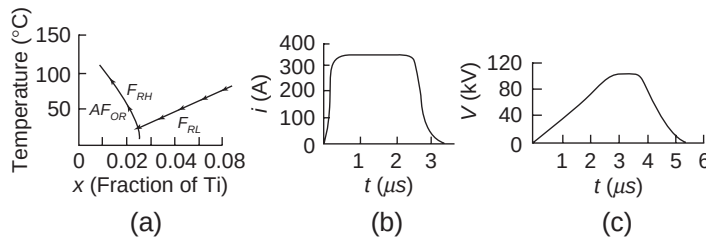


Figure 4-35 (a) Phase diagram of the $\text{PbZr}_{1-x}\text{Ti}_x\text{O}_3 + 1 \text{ Wt\% Nb}_2\text{O}_5$ system. (b) Short-circuit current-time curve. (c) Voltage-time curve.

- Electromechanical coupling factor $k_p = 0.1$
- Piezoelectric coefficient $d_{33} = 7 \times 10^{-12}$ coulomb/newton

The shock wave of the compressive stress was produced by an explosive system consisting of an initiating detonator, a plane wave generator, and a spacer to control the magnitude of the stress. The stress used for this investigation was about 3 Gpa, the shock wave velocity was $0.385 \text{ cm } \mu\text{s}^{-1}$, and the width of the wave was much longer than τ . This compressive stress was sufficient to cause all remanent polarization (bound charge) to be released. Some typical results are briefly described as follows.

Short-Circuit Current

We used specimens with $x_o = 0.5 \text{ cm}$, $y_o = 28 \text{ cm}$, and $z_o = 1 \text{ cm}$, so $\tau = z_o/u = 2.6 \mu\text{s}$. Since $P_r = 32.3 \mu\text{C cm}^{-2}$, the total charge on the electrode is $Q_b = P_r y_o z_o = 904.4 \mu\text{C}$. For this system, $C_o = \epsilon_r \epsilon_o y_o z_o x_o^{-1} = 8.4 \times 10^{-9} \text{ F}$. Thus, the charge $Q(t) = Q_b t \tau^{-1}$ and the current $i = dQ/dt = Q_b \tau^{-1}$ for $0 < t < \tau$, and $i = 0$ for $t \geq \tau$. The result is shown in Figure 4-35(b). The experimental value of 337 A for i is close to the calculated value of $904.4/2.6 = 348 \text{ A}$.

Voltage across Load Impedance Consisting of C_L and R_L

For this experiment we used specimens with $x_o = 4 \text{ cm}$, $y_o = 4 \text{ cm}$, and $z_o = 1 \text{ cm}$. In this case, $C_o = 1.5 \times 10^{-10} \text{ F}$, we used $C_L = 9.75 \times 10^{-10} \text{ F}$ and $R_L = 39 \text{ k}\Omega$. The result is shown in Figure 4-35(c). The experimental value of V across R_L

is 107 kV, which again is very close to the calculated value of 111 kV.

It should be noted that most ceramics are not good insulators. They consist of many pores and other defects, and generally they have a low resistivity and low breakdown strength. During depoling, a high voltage would be developed in the depoled section, which may enhance filamentary carrier injection from the electrodes, leading to local heating in certain regions. All of these factors may operate simultaneously to limit and affect the performance of the system.

Memories

The bistable polarization of ferroelectrics can be used for binary memory systems in the same way as the bistable magnetization of ferromagnetics. The memory is nonvolatile and does not require a holding voltage. The idea of using ferroelectric bistable polarization for memories was first suggested by Anderson.¹⁰⁶

The basic principle is simple. A ferroelectric matrix store is shown schematically in Figure 4-36 (a). The row and column electrodes are on opposite surfaces of the ferroelectric specimen. Thus, there are a number of square portions of the specimen having electrodes on both surfaces; each portion is one cell of the memory. The cell may have two remanent polarization states ($+P_r$ or $-P_r$). To write can be achieved by applying an electric field pulse F_x to the row electrode and a pulse F_y to the column electrode, both additively to produce a polarization of full magnitude $2P_r$ in the cell, and so it is dressed. To read, similar pulses F_x and F_y (but

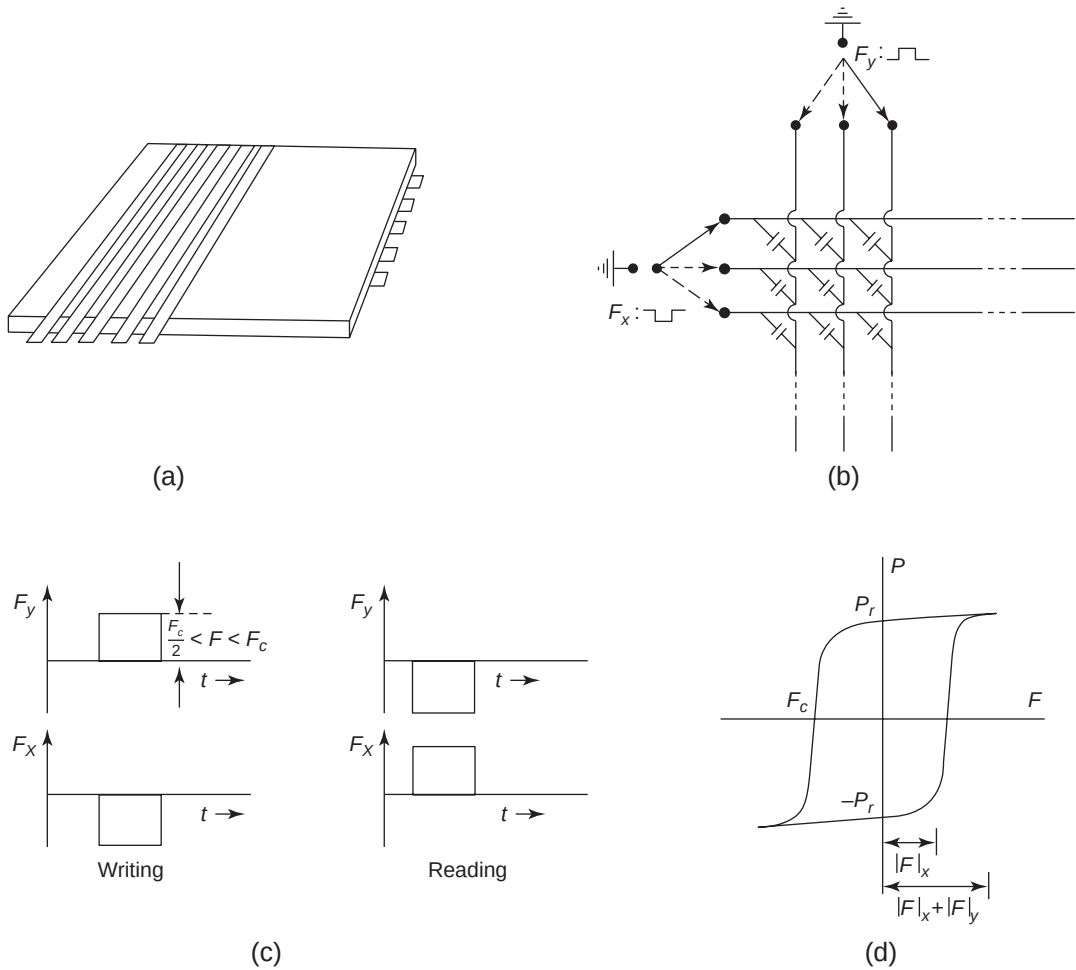


Figure 4-36 (a) Electrode arrangement for a ferroelectric matrix store, (b) electrical circuit for the application of voltages (or electric field) pulses to the electrodes, (c) the polarities of electric field pulses for *Writing* and for *Reading*, and (d) the total polarization $2P_r$ attained by simultaneous application of F_x and F_y pulses additively.

with the polarities opposite to those for writing) are applied to the electrodes to depole the polarized cell, so the stored charge $2P_r$ will be released generating a current or a voltage pulse across a registered resistor.

Figure 4-36 (b) through (d) illustrates this writing and reading process. It can be seen that the cells receiving a single pulse F_x or F_y are hardly affected; the two pulses must be coincident for both the writing and the reading. It should be noted that for a nondestructive readout, arrangements must be made to restore the original state of a cell after reading it. Fer-

roelectric stores are similar to ferrite stores, but the former have some major shortcomings, mainly the low switching speed, the decay of the stored signal with repeated uses due to the inherent fatigue behavior, and the aging effect.

Many suggestions have been put forward to deal with these shortcomings.^{107–110} For example, the combination of thin ferroelectric film technology with silicon-based memory technology offers the potential of combining nonvolatility with the fast read and write characteristics of dynamic random access memory (DRAM) and the small cell size of the

memory.^{108,110} For certain applications, such as for satellites, the radiation hardness of ferroelectric materials is a very attractive feature.^{108,109} Depending on the requirements of the particular application, ferroelectric memories have their own merit and deserve further development. By comparison, however, it appears that in some respects ferroelectric memories cannot compete with their well established magnetic and semiconductor counterparts. As a result, attention has been shifted to optical memories. For some excellent literature on optical memories, see references.^{98,111–114}

4.3 Piezoelectric Phenomena

Crystals formed by polar molecules with a non-centrosymmetric structure (i.e., without a center of symmetry) will exhibit a piezoelectric effect. This implies that a mechanical stress applied to the crystal specimen will create an overall polarization and hence a voltage across it. The reverse of the stress direction will cause the reverse of the polarity of the polarization and hence the voltage. The piezoelectric effect is convertible. This means that an applied electric field will create a mechanical strain, expansion, or contraction, depending on the direction of the field.

In general, an applied electric field always causes mechanical distortion in the geometric shape of the material, because matter is constructed of charged nuclei surrounded by a compensating electron cloud. The polarization induced by the applied field will cause changes in charge distribution and hence mechanical distortion. The strain resulting from the mechanical distortion is proportional to the square of the field, and this phenomenon, called *electrostriction*, is not inversive. This means that a mechanical stress acting on mass points cannot induce dipole moments from the neutral state of the material. Electrostriction occurs in all materials, although its effect, for most practical cases, is extremely or negligibly small.

The direct effect of the piezoelectricity is the generation of electric polarization by a mechanical stress (acting like a generator), while the converse or inverse effect is the mechanical

movement actuated by an electric field (acting like a motor). There are two principal mechanisms for piezoelectricity. Based on the first mechanism, the dipole moments may just mutually cancel each other in the material under the unstrained condition. Piezoelectricity may occur if the crystal has no center of symmetry, and in this case the relation between the electric field and the mechanical strain is linear in the first approximation. This linear relation is sometimes referred to as the *linear piezoelectric effect*. However, based on the second mechanism, the dipole moment components may remain, but they add to a resulting moment along a polar axis of the unit cell, so the occurrence of piezoelectricity is accompanied by pyroelectricity, involving spontaneous polarization. Thus, for ferroelectric piezoelectricity, the variation of the mechanical strain with the applied electric field follows the change of polarization in the hysteresis loop, as shown in Figure 4-37. During the poling process, there is a small expansion of the material along the poling direction and a contraction in directions (lateral directions) perpendicular to it. So the strain along the poling direction is positive. In Figure 4-37, the solid line traces the poling strain from a virgin state 0 to saturation state C and remanent state D. The lateral strain is negative. Both the poling and the lateral strain-field relations form a butterfly-shaped hysteresis. Figure 4-37 shows the basic differences among electrostriction, linear piezoelectricity, and ferroelectric piezoelectricity.

In general, ferroelectric piezoelectrics, such as ceramics, have advantages for use in transducers because they have a large piezoelectric coefficient, especially near the transition temperature, and a high dielectric constant that allows the electro-mechanical coupling factor to approach unity. In comparison with non-ferroelectric materials, ferroelectrics may have a high coupling constant, but they also have high dielectric losses.

4.3.1 Phenomenological Approach to Piezoelectric Effects

Thermodynamically reversible interactions occur among the electrical, mechanical, and

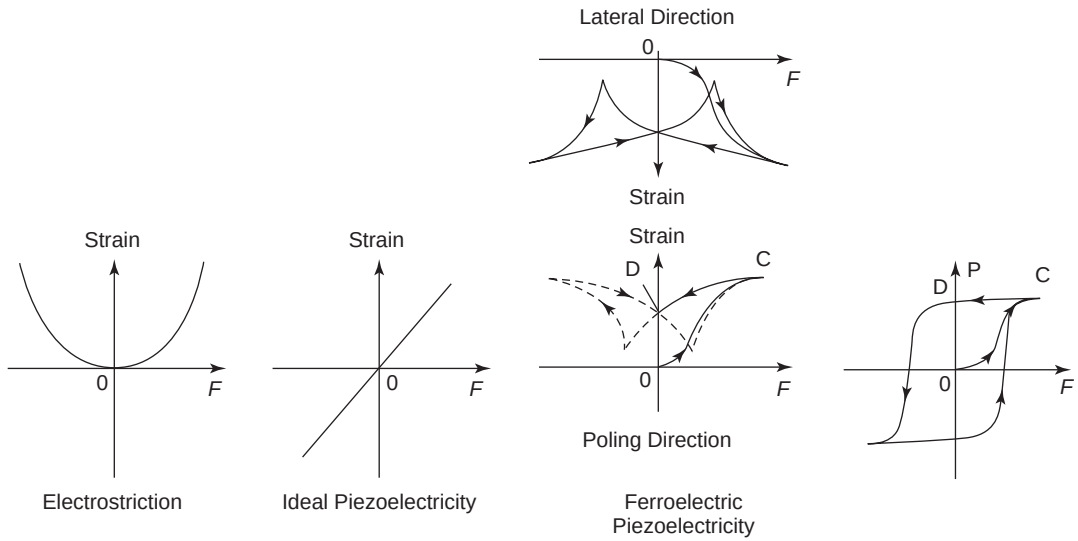


Figure 4-37 Schematic illustration of the variation of mechanical strain with electric field for electrostriction, ideal piezoelectricity, and ferroelectric piezoelectricity in poling and lateral directions.

thermal properties, as illustrated in Figure 4-38. The lines joining pairs of circles indicate that a small change in one of the variables causes a corresponding change in the others. The three direct relations are as follows.

The relation between electric field F and displacement D defines the permittivity ϵ

$$D = \epsilon F \quad (4-65)$$

The relation between mechanical stress X and strain h defines the elastic compliance or elastic constant y ($y = 1/Y$, where Y is the modulus of elasticity or the elastic stiffness coefficients).

$$h = yX \quad (4-66)$$

The relation between temperature T and entropy S defines the heat capacity Q

$$S = QT^{-1} \quad (4-67)$$

The lines joining the circles at different corners define the coupling effects. For ordinary solids, the properties follow Equations 4-65 through 4-67. But for piezoelectric materials, additional terms are required. For example, a mechanical stress causes not only a strain, but also electric polarization, even at a constant temperature.

Thus, the displacement induced by the stress can be expressed as

$$D = dX \quad (4-68)$$

where d is one of the piezoelectric constants (or coefficients) whose unit is coulombs/newton.

In the converse piezoelectric effect, we can also express the relation between electric field and the strain as

$$h = dF \quad (4-69)$$

In this case, the unit for d is meters/volt. In fact, for both the piezoelectric and the converse piezoelectric effects, d is identical, so

$$d = \frac{D}{X} = \frac{h}{F} \quad (4-70)$$

In terms of the electric field produced by a mechanical stress, the relation becomes

$$F = gX \quad (4-71)$$

In this case, g is another one of the piezoelectric constants. Its unit is [volts/meter]/[newtons/meter²]. It can be expressed as

$$g = \frac{d}{\epsilon} = \frac{d}{\epsilon_r \epsilon_0} \quad (4-72)$$

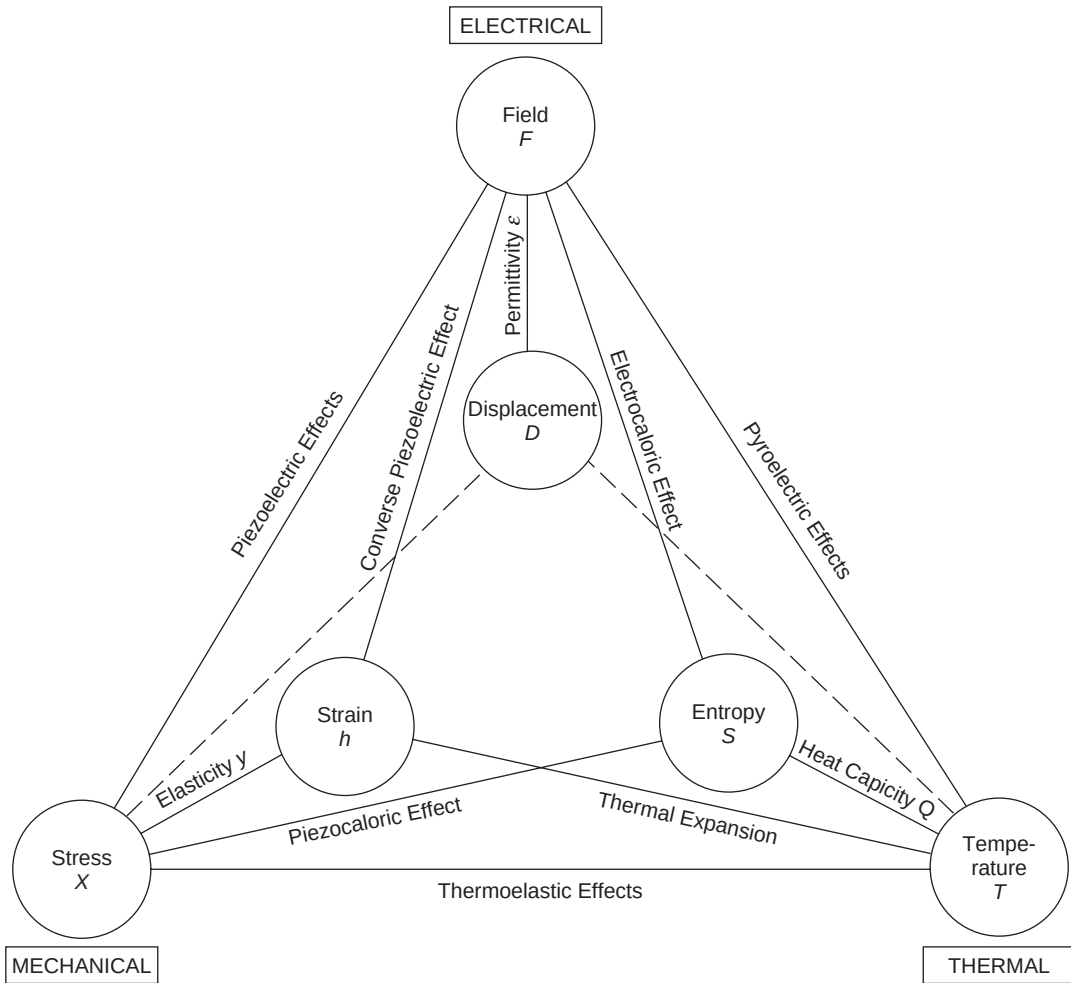


Figure 4-38 The relations between the mechanical, electrical, and thermal properties for crystals with a noncentrosymmetric structure (without a center of symmetry).

There are two other piezoelectric constants, namely e and f , which are only occasionally used. Their relation is

$$X = -eF \quad (4-73)$$

$$F = -fh \quad (4-74)$$

Rigorous analysis and discussion of these piezoelectric constants (or coefficients) has been reported by Mason¹¹⁵ and Jaffe and Berlincourt.⁵³ These four piezoelectric constants can be defined in partial derivatives as

$$\begin{aligned} d &= \left(\frac{\partial D}{\partial X} \right)_{F,T} \\ g &= - \left(\frac{\partial F}{\partial X} \right)_{D,T} \\ e &= \left(\frac{\partial D}{\partial h} \right)_{F,T} \\ f &= - \left(\frac{\partial F}{\partial h} \right)_{D,T} \end{aligned} \quad (4-75)$$

Similarly, the piezoelectric constants for the converse piezoelectric effects can be written as

$$\begin{aligned}
d' &= \left(\frac{\partial h}{\partial F} \right)_{X,T} \\
g' &= \left(\frac{\partial h}{\partial D} \right)_{X,T} \\
e' &= - \left(\frac{\partial X}{\partial F} \right)_{h,T} \\
f' &= - \left(\frac{\partial X}{\partial D} \right)_{h,T}
\end{aligned} \tag{4-76}$$

Based on the thermodynamic argument, $d = d'$, $g = g'$, $e = e'$, and $f = f'$.

The interactions between the electrical and the elastic variables describe the piezoelectric effects. The equations of state relating the electrical and the elastic variables can be written in general form as

$$D = dX + \epsilon^X F \tag{4-77}$$

$$h = y^F X + dF \tag{4-78}$$

where the superscripts denote the parameters held constant, so ϵ^X denotes the permittivity at constant stress X and y^F denotes the elastic compliance at constant electric field F . For the ferroelectric piezoelectricity, the properties are nonlinear, so the equations of state should be written in differential form as

$$\delta D = d\delta X + \epsilon^X \delta F \tag{4-79}$$

$$\delta h = y^F \delta X + d\delta F \tag{4-80}$$

The piezoelectric coefficients are interdependent. For example,

$$\begin{aligned}
\frac{d}{g} &= \epsilon^X \\
\frac{e}{f} &= \epsilon^X
\end{aligned} \tag{4-81}$$

To specify the overall strength of the piezoelectric effect, it is convenient to use the electro-mechanical coupling factor k , which can be considered the direct way to measure the ability of piezoelectric materials to convert one form of energy to another. This coupling factor k is defined as

$$k^2 = \frac{\text{Electrical energy from conversion}}{\text{Total input mechanical energy}} \tag{4-82}$$

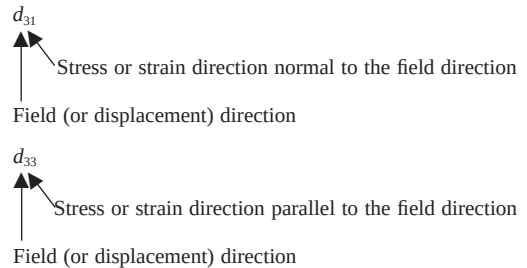
resulting from the direct piezoelectric effect, or

$$k^2 = \frac{\text{Mechanical energy from conversion}}{\text{Total input electrical energy}} \tag{4-83}$$

resulting from the converse piezoelectric effect. Since the electro-mechanical conversion is always incomplete, k is always less than unity. Typical values of k are 0.1 for quartz, 0.4 for barium titanate, 0.7 for PZT, and close to 0.9 for Rochelle salt.

In general, piezoelectric properties are dependent on orientational direction, so they must be described in terms of tensors. A convenient way to specify the directional properties is to use subscripts that define the direction and orientation, as illustrated in Figure 4-39. The subscript 3 refers to the polar axis (or poling axis); 1 and 2 refer to arbitrarily chosen orthogonal axes perpendicular to 3. Subscripts 4, 5, and 6 refer the shear planes of the mechanical stress and strain perpendicular to axes 1, 2, and 3, respectively. For example, subscript 4 indicates the change of angle of the stress or the strain between the two initially orthogonal axes 2 and 3 in the shear plane (subscript 4) normal to axis 1. Similar meaning is applied to subscripts 5 and 6. Stress and strain due to shearing action are referred to as the *shear stress* and *shear strain*.

Piezoelectric coefficients are usually indicated with two subscripts denoting the direction of the properties. The first subscript refers to the direction of the electric field F (or the displacement D). The second subscript refers to the direction of the mechanical stress X (or the strain h). For example,



In most cases, it can be assumed that the plane perpendicular to the polar axis is isotropic. This

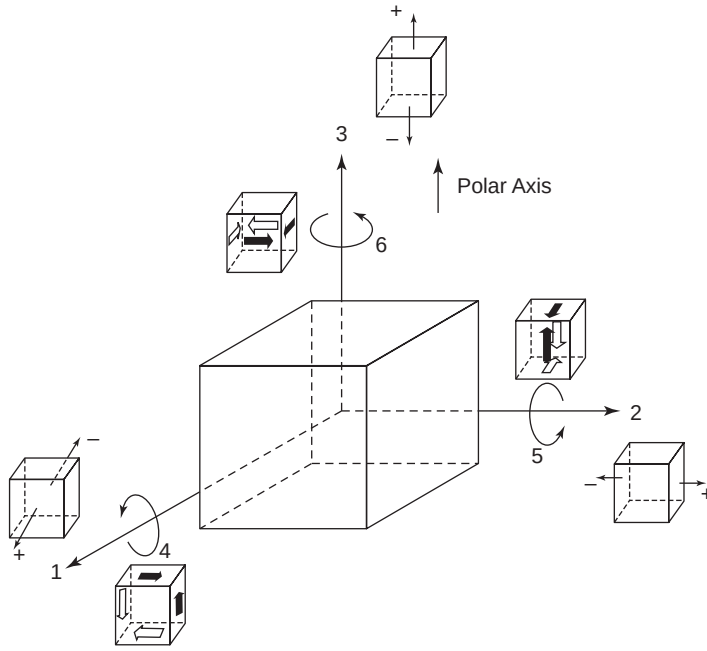


Figure 4-39 Subscript symbols for the notation of the directional properties for poled piezoelectric materials.

implies that an electric field parallel to polar axis 3 interacts in the same way, with the mechanical stress either along axis 1 or along axis 2. Thus, $d_{31} = d_{32}$, $d_{15} = d_{24}$, $y_{13} = y_{23}$, etc. It should be noted that shear can only occur when an electric field is applied at right angle to the polar axis, so there is only one shear piezoelectric coefficient d_{15} (since $d_{15} = d_{24}$). There are also piezoelectric coefficients corresponding to hydrostatic stress. In this case, $d_h = d_{33} + d_{31} + d_{32} = d_{33} + 2d_{31}$. Similar notations are used for other piezoelectric coefficients, such as g_{33} , g_{31} , g_{15} , g_h , etc.

Elastic properties can also be specified in terms of elastic compliance y , such as y_{11} , y_{12} , y_{13} , y_{33} , y_{44} , y_{66} , to denote the interaction of a strain and a stress under the condition of constant field. Thus y should be written as y_{11}^F , y_{12}^F , y_{13}^F , y_{33}^F , y_{44}^F , y_{66}^F (usually for $F = 0$). Each y relates to the application of a single stress in one direction; the other y s are kept fixed so that there is no lateral restraint. Stress and strain are interchangeable, so $y_{13} = y_{31}$. The permittivity ϵ should also be a tensor, but for most cases, the field has only one direction, oblique

components being negligible, so one subscript is sufficient. For example, $\epsilon_{33} = \epsilon_3$, $\epsilon_{31} = \epsilon_1$, $\epsilon_{32} = \epsilon_2$, etc. indicating that ϵ_3 is the permittivity along the polar axis 3, and ϵ_1 and ϵ_2 are the permittivities along axes 1 and 2, respectively.

Following the above conventions, Equation 4-81 becomes

$$\begin{aligned} \frac{d_{ij}}{g_{ij}} &= \frac{(\partial h_j / \partial F_i)_{x,T}}{(\partial h_j / \partial D_i)_{x,T}} \\ &= \left(\frac{\partial D_i}{\partial F_i} \right)_{x,T} = \epsilon_i^x \end{aligned} \quad (4-84)$$

For example,

$$\begin{aligned} \frac{d_{33}}{g_{33}} &= \epsilon_{33}^x = \epsilon_3^x \\ \frac{d_{31}}{g_{31}} &= \epsilon_{31}^x = \epsilon_1^x \end{aligned}$$

Thus, Equation 4-79, due to the direct piezoelectric effect, becomes

$$\begin{aligned} \delta D_1 &= d_{15} \delta X_3 + \epsilon_1^x \delta F_1 \\ \delta D_2 &= d_{15} \delta X_4 + \epsilon_2^x \delta F_2 \\ \delta D_3 &= d_{31} (\delta X_1 + \delta X_2) + d_{33} \delta X_3 + \epsilon_3^x \delta F_3 \end{aligned} \quad (4-85)$$

Equation 4-80, due to the converse piezoelectric effect, becomes

$$\begin{aligned}
 \delta h_1 &= y_{11}^F \delta X_1 + y_{12}^F \delta X_2 + y_{13}^F \delta X_3 + d_{31} \delta F_3 \\
 \delta h_2 &= y_{11}^F \delta X_2 + y_{12}^F \delta X_1 + y_{13}^F \delta X_3 + d_{32} \delta F_3 \\
 \delta h_3 &= y_{13}^F (\delta X_1 + \delta X_2) + y_{33}^F \delta X_3 + d_{33} \delta F_3 \\
 \delta h_4 &= y_{44}^F \delta X_4 + d_{15} \delta F_2 \\
 \delta h_5 &= y_{44}^F \delta X_5 + d_{15} \delta F_1 \\
 \delta h_6 &= y_{66}^F \delta X_6
 \end{aligned} \quad (4-86)$$

4.3.2 Piezoelectric Parameters and Their Measurements

The best-known crystal showing a piezoelectric effect is quartz (SiO_2). Lord Kelvin¹¹⁵ was the first to develop a model to explain this phenomenon. For a quantitative analysis of the piezoelectric effect, it is necessary to know first the piezoelectric and dielectric properties of the material. In the previous section, it was shown that piezoelectric properties are primarily governed by the electro-mechanical coupling coefficient k ; the piezoelectric coefficients, mainly d and g ; the elastic compliance y ; and the dielectric constant ϵ_r . These parameters are generally dependent on the crystal structure and the direction of the crystal axes, so it is important to know these parameters. The convenient way to determine these parameters is to use an equivalent circuit which consists of two portions: one involving an analog between mechanical and electrical quantities representing mechanical impedance, and the other representing the electrical impedance.^{4,116-118} The analog between mechanical and electrical quantities follows.

Mechanical Quantity	Electrical Quantity
Elastic compliance: y	Capacitance: C_m
Mechanical mass: m	Inductance: L_m
Viscous damping: Γ	Resistance: R_m

First, we measure the resonance frequencies of the piezoelectric specimen under investigation with a sinusoidally varying electric field, as shown in Figure 4-40(a). By comparing the phase of the AC voltage with that of the AC current, it is easy to determine the frequencies at which $X_s = X_L - X_c > 0$ or $X_s < 0$ or $X_s = 0$. The various frequencies are defined as follows:

f_R : The resonant frequency at which $X_s = 0$ ($X_L = X_c$); in this case, $X_L - X_c$ increases with frequency for frequencies below f_R

f_A : The antiresonant frequency at which $X_s = 0$ ($X_L = X_c$), but in this case, $X_L - X_c$ decreases with the frequency for frequencies around f_A

f_p : The frequency at which the resistive component R_s is a maximum

f_s : The frequency at which the series arm of the upper portion of Figure 4-40(b) has zero reactance ($X_m = 0$)

It can be seen from the equivalent circuit and its frequency response in Figure 4-40 that the piezoelectric specimen stores energy mechanically in one part of the cycle and feeds it back into the electrical portion in another part of the cycle. The specimen operates in sequence as an electro-mechanical transducer.

It can be shown from the equivalent circuit Figure 4-40(b) and (c) that

$$f_R = \frac{1}{2\pi\sqrt{L_m C_m}} \quad (4-87)$$

$$f_A = \frac{1}{2\pi} \sqrt{\frac{C_m + C_e}{L_m C_m C_e}} \quad (4-88)$$

The ratio of f_A/f_R is related to the electro-mechanical coupling factor by the following equation¹¹⁹

$$\left(\frac{f_A}{f_R}\right)^2 = 1 + \frac{8}{\pi^2} \left(\frac{k^2}{1 - k^2}\right) \quad (4-89)$$

Thus, from Equations 4-87 through 4-89, we obtain

$$k^2 = \frac{C_m}{C_e + C_m} \approx \frac{f_A^2 - f_R^2}{f_R^2} \quad (4-90)$$

Equation 4-90 gives good approximations, provided that the quality factor Q of the specimen is sufficiently high. Q is the ratio of the energy stored to the energy dissipated per half-cycle, which is equal to $(\tan \delta)^{-1}$. For most dielectric materials, $\tan \delta$ is very small. For example, when $\tan \delta < 0.01$, $Q > 100$. In Figure 4-40(b), we have already assumed $Q > 100$, so only the capacitance C_e is included in the electrical portion, the leakage resistance being ignored.

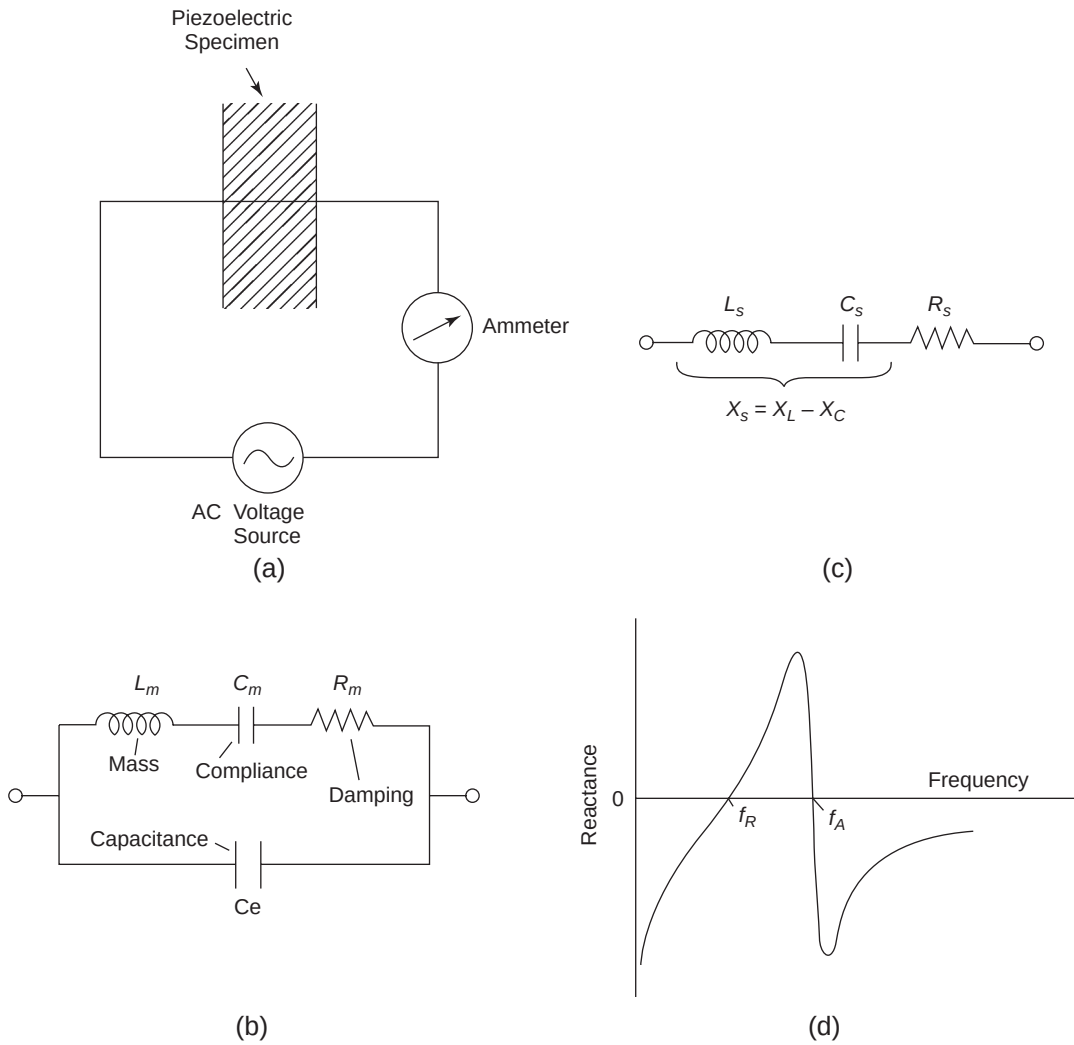


Figure 4-40 (a) A piezoelectric specimen under an alternating electric field, (b) equivalent circuit of (a) near the fundamental resonances (the top branch of the circuit represents the mechanical portion and the bottom branch represents the electrical portion), (c) the equivalent series circuit of the impedance of (b), and (d) frequency response near the fundamental resonances.

Once the parameter k has been determined, the other parameters can be determined easily. The mechanical energy stored in the specimen per unit volume is given by

$$U_m = \frac{1}{2} |X| |h| = \frac{1}{2} \frac{h^2}{y} \quad (4-91)$$

and the electrical energy stored per unit volume is

$$U_e = \frac{1}{2} |D| |F| = \frac{1}{2} \epsilon F^2 \quad (4-92)$$

Thus, we can write

$$k^2 = \frac{U_m}{U_e} = \frac{d^2}{y\epsilon} \quad (4-93)$$

The relations between f_R and f_A and the elastic compliance y are given by

$$f_R = \frac{1}{2\ell(\rho y^F)^{1/2}} \quad (4-94)$$

$$f_A = \frac{1}{2\ell(\rho y^D)^{1/2}} \quad (4-95)$$

where ρ is the density of the specimen material and ℓ is the length of the specimen. The superscript F in y^F means y measured at constant field, signifying that the specimen is short-circuited, while the superscript D refers to y at constant displacement, signifying that the specimen is open-circuited. The permittivity ϵ can be determined from the capacitance measured at a frequency well below resonance. Thus, from Equations 4-93 through 4-95, we can determine y , d , and g , since $g = d/\epsilon$. However, the piezoelectric parameters depend on the geometric shape and dimensions of the specimen.

The geometry commonly used is a thin disc of diameter dia and thickness t , which is much smaller than dia . Metallic electrodes are deposited on both surfaces of the specimen, and the specimen is poled in a direction perpendicular to the surfaces. In this case, the planar (radial) electromechanical coupling factor k_p is of a radial mode, excited through the piezoelectric effect across the thickness of the disc. k_p is related to the resonant frequencies f_p and f_s , defined by the following relation

$$\frac{k_p^2}{1-k_p^2} = F \left(J_0, J_1, \nu \frac{f_p - f_s}{f_s} \right) \quad (4-96)$$

where J_0 and J_1 are the Bessel functions of the orders of zero and one, respectively; and ν is Poisson's ratio. The other piezoelectric parameters d and g can be determined in the same manner as for the thick specimens, but the expressions for the thin disc are more complex.⁹⁶ For more details about piezoelectric parameters, see references 44, 59, 96, 120, 121.

4.3.3 Piezoelectric Materials

The most important piezoelectric materials are the ceramics consisting of crystallites of Perovskite structure, such as PZT and PLZT; the synthetic polymers, such as PVDF; and the ceramic-polymer composites. Table 4-4 lists the typical values of some important parameters of some commonly used piezoelectric materials. All ferroelectric materials are piezoelectric, but not all piezoelectric materials are ferroelectric. We will describe briefly

some special features of some piezoelectric materials.

Quartz is a commonly used piezoelectric crystal, but crystals of the water-soluble type have a much larger electromechanical coupling factor k . For example, ammonium dihydrogen phosphate (ADP) has $k \approx 0.30$. Still higher coupling can be achieved with ferroelectrics such as PZT (a PbTiO_3 and PbZrO_3 mixture). The PZT specimen, after being poled with a sufficiently strong electric field, exhibits high values of k_p and d depending on the composition, as shown in Figure 4-41. This PZT system has been studied by several investigators.^{44,120-124} Aliovalent substituents would modify the properties of ceramics with a Perovskite structure. In order to achieve a given set of properties, PZT-based ceramics may require composition of more than one type of additive. One system that has been widely used for piezoelectric devices is PZT doped with lanthanum to form a new PLZT system. Such systems usually contain 3–12 mol.% La and 5–30 mol.% Ti. For the system with the composition of La:Zr:Ti=7:60:40, the values of k_p and d_{33} have reached 0.72 and 7.1×10^{-10} m/V.^{125,126} Such high values of k_p and d_{33} have been attributed to the compositions located within the boundary region between the rhombohedral and tetragonal phases.¹²⁷⁻¹²⁹ PLZT may also contain vacancies on B (as

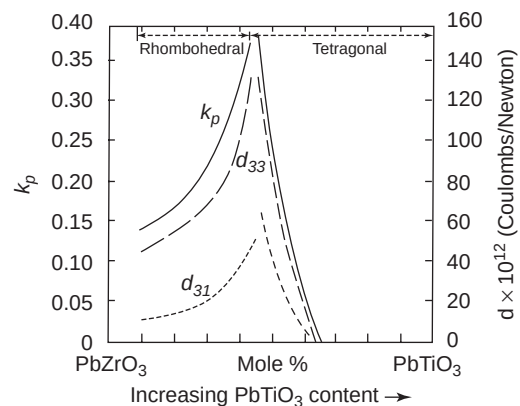


Figure 4-41 The planar electromechanical coupling factor k_p and piezoelectric coefficient d as functions of the composition of the PZT ceramics.

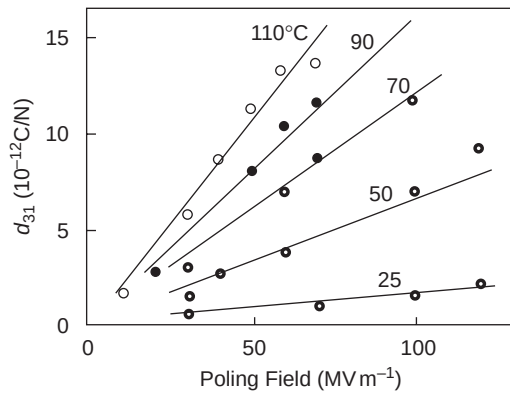
Table 4-4 Selected properties of some piezoelectric materials: d = piezoelectric constant, k = electro-mechanical coupling factor, and ρ = density.

Material (Abbreviation)	Chemical Formula	d (cm V ⁻¹)	k	ρ (g cm ⁻³)
Silicon Dioxide (Quartz)	SiO ₂	2.25×10^{-10}	0.099	2.65
Potassium-Sodium Tartrate-Tetrahydrate (Rochelle Salt)	KNaC ₄ H ₄ O ₆ •4H ₂ O	5.30×10^{-9}	0.78	1.77
Barium Titanate	BaTiO ₃	1.90×10^{-8}	0.38	5.70
Ammonium Dihydrogen Phosphate (ADP)	AH ₂ PO ₄	2.46×10^{-9}	0.29	1.80
Potassium Dihydrogen Phosphate (KDP)	KH ₂ PO ₄	1.07×10^{-9}	0.12	2.31
Triglycine Sulfate (TGS)	(NH ₂ CH ₂ COOH) ₃ •H ₂ SO ₄	5.00×10^{-9}	—	1.69
Potassium Niobate or Sodium Niobate	KNbO ₃ or NaNbO ₃	4.90×10^{-9}	0.42	4.45
Lithium Niobate	LiNbO ₃	0.85×10^{-10}	0.035	4.64
Lithium Tantalate (LT)	LiTaO ₃	3.00×10^{-10}	0.10	7.46
Lead Titanate	PbTiO ₃	7.40×10^{-10}	—	7.12
Lead Zirconate Titanate (PZT)	Pb(Zr _{1-x} Ti _x)O ₃	2.34×10^{-8}	0.66	7.70
Polyvinyl Chloride (PVC)	see Table 5-2	0.70×10^{-10}	—	1.40
Polyvinyl Fluoride (PVF)	see Table 5-2	1.00×10^{-10}	0.03	—
Polyvinylidene Fluoride (PVDF or Kynar)	see Table 5-2	4.00×10^{-10}	0.12	1.78

well as on A) sites, providing a favorable way to change their polar states under the influence of applied fields.

Polymer films have the following advantages: they are flexible and tough, they can be made very thin ($<10 \mu\text{m}$) and large in area, they can be shaped into any geometric forms, they have a low mechanical impedance, etc. Polymers, particularly PVDF, have been used widely for various transducers and sensors. For PVDF, the piezoelectric coefficient d and the pyroelectric coefficient p are closely related to the poling parameters. They increase linearly with increasing poling field F_p , poling temperature T_p , and poling time t_p , as shown in Figure 4-42. The data are from Murayama et al.^{130,131}

Apart from PVDF, other polymers, such as polyvinyl fluoride (PVF), polyvinyl chloride (PVC), PTFE, etc., also exhibit piezoelectric effects. PVC is a noncrystalline polymer but has relatively high piezoelectric and pyroelec-

**Figure 4-42** The piezoelectric coefficient d_{31} of PVDF at room temperature as a function of poling field and poling temperature for the fixed poling time of 30 minutes.

tric coefficients when poled at high temperatures. It should be noted that the properties of polymers depend strongly on the preparation techniques and conditions.¹³²

To produce materials with optimal properties for a special application, we may choose a suitable ceramic–polymer composite. In general, a composite refers to a mixture in which ferroelectric ceramic particles are dispersed in a polymer matrix. The properties are largely determined by the choice of components, their relative composition, and the manner in which they are interconnected. The composites made by mixing PZT ceramic, PVDF polymer, and fluorinated rubber have a good piezoelectric coefficient, depending on the composition.¹³³ Ceramic–polymer composites have been developed mainly to search for a suitable combination that can be adapted for piezoelectric or pyroelectric devices to be used in stringent environments, such as sonar devices in water, sensors for medical diagnostics, etc.

4.3.4 Applications of Piezoelectrics

Both the direct and converse piezoelectric effects can be used for many practical applications, such as high-voltage generators, cigarette lighters, gas igniters, etc.; transducers for high-intensity ultrasounds employed in ultrasonic cleaning, ultrasonic therapy, medical diagnosis, delay lines, etc.; resonators for filters and oscillators; microphones; various sensors; etc. The following sections use four simple and typical examples to demonstrate the importance of piezoelectric effects.

Gas Igniters

A schematic diagram of a poled piezoelectric ceramic igniter is shown in Figure 4-43, which is self-explanatory. The principle is simple. When a force is applied to the poled ceramic, a voltage will be generated between electrodes. It is usual to use two pieces of poled ceramic elements connected back to back in order to double the charges released for the spark. On releasing the force, a voltage of about the same magnitude, but of the opposite sign, will be produced, yielding another spark across the gas gap. If the force were applied too slowly, no spark would occur because the charges released

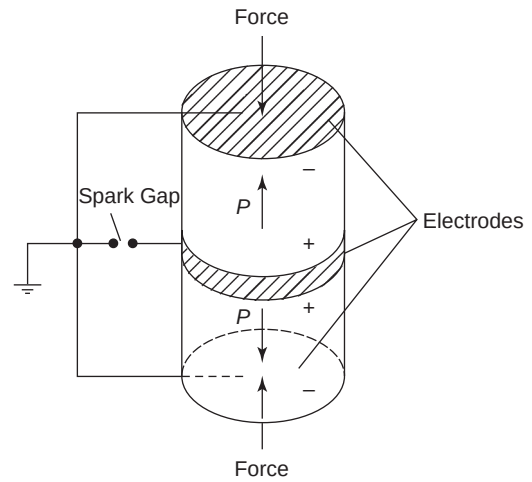


Figure 4-43 Schematic diagram illustrating the principle of a gas igniter.

gradually would leak away through the leakage across the electrode, instead of building up to generate a high voltage across the spark gap. This is why igniters are of the momentary impact type rather than the slow squeeze type.

Since most igniters have two pieces of poled ceramic elements connected back to back, we need consider only one for our analysis. Based on Equation 4-71, the voltage developed across the ceramic element when an impact force is applied is given by

$$V = F\ell = g_{33}\ell X = g_{33}\ell \frac{\mathcal{F}}{A} \quad (4-97)$$

where $\mathcal{F}$ is the force and ℓ and A are, respectively, the length and the area of the ceramic element. Ignoring all possible losses in the operation, the energy that may be dissipated in the spark is

$$W = \frac{1}{2} \epsilon_{33} \frac{A}{\ell} \left(g_{33} \ell \frac{\mathcal{F}}{A} \right)^2 = \frac{1}{2} d_{33} g_{33} \frac{\mathcal{F}^2 \ell}{A} \quad (4-98)$$

If two elements are connected back to back (in fact, they are connected in parallel), the total energy available for the spark is two times that shown in Equation 4-98. PZT and PLZT are commonly used for gas igniters.

Delay Lines

The transmission of a microwave signal can be delayed by the following steps:

1. The microwave signal is guided through a coupling network to a piezoelectric transducer, which converts it to an acoustic wave.
2. This acoustic wave then propagates in a dielectric medium and reflects when it reaches the end.
3. This reflected acoustic wave will be converted back to the original microwave signal by the transducer, as shown in Figure 4-44.

The major advantage of this type of delay line is the use of the transducer to convert the high-speed microwave signal to a low-speed acoustic wave, and of the dielectric medium to produce the time delay. Since the speed of an acoustic wave is about 10^5 times slower than an electromagnetic wave, this type of delay line can be made much smaller in size and weight than a conventional one.¹³⁴ However, for the design of such delay lines, it is important to consider the following factors:

- The dielectric medium must have a low acoustic attenuation for the acoustic wave to propagate.
- The transducer must have a high electro-mechanical coupling factor k and a high longitudinal mode acoustic wave velocity for its use at microwave frequencies.
- The acoustic impedance matching between the dielectric medium and the transducer must be good, and the insertion loss must be low.

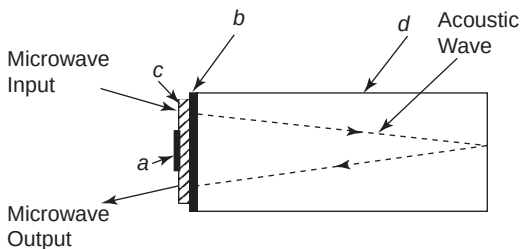


Figure 4-44 The structure of a microwave delay line; a : electrode A, b : electrode B, c : transducer, and d : dielectric medium.

Piezoelectric Positioners and Actuators

Piezoelectric positioners and actuators have been used widely in areas related to precision position controlling, vibration damping, relays, phonograph pickup, pressure sensing, etc.¹³⁵⁻¹³⁷

The basic principle is simple. A thin piezoelectric plate with metallic electrodes deposited on both surfaces can be used as a mechanical positioner or relay when the required movements are very small (e.g., a few micrometers or less).

Figure 4-45 shows schematically a poled rectangular piezoelectric plate with the following dimensions: length a , width b , and thickness c , which are in the x , y , and z directions, respectively. When a DC voltage V is applied across the thickness in the z direction (which is also the poled direction), then the plate will expand or contract in the z direction, and contract or expand in the x and y directions, depending on the polarity of the voltage with respect to the poled direction. The magnitude of such movements can be easily calculated. From Equation 4-70, we can write

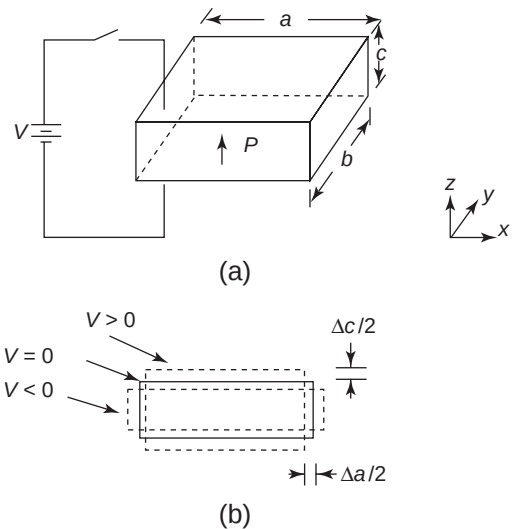


Figure 4-45 Schematic illustration of the deformation of a poled piezoelectric plate under an electric field: (a) the original shape of the plate at $V = 0$ and (b) the deformation at $V > 0$ and $V < 0$. (The deformation is exaggerated for clarity.)

$$d_{33} = \frac{h}{F} = \frac{\Delta c/c}{V/c} = \frac{\Delta c}{V} \quad (4-99)$$

or

$$\Delta c = d_{33}V$$

This indicates that the change in thickness (i.e., the movement) depends on the piezoelectric coefficient d_{33} and the applied voltage V , but is independent of the thickness.

For movements in the direction normal to the poled direction (for example, in the x direction), we have

$$d_{31} = \frac{h}{F} = \frac{\Delta a/a}{V/c} = \left(\frac{c}{a}\right) \frac{\Delta a}{V} \quad (4-100)$$

$$\Delta a = \left(\frac{a}{c}\right) d_{31}V$$

It can be seen that in this case, a can be made much larger than c if $a \gg c$. For a very small adjustment in the position, we may use a small and more precise movement in the direction parallel to the poled direction. But for a larger movement, such as for relays, sideways movement in the direction normal to the poled direction may be preferable.

Elastic compliance can be greatly increased by using long, thin piezoelectric strips or plates and mounting them as cantilevers. Bending such a strip or plate causes one half to stretch and the other half to compress, so there can be no net electrical response. However, if there are two such strips or plates with an intervening electrode and electrodes on outer surfaces to form a bimorph, as shown in Figure 4-46, then the bending will generate a voltage between the outer electrodes. By the converse piezoelectric effect, a bimorph will bend when a voltage is applied between the outer electrodes. The bending is governed by $d' = \left(\frac{\partial h}{\partial F}\right)$ (see Equation 4-76). The applied field will also produce uniform strains along the length of the bimorph so that it will be in the form of a circular arc. Detailed analysis of the behavior of the bimorph under static conditions has been done by several investigators.^{138,139} The reflection at the free end of the bimorph cantilever is given by

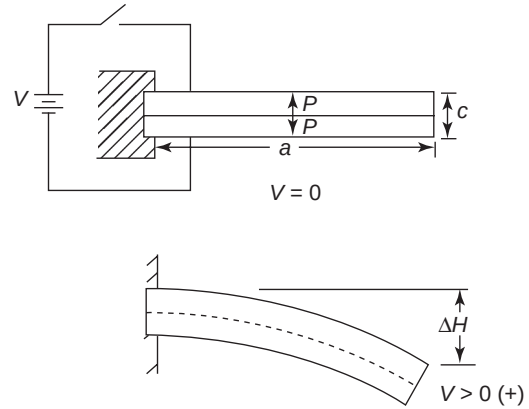


Figure 4-46 Schematic illustration of the bending of a bimorph cantilever-piezoelectric actuator. (The bending magnitude is highly exaggerated for clarity.)

$$\Delta H = \frac{3}{2} \left(\frac{a^2}{c^2} \right) d_{31}V \quad (4-101)$$

In this case, ΔH is proportional to $(a/c)^2$. This is why bending actuators are widely used for the applications requiring large movements, because they provide the largest movement (deflections) for a given applied voltage. It should be noted that bending bimorphs are not suitable for some applications because they have high mechanical inertia and cannot produce significant forces.

Piezoelectric Transformers

A transformer can be formed simply by a piezoelectric plate with electrodes deposited on half of its two flat surfaces and an electrode on the edge of the plate, as shown in Figure 4-47(a). One half of the plate is poled in the direction perpendicular to the surfaces, and the other half is poled in the direction parallel to the surfaces. The operation requires the application of an alternating voltage, like the conventional transformer, at a frequency that can excite a length-mode resonance. In the first half of the plate, the input electrical energy through the input terminal is converted into mechanical energy, which results in oscillatory vibration. This mechanical energy is then converted back to electrical energy and picked up at the output

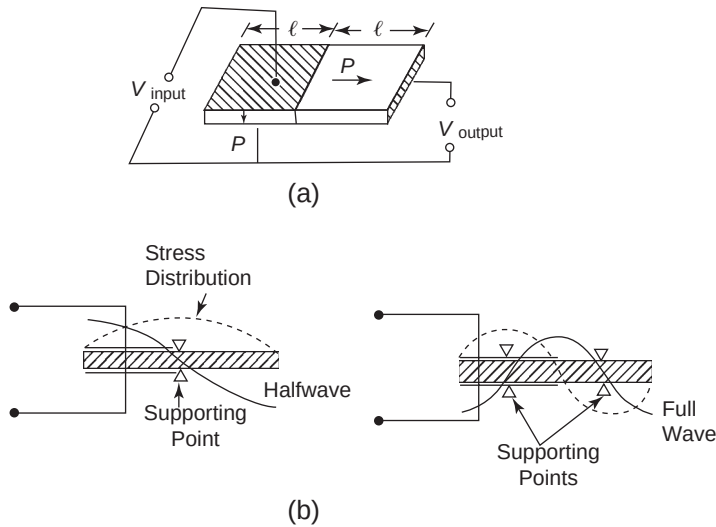


Figure 4-47 (a) Schematic illustration of a piezoelectric transformer and (b) the half- and full-wave resonances.

terminal. It is important that the length of the plate be equal to a number of half-wavelengths of the acoustic wave to allow standing waves to form. In short, the mechanical oscillations in the first part are transmitted at resonance to the second part, which reconverts the mechanical energy into electrical energy, resulting in the generation of a higher voltage. The resonant frequency is determined by the acoustic velocity v in the piezoelectric material. If the total length of the transformer is 2ℓ , then for the transformer length equal to one full wavelength, the full wave resonant frequency is

$$f = \frac{v}{2\ell} \quad (4-102)$$

For the transformer length equal to a half-wavelength, the half-wave resonant frequency is

$$f = \frac{v}{4\ell} \quad (4-103)$$

Figure 4-47(b) shows that half-wave and the full-wave resonance. Since the operation of the transformer is based mainly on the resonance conditions, the location of the supporting points in the installation of such transformers is very important. In general, supporting points must be located at places where the oscillatory vibration is zero, as shown in Figure 4-47(b).

Based on the equivalent circuit shown in Figure 4-40, when the resonance occurs, the reactance due to L_m and C_m falls to zero and the overall impedance becomes very small if R_m is small. So the piezoelectric material can be used for wave filters to allow only a selected frequency band close to the resonance frequency to pass and to stop waves of other frequencies. The efficiency of such wave filters depends on the value of the electro-mechanical coupling factor k .

4.4 Pyroelectric Phenomena

Of 20 noncentrosymmetric crystal classes, 10 contain a unique polar axis under the unstrained condition. This implies that such crystals are already spontaneously polarized in a certain temperature range. Thus, they exhibit both piezoelectric and pyroelectric effects. For the latter, the crystal will develop free electric charges at the surfaces when uniformly heated because the polarization inside the crystal is changed by the heat. It should be noted that if a crystal specimen is heated nonuniformly (for example, with a temperature gradient along the specimen), then the temperature gradient will also develop a mechanical stress. So in this

case, the specimen may have mixed piezoelectric and pyroelectric effects. However, under a uniform heating condition (i.e., a hydrostatic condition), the resultant piezoelectric effect vanishes.

Since pyroelectricity results mainly from the temperature dependence of the spontaneous polarization of the polar materials, it occurs in ferroelectric materials whether they are single-domain crystals or poled ceramics. A change in temperature will cause a corresponding change in polarization and hence a change in the compensating charges on the metallic electrodes deposited on its surfaces. This change will produce a current in an external circuit. If the pyroelectric specimen is not connected to an external circuit and kept perfectly insulated from its surroundings, the surface charges on the electrodes released due to a change in temperature will flow inside the material through the intrinsic electrical conductivity of the material σ , but it may take ϵ/σ time for the released charges to be neutralized.

4.4.1 Phenomenological Approach to Pyroelectric Effects

The interactions between electrical and thermal variables are the direct concern of this section. A change in a crystal's temperature can produce a number of electrical effects, which are referred to as *pyroelectricity*. These effects depend upon the thermal, mechanical, and electrical constraints on the crystal. The relation between D and T defines the pyroelectric coefficient (see Figure 4-38, the dashed line joining circle D and circle T) or

$$\begin{aligned} dD &= p^x dT \\ p^x &= \frac{dD}{dT} \end{aligned} \quad (4-104)$$

The unit of p is coulomb $\text{m}^{-2}\text{K}^{-1}$.

Pyroelectricity occurs only when the material exhibits spontaneous polarization. If the temperature of the specimen remains constant for a sufficient length of time, the charges accumulated on the surfaces will compensate for the internal polarization, so there will be no charge flow between the surfaces when they are short-

circuited. However, a change in temperature will cause a change in the resultant dipole moment of all dipoles and hence a change in the spontaneous polarization. The surface charges will redistribute themselves to compensate for the new internal polarization, thus producing a flow of charges when the specimen is short-circuited.

Since

$$\begin{aligned} D &= \epsilon_o F + P_{\text{total}} \\ &= \epsilon_o F + (P_{\text{induced}} + P_s) \end{aligned} \quad (4-105)$$

where P_{induced} and P_s are, respectively, the field-induced polarization and the spontaneous polarization. P_{induced} is much smaller than P_s and less temperature dependent. So p^x can be expressed as

$$p^x = \left. \frac{\partial P_s}{\partial T} \right|_{X, F} \quad (4-106)$$

The pyroelectric coefficient under the condition of constant mechanical stress implies that the crystal is not clamped, that is, it is free to expand or contract thermally. Based on the relations between F , D , X , and T shown in Figure 4-38, we can write

$$dh = \underbrace{y^{F,T} dX}_{\text{Elasticity}} + \underbrace{d^T dF}_{\text{Converse Piezoelectricity}} + \underbrace{\alpha^F dT}_{\text{Thermal expansion}} \quad (4-107)$$

$$dD = \underbrace{d^T (dX)}_{\text{Direct Piezoelectricity}} + \underbrace{\epsilon^{X,T} dF}_{\text{Permittivity}} + \underbrace{p^x dT}_{\text{Pyroelectricity}} \quad (4-108)$$

where y , d , ϵ , α , and p are, respectively, elastic compliance, piezoelectric coefficient, permittivity, thermal expansion coefficient, and pyroelectric coefficient. These equations should be written in tensor form, like those for piezoelectric effects, with subscripts to denote the direction and rotation of the vectors. For simplicity, we will ignore the complicated mathematical expressions and emphasize the physical concepts of the equations.

Under the constraint condition of constant electrical field (i.e., $dF = 0$), from Equations 4-107 and 4-108, we obtain

$$\begin{aligned} dX &= \frac{dh}{y^{F,T}} - \frac{\alpha^F}{y^{F,T}} dT \\ &= \frac{dD}{dT} - \frac{p^x}{dT} dT \end{aligned} \quad (4-109)$$

Under the condition of constant strain, $dh = 0$. This implies that the crystal is rigidly clamped, so expansion or contraction is not possible. Thus, from Equation 4-109, we obtain

$$\begin{aligned} p^x &= \left. \frac{\partial D}{\partial T} \right|_h + \frac{d^T \alpha^F}{y^{F,T}} \\ &= p^h + \frac{d^T d^F}{y^{F,T}} \end{aligned} \quad (4-110)$$

Equation 4-110 indicates that the total pyroelectric coefficient at constant stress p^x consists of two components:

- The pyroelectric coefficient at constant strain p^h due to the primary pyroelectric effect
- The secondary pyroelectric effect, which may occur when the crystal is free to react

As shown in Figure 4-1, all pyroelectric materials are piezoelectric, but not all piezoelectric materials are pyroelectric, because the latter requires spontaneous polarization for its occurrence. For example, α -quartz possesses three twofold axes, so a stress applied along any of these axes would cause a piezoelectric effect. But with heating, the material will expand equally along each of these axes. The angle between two adjacent axes is 120 degrees, so the pyroelectric effects tend to cancel each other. Hence, α -quartz is not pyroelectric. Similarly, zinc blend ZnS (zinc sulfide) possesses four threefold axes and hence is piezoelectric and not pyroelectric. On the other hand, the hexagonal zinc sulfide or cadmium sulfide with a wurtzite-type structure possesses one sixfold axis, so it is piezoelectric and also pyroelectric. In general, pyroelectric materials also exhibit the piezoelectric effect under hydrostatic pressure, and pyroelectric materials have secondary pyroelectric and piezoelectric effects.

It should be noted that, theoretically, all ferroelectric materials are pyroelectric, but for applications only those with the second-order transition at the Curie point are usable. From the $P_s - T$ characteristics shown in Figure 4-9 it can be seen that the pyroelectric coefficient is high just below the Curie point for the case with a second-order transition because dP_s/dT is large. For example, TGS exhibits a second-order transition and has a high pyroelectric coefficient just below the Curie point. The materials exhibiting a steep first-order transition at the Curie point are not suitable for pyroelectric applications because they exhibit a thermal hysteresis and it is difficult in most applications to maintain a sufficiently constant temperature environment (see also Ferroelectric Phenomena in Chapter 4).

The electrocaloric effect is the converse of the pyroelectric effect. The change of entropy can be expressed as

$$dS = \alpha^F dX + p^x dF + \frac{1}{Ty^{x,F}} dT \quad (4-111)$$

Under the condition of constant entropy, which is an adiabatic process, $dS = 0$ and constant stress, $dX = 0$, then the electrocaloric coefficient q^x can be written as

$$q^x = \left. \frac{dT}{dF} \right|_{s,x} = -Ty^{x,F} p^x \quad (4-112)$$

The unit of q^x is KmV^{-1} .

4.4.2 Pyroelectric Parameters and Their Measurements

The major parameter is the pyroelectric coefficient p^x since the electrocaloric coefficient is usually very small and scarcely used for the evaluation of pyroelectric effects. Several methods can be used for measuring p^x . Here, we describe only two simple ones.

One is the static method. In this method, the pyroelectric specimen is heated in a constant temperature chamber in which the temperature can be varied. Figure 4-48 shows the arrangement for measuring the charges developed on the pyroelectric specimen. These charges are then compensated by static charges induced on

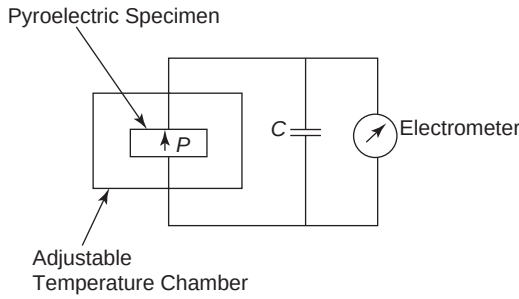


Figure 4-48 Schematic diagram illustrating the static method for the measurement of pyroelectric coefficient p . For the dynamic method, capacitor C is replaced by a resistor R .

capacitor C , which can be measured simply by an electrometer. Based on Equation 4-104, p^x can be evaluated by

$$p^x = \frac{Q_2 - Q_1}{A(T_2 - T_1)} \quad (4-113)$$

where Q_2 and Q_1 are the charges measured at temperatures T_2 and T_1 , respectively, and A is the surface area of the specimen.

The other method is the dynamic method. In this method, we can use the same arrangement shown in Figure 4-48 but with capacitor C replaced by a resistor R . In this case, the specimen is heated at a uniform rate of temperature rise, and it generates a current flowing through R measured also by an electrometer. We can write

$$i = \frac{dQ}{dt} = A \frac{dD}{dt} = A \frac{dD}{dT} \frac{dT}{dt} = A p^x \frac{dT}{dt}$$

Thus,

$$p^x = \frac{i}{A(dT/dt)} \quad (4-114)$$

where i is the current and dT/dt is the rate of temperature rise.

4.4.3 Pyroelectric and Thermally Sensitive Materials

Similar to piezoelectric materials, the most important pyroelectric materials are ceramics, synthetic polymers, and ceramic-polymer composites. In general, there are two types of pyro-

electric materials: one is ferroelectric, and the other is not ferroelectric but possesses spontaneous polarization, which may not be reversed when the electric field is reversed. For poled specimens, the remanent polarization is temperature dependent and falls to zero as the Curie temperature is approached. This $P_s - T$ relation below T_c is the cause of the pyroelectric effect. However, ferroelectric materials with the second-order transition at T_c generally have higher pyroelectric coefficients than nonferroelectric ones, particularly at temperatures just below T_c .

Table 4-5 lists the important parameters of some commonly used pyroelectric materials. The selection of pyroelectric materials for a special application should not be based solely on the figure of merit (see Equation 4-138); other properties, such as chemical stability, mechanical strength, power handling ability, and fabrication processes, should also be considered. For example, triglycine sulfate (TGS) has a high figure of merit and is therefore a commonly used material for low-power detection applications, but, it has its inherent shortcomings. It cannot be subjected to high heat—even sunlight may make it hot enough to depole. It is also mechanically weak and hygroscopic. However, the incorporation of alanine dopants into TGS greatly reduces its dielectric loss and dielectric constant, improving its performance as a radiation detector.^{140,141} Anyhow, this material is not suitable for devices to be used in an environment involving high vacuum or high humidity. Other materials with better mechanical strength and chemical stability than TGS, such as LiTaO_3 and SBN , have already been used widely for pyroelectric devices. They are also insensitive to humidity.

Lead zirconate (PZ)-based ceramics, such as PZT and PLZT, have a high value of pyroelectric coefficient and also of dielectric constant. It has been found that PZ doped with Fe, Nb, Ti, and U, such as $\text{Pb}_{1.02}(\text{Zr}_{0.58}\text{Fe}_{0.20}\text{Nb}_{0.20}\text{Ti}_{0.02})_{0.994}\text{U}_{0.006}\text{O}_3$ (PZFNTU), gives a higher pyroelectric coefficient and a lower dielectric constant.¹⁴²

As mentioned earlier, pyroelectric polymers are now widely used for many applications

Table 4-5 Selected properties of some pyroelectric materials: p = pyroelectric coefficient at constant stress and electric field, and c_p = heat capacity.

Material (Abbreviation)	Chemical Formula	p (C cm ⁻² K ⁻¹)	c_p (J cm ⁻³ K ⁻¹)	Temperature Range (K)
Triglycine sulfate (TGS)	(NH ₂ CH ₂ COOH) ₃ •H ₂ SO ₄	3.0×10^{-8}	1.70	273–321
Deuterated Triglycine Sulfate (DTGS)	(ND ₂ CD ₂ COOD ₃)•D ₂ SO ₄	3.0×10^{-8}	2.40	243–334
Lithium Tantalate (LT)	LiTaO ₃	1.9×10^{-8}	3.19	273–891
Strontium Barium Niobate (SBN)	Sr _{0.6} Ba _{0.4} Nb ₂ O ₆	8.5×10^{-8}	2.34	248–303
Lead Zirconate Titanate (PZT)	Pb(Zr _{0.52} Ti _{0.48})O ₃	5.5×10^{-8}	2.60	298–523
Lanthanum Modified PZT (PLZT)	Composition Ratio: La/Zr/Ti = 6/80/20	7.6×10^{-8}	2.57	—
Barium Titanate	BaTiO ₃	3.3×10^{-8}	—	293–320
Polyvinyl Chloride (PVC)	see Table 5-2	0.4×10^{-9}	2.40	—
Polyvinyl Fluoride (PVF)	see Table 5-2	1.8×10^{-9}	2.40	—
Polyvinylidene Fluoride (PVDF or Kynar)	see Table 5-2	4.0×10^{-9}	2.40	—

because of their ease of preparation in any shape and size. For example, polyvinyl fluoride (PVF) and polyvinylidene fluoride (PVDF) can be made extremely thin ($< 2 \mu\text{m}$). Although they have a low figure of merit, a low pyroelectric coefficient, and also a relatively high dielectric loss, they have a low heat conductivity and a low permittivity, facilitating both thermal and electrical coupling between neighboring elements. Their mechanical strength, thermal stability, and chemical stability are also good.

Apart from pyroelectric materials, there are some materials whose resistivity is very sensitive to temperature based on other mechanisms and not pyroelectric effects. These materials are usually called *negative temperature coefficient* (NTC) materials, whose resistivity increases with decreasing temperature, or *positive temperature coefficient* (PTC) materials, whose resistivity increases with increasing temperature. Obviously, there are numerous applications of such materials with a high temperature coefficient of resistivity (TCR) in temperature indicators (e.g., thermometers), temperature

controllers (e.g., thermostats, thermisters), current-limiting devices, etc.

NTC Materials

The most commonly used NTC materials are based on solid solutions of metallic oxides with a spinal structure, such as Fe₃O₄ – ZnCr₂O₄ and Fe₃O₄ – MgCr₂O₄, and also those based on Mn₃O₄ with partial replacement of Mn with Ni, Co, or Cu.⁹⁶ On the basis of electrical conduction in semiconductors due to transport of charge carriers in the bands or hopping between localized states in the band gap, the temperature dependence of the resistivity follows the relation

$$\rho_{NTC}(T) = A \exp\left(\frac{B}{kT}\right) \quad (4-115)$$

where k is the Boltzmann constant and A and B are constant. A is related to the resistivity at $T \rightarrow \infty$, and B is the activation energy for the

carrier movement. The negative TCR is

$$\alpha_{NTC} = \frac{1}{\rho_{NTC}} \frac{d\rho_{NTC}}{dT} = -\frac{B}{kT^2} \quad (4-116)$$

PTC Materials

The behavior of the DC electrical resistivity of ferroelectric ceramics is anomalous in the neighborhood of the Curie temperature T_C . A typical example is BaTiO_3 ceramic doped with 0.05×10^{-2} mol of Sn^{3+} or Nb^{3+} . When the trivalent donor Sn^{3+} or Nb^{3+} substitutes for Ba^{2+} , the extra positive charge in the material is compensated by an electron in the conduction band. The material behaves like a semiconductor, but electron mobility is also changed at the Curie temperature.¹⁴³ As a result, the resistivity of the

material increases with increasing temperature and reaches a peak at $T > T_C$, with the magnitude of several orders higher than that at T_C , as shown in Figure 4-49(a). This phenomenon is not observed in material with a single crystal structure. It occurs only in polycrystalline ceramics. Heywang¹⁴⁴ has proposed that this phenomenon is caused by the Schottky barrier at the grain boundaries in polycrystalline structures.

The energy band diagram near the grain boundary is shown in Figure 4-49(b). In the paraelectric phase, the band bending near the boundary is determined by the number of interface states N_{int} at the boundary. Similar to the Schottky barrier at the metal–semiconductor junction with the interface states at the metal–semiconductor interface, there is a

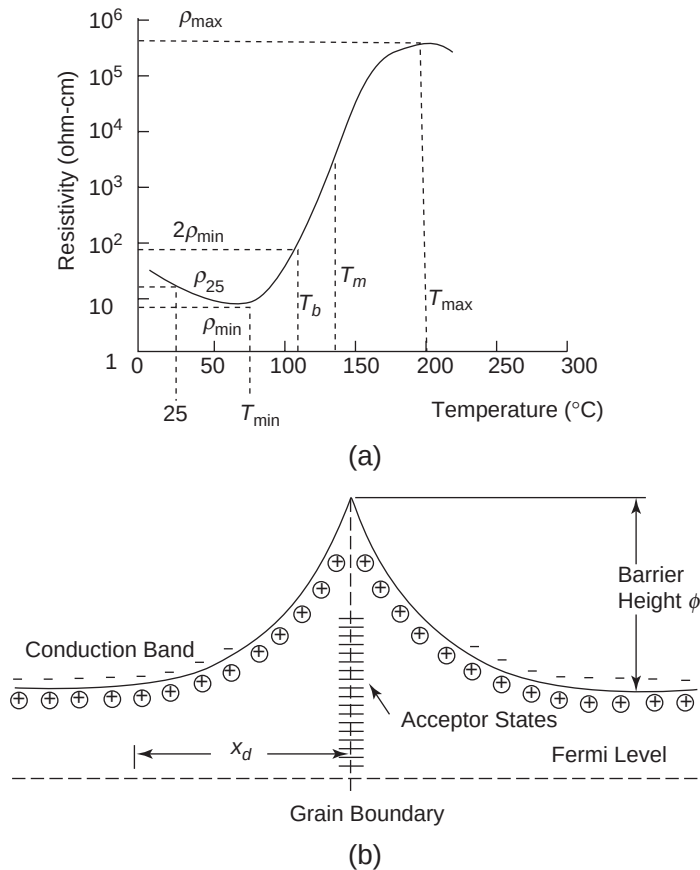


Figure 4-49 (a) The resistivity ρ as a function of temperature for BaTiO_3 based PTC material and (b) the energy band diagram near a grain boundary of a polycrystalline material.

depletion region x_d and a barrier height ϕ . Assuming that all donors are ionized, it can be shown that the barrier height is given by

$$\phi = \frac{q^2 N_D x_d^2}{2\epsilon_r \epsilon_o} \quad (4-117)$$

where N_D is the concentration of donors. The number of interface states at the boundary that accept electrons from the donors, acting as acceptor-like traps, must be $N_{int} = N_D x_d$. The barrier height depends on the size of grains because it affects the energy levels of the interface states.¹⁴⁵ The resistivity of the ceramic material is determined by the magnitude of ϕ . Thus, we can write

$$\rho_{PTC} \propto \exp\left(-\frac{\phi}{kT}\right) \quad (4-118)$$

For temperatures $T > T_c$, the dielectric constant decreases with temperature following the Curie–Weiss relation (see Equation 4-1). So, from Equations 4-1, 4-117, and 4-118, we obtain

$$\rho_{PTC} = A' \exp\left[-\frac{B'}{k}\left(1 - \frac{T_c}{T}\right)\right] \quad (4-119)$$

where A' and B' are constants related to doping concentration and structure of the material. It can be seen that for $T > T_c$ the resistivity increases with increasing temperature. The TCR for this case is thus

$$\alpha_{PTC} = \frac{1}{\rho_{PTC}} \frac{d\rho_{PTC}}{dT} = \frac{B'}{k} \frac{T_c}{T^2} \quad (4-120)$$

4.4.4 Applications of Pyroelectrics

Pyroelectric materials respond to changes in temperature either by continuous heating or by the absorption of sinusoidally modulated radiation. Pyroelectric materials may be used for thermal detectors and calorimeters, but they do not respond to a temporally steady temperature or to radiation intensity. To obtain a response from a stationary temperature or radiation, it is necessary to make the heat or radiation periodically interrupted by a rotating chopper placed between the heat or radiation source and the pyroelectric element. Obviously, there are many applications of pyroelectric phenomena.

In the following sections, we discuss some typical cases as examples to illustrate the basic principles of pyroelectric devices.

Pyroelectric Radiation Detectors

Radiation detectors are mainly used for the detection of infrared radiation. A poled pyroelectric element is typically a thin plate of rectangular shape with metallic electrodes deposited on both surfaces. It is important that the element be very thin, so that it heats up quickly and uniformly, and that at least one of the electrodes is a good absorber of radiation, so that it can convert radiation energy to heat. Figure 4-50(a) shows schematically the basic arrangement for the detection of infrared radiation. Under a short-circuit condition, there is a current flow due to the pyroelectricity, which is given by

$$i_p = pA \frac{dT}{dt} \quad (4-121)$$

where p and A are, respectively, the pyroelectric coefficient and the area of the element. In the following discussion, the superscript x of p^x is not shown for simplicity. The pyroelectric element behaves like a current generator with an internal resistance R_i and an internal capacitance C_i . The equivalent circuit to Figure 4-50(a) is shown in Figure 4-50(b). The detector also includes an amplifier represented by a shunt resistor R_d , a shunt capacitor C_d , and a voltmeter V .

It is assumed that all incident infrared radiation energy is absorbed by the element in time dt and rapidly distributed throughout the whole volume of the element, so thermal diffusion can be ignored. The total power W incident to the element may be written as

$$W = W_o + W_i \exp(j\omega t) \quad (4-122)$$

where W_o is the input power from the surrounding environment corresponding to temperature T_o , $W_i \exp(j\omega t)$ is the incident infrared power, and ω is its frequency. When the element temperature is increased from T_o to T , part of the absorbed power will be lost to the surroundings by re-radiation, conduction, or

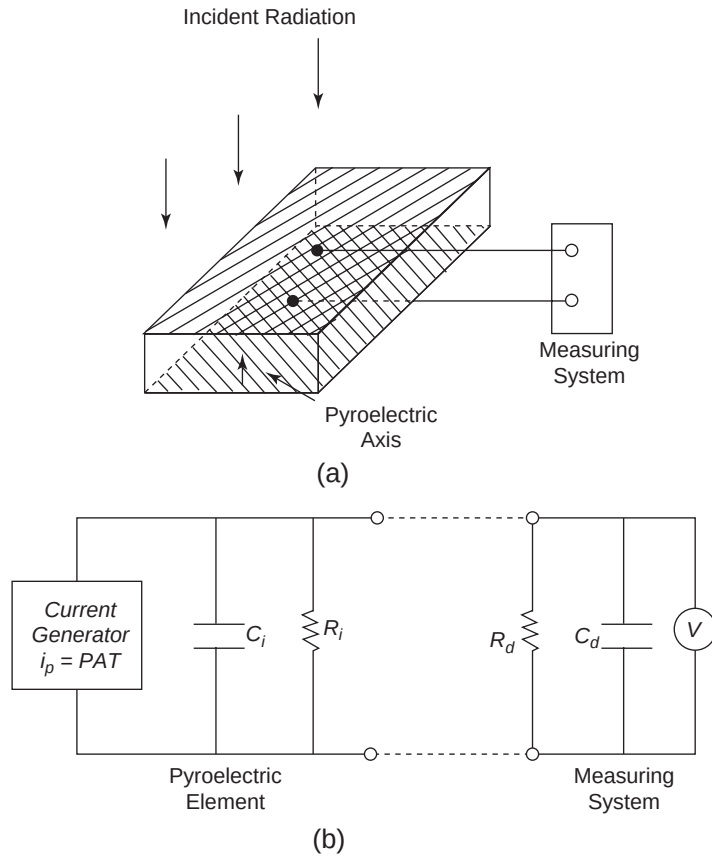


Figure 4-50 (a) Experimental arrangement for infrared radiation detection and (b) the equivalent circuit of the radiation detector shown in (a).

convection at a rate of G per unit temperature rise in the element due to incident radiation. Thus, we can write the energy balance equation as

$$\eta W_i \exp(j\omega t) - G(T - T_o) = H \frac{dT}{dt} \quad (4-123)$$

where η is the fraction of the incident power that heats the element; H is the heat capacity of the element, which is equal to ρcAL in which ρ is the density; c is the specific heat; and A and L are, respectively, the area and the thickness of the element. If the radiation is switched on at $t = 0$ and switched off at $t = t_1$, then when $t = t_1$, $W_i = 0$, Equation 4-123 becomes

$$H \frac{dT}{dt} + G(T - T_o) = 0 \quad (4-124)$$

Using the boundary condition (when $t = 0$, $T = T_o$ and when $t = t_1$, $T = T_1$), the solution of Equation 4-124 yields

$$T - T_o = (T_1 - T_o) \exp\left(-\frac{t}{\tau_T}\right) \quad (4-125)$$

where τ_T is the thermal relaxation time, which is

$$\tau_T = \frac{H}{G} \quad (4-126)$$

Assuming that the temperature of the element varies sinusoidally following the incident radiation frequency ω , then the solution of Equation 4-123 gives the continuous response of the pyroelectric element^{98,146} as follows:

$$T - T_o = \eta W_i (G^2 + \omega^2 H^2)^{-1/2} \exp[j(\omega t + \theta)] \quad (4-127)$$

where θ is the phase angle between the radiation and the temperature oscillation, which is given by

$$\theta = \tan^{-1}(\omega H/G) \quad (4-128)$$

Based on the equivalent circuit shown in Figure 4-50(b), the total resistance R_T and the capacitance C_T are

$$R_T = \frac{R_i R_d}{R_i + R_d} \quad (4-129)$$

$$C_T = C_i + C_d \quad (4-130)$$

So the element-generated current i_p due to the variation of temperature can be written as^{146,147}

$$\begin{aligned} i_p &= pA \frac{dT}{dt} = C_T \frac{dV}{dt} + \frac{V}{R_T} \\ &= \eta W_i pA \left(\frac{\omega}{G} \right) [1 + \omega^2 \tau_T^2]^{-1/2} \exp[j(\omega t + \theta)] \end{aligned} \quad (4-131)$$

The useful criteria for pyroelectric detectors are the current responsivity r_i and the voltage responsivity r_v , which are defined as follows:

$$r_i = \left| \frac{i_p}{W_i} \right| = \frac{pA\eta\omega}{G} (1 + \omega^2 \tau_T^2)^{-1/2} \quad (4-132)$$

From Equations 4-129 and 4-130, we can easily find the total admittance Y_T of the circuit. The voltage due to the change in temperature across R_d is given by

$$\begin{aligned} V &= \left| \frac{i_p}{Y_T} \right| \\ &= \eta W_i pA R_T \left(\frac{\omega}{G} \right) (1 + \omega^2 \tau_T^2)^{-1/2} (1 + \omega^2 \tau_E^2)^{-1/2} \end{aligned} \quad (4-133)$$

where τ_E is the electrical relaxation time, which is

$$\tau_E = R_T C_T \quad (4-134)$$

Thus, the voltage responsivity can be expressed as

$$r_v = \left| \frac{V}{W_i} \right| = \frac{\eta p A R_T \omega}{G(1 + \omega^2 \tau_T^2)^{1/2} (1 + \omega^2 \tau_E^2)^{1/2}} \quad (4-135)$$

Assuming that the circuit parameters are independent of frequency, we can see the following features:

From Equation 4-127, $T - T_o$ is independent of frequency for frequencies $\omega \ll \tau_T^{-1}$ and decreases with increasing frequency for $\omega \gg \tau_T^{-1}$.

From Equations 4-132 and 4-135, r_v is maximal at the frequency $\omega = (\tau_T \tau_E)^{-1/2}$. Both r_i and r_v increase with increasing frequency for $\omega < \tau_T^{-1}$ and decrease with increasing frequency for $\omega > \tau_T^{-1}$.

The maximum value of r_v occurs at $\omega = (\tau_T \tau_E)^{-1/2}$ and is given by

$$r_{v(\max)} = \frac{\eta p A R_T}{G(\tau_E + \tau_T)} \quad (4-136)$$

For high frequencies, $\omega^2 \tau_T^2 \gg 1$, $\omega^2 \tau_E^2 \gg 1$ and $C_i > C_d$, so Equation 4-135 can be simplified to

$$\begin{aligned} r_v &= \frac{\eta p A}{\omega H C_i} \\ &= \frac{\eta p}{A \omega c_p \epsilon} \end{aligned} \quad (4-137)$$

where $c_p = \rho c$. Since the sensitivity of the radiation detector increases with increasing r_v , the quantity $\frac{p}{c_p \epsilon}$ in Equation 4-137 may be used as a figure of merit for the pyroelectric radiation detector F_v , which is

$$F_v = \frac{p}{c_p \epsilon} \quad (4-138)$$

This is mainly a function of the material properties. In general, we would like the material to have a high pyroelectric coefficient and a low dielectric constant.

However, all signal detectors encounter the effect of noise, which is always present in any detector circuit. So, the sensitivity of any detector is determined by the level of noise in the amplifier output signal. Therefore, the signal-to-noise ratio must be made as large as possible. For pyroelectric detectors, the principal sources of noise are Johnson noise, thermal fluctuations, and amplifier noise. The noise level can also be described by the so-called *noise equivalent power* (NEP), which is defined as

$$NEP = \frac{\Delta V_N}{r_v(\Delta f)^{1/2}} \quad (4-139)$$

where ΔV_N is a signal voltage equal to the noise voltage, produced by a necessary amount of power input to the detector, and Δf is the bandwidth of the amplifier. It is often convenient to describe the minimum detective power of a detector by the so-called *detectivity* (D) which is defined as

$$D = \frac{1}{NEP} \quad (4-140)$$

Pyroelectric Burglar Alarm Systems

A blackbody at 300 K would radiate infrared rays of about $9.8\mu\text{m}$ in wavelength. Human or animal bodies whose temperature is around 310 K are expected to generate infrared radiation of about $10\mu\text{m}$ in wavelength. Thus, a moving intruder produces variable infrared radiation which, when reaching a pyroelectric detector, will switch on the alarm system. Figure 4-51 shows the arrangement of the pyroelectric elements and the alarm system. Since pyroelectric materials are also piezoelectric, they would produce electric charges due to external mechanical stresses caused by the expansion or contraction when the ambient temperature of the surroundings rises or falls. These stress-induced charges may interfere with the response of the active pyroelectric element to radiation. To prevent this interference, we need a dummy compensating element similar to the active element, but which does not respond to infrared radiation because the electrode facing the radiation is a reflecting electrode. The con-

nection of this compensating element and the active element is back to back, so the response of the compensating element to the ambient temperature fluctuation will cancel that of the active element. In general, such compensation is essential for many applications. The basic principle is the same as that described in the previous section. The output voltage of the system is to switch on the alarm system.

Pyroelectric Thermometry

In contrast to radiation detectors, pyroelectric thermometers receive energy through conduction or convection. As a result, noise due to thermal fluctuations is low. The principle is simple.^{148,149} Based on the equivalent circuit shown in Figure 4-50, Equation 4-131 can be rewritten as

$$pA \frac{dT}{dt} = C_T \frac{dV}{dt} + \frac{V}{R_T} \quad (4-141)$$

This equation indicates that pyroelectric thermometers can be used to measure the rate of temperature changes or the steps of steady temperature changes.

Measuring the Rate of Temperature Changes

In this case, the electrical relaxation time must be small (i.e., $\tau_E = R_T C_T \ll 1$), so the first term on the right side of Equation 4-141 can be neglected. Thus, we have

$$\frac{dT}{dt} = \frac{V}{pAR_T} \quad (4-142)$$

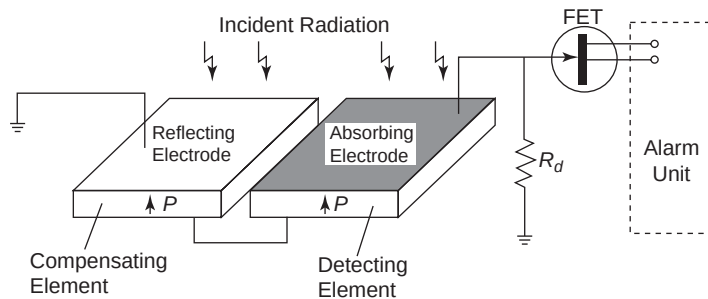


Figure 4-51 Schematic diagram of a pyroelectric burglar system alarm consisting of an active detecting element, a compensating (dummy) element, and an alarm unit.

The thermometer directly measures the rate of temperature changes by measuring the voltage across the pyroelectric element.

Measuring Steady Temperature Changes

In this case, the electrical relaxation time must be large (i.e., $\tau_E = R_T C_T \gg 1$), so the second term on the right side of Equation 4-141 can be neglected. Thus, we have

$$\frac{pA}{C_T} \frac{dT}{dt} = \frac{dV}{dt} \quad (4-143)$$

Integrating Equation 4-143 with the initial condition, $V(0) = V_o$ and $T(0) = T_o$, we obtain

$$T - T_o = (V - V_o) \frac{C_T}{pA} \quad (4-144)$$

The thermometer measures the steady temperature changes in steps by measuring V at T and V_o at T_o .

For more about the limitations of this method and the factors involved in temperature measurements, see references 146–149.

Pyroelectric Energy Conversion

Thermal energy input to a pyroelectric element will cause changes in the spontaneous polarization and hence generate electric charges on the electrodes. If a load is connected between the electrodes, a current will flow through the load, and electrical energy will be consumed there. The energy conversion efficiency is defined as

$$\begin{aligned} \psi &= \frac{\text{Electrical energy output}}{\text{Thermal energy input}} \\ &= \frac{i_p V}{W_i} \\ &= W_i r_i r_v \end{aligned} \quad (4-145)$$

From Equations 4-132 and 4-135, ψ can be written as

$$\psi = \frac{W_i p^2 R_T (\omega \tau_T)^2}{c_p^2 L^2 (1 + \omega^2 \tau_T^2) (1 + \omega^2 \tau_E^2)^{1/2}} \quad (4-146)$$

When $R_i \ll R_d$, $C_i \gg C_d$, $C_i = \epsilon A/L$ and $\sigma = L/AR_i = \omega \epsilon'$, we obtain the internal energy conversion efficiency of the pyroelectric element as

$$\psi_{in} = \frac{p^2}{c_p^2 (1 + \omega^2 \tau_T^2) (\epsilon^2 + \epsilon'^2) \omega L} \left(\frac{W_i}{A} \right) \quad (4-147)$$

ψ_{in} becomes a maximum for $\omega \tau_T = 1$. Using the average temperature change given by Equation 4-127, $\Delta T = T - T_o = W_i / c_p A$, we have

$$\psi_{in} = \frac{p^2 \Delta T}{2(\epsilon^2 + \epsilon'^2) c_p} \quad (4-148)$$

In practice, $\epsilon \gg \epsilon'$. So, we may consider $p^2 / c_p \epsilon$ as the figure of merit for energy conversion of pyroelectric materials. For typical values of $\tau_T = 1$ s and $W_i = 100$ mW cm⁻² and the temperature change $\Delta T = T - T_o = 50^\circ\text{C}$, the internal conversion efficiency is $\psi_{in} \approx 10^{-3}$. It has been found that $p^2 / c_p \epsilon$ does not vary much for most pyroelectric materials.^{98,150} So it is unlikely that the energy conversion efficiency would greatly exceed 10^{-3} under practical conditions, based on pyroelectric effects.

References

1. J. Valasek, Phys. Rev., 17, 475 (1921).
2. J. Valasek, Phys. Rev., 19, 478 (1922).
3. J. Valasek, Phys. Rev., 24., 560 (1924).
4. W.G. Cady, *Piezoelectricity*, (Dover, New York, 1964).
5. J. Curie and P. Curie, Compt. Rend., 91, 294 (1880) and 95, 1137 (1881).
6. Bunget and M. Popesco, *Physics of Solid Dielectrics*, (Elsevier, Amsterdam 1984), p. 383.
7. M. Born and K. Huang, *Dynamical Theory of Crystal Lattices*, (Clarendon, Oxford, 1954).
8. J.F. Nye, *Physical Properties of Crystals*, (Oxford University Press, Oxford, 1960).
9. C.S. Smith, "Macroscopic Symmetry and Properties of Crystals," in *Solid State Physics*, Edited by F. Seitz and D. Turnbull, Vol. 6 (Academic Press, New York, 1958), p. 175.
10. K.R. Brain, Proc. Phys. Soc., 36, 81 (1924).
11. E.L. Kern and S.M. Skinner, J. Appl. Polym. Sci., 6, 404 (1962).
12. S.N. Levine, J. Appl. Polym. Sci., 9, 3351 (1965).
13. G.E. Hauver, J. Appl. Phys., 36, 2113 (1965).
14. E. Fukada, M. Date, and N. Hirai, Nature, 211, 1079 (1966).

15. H. Kawai, Jpn. J. Appl. Phys., *8*, 975 (1969).
16. C.B. Sawyer and C.H. Tower, Phys., Rev., *35*, 269 (1930).
17. W.J. Merz, Phys. Rev., *76*, 1221 (1949).
18. W.J. Merz, Phys. Rev., *91*, 513 (1953).
19. E.J. Huibregtse and D.R. Young, Phys. Rev., *103*, 1705 (1956).
20. A.F. Davenshire, Phil. Mag., *42*, 1065 (1951).
21. H.H. Weider, Phys. Rev., *99*, 1161 (1955), and J. Appl. Phys., *26*, 1479 (1955).
22. W.J. Merz, Phys. Rev., *95*, 690 (1954).
23. R. Landauer, D.R. Young, and M.E. Drougard, J. Appl. Phys., *27*, 752 (1956).
24. H.H. Weider, J. Appl. Phys., *28*, 367 (1957).
25. D.S. Campbell, J. Electronics and Control, *3*, 330 (1957).
26. W. J. Merz, J. Appl. Phys., *27*, 938 (1956).
27. V. Janovec, Czechoslov. J. Phys., *8*, 3 (1956).
28. W.J. Merz and J.R. Anderson, Bell Lab. Record, *33*, 335 (1955).
29. J.R. Anderson, G.W. Brady, W.J. Merz, and J.P. Rameika, J. Appl. Phys., *26*, 1387 (1955).
30. E. Fatuzzo and W.J. Merz, *Ferroelectricity*, (North-Holland, Amsterdam, 1967).
31. A.G. Chynoweth, Phys. Rev., *110*, 1316 (1958).
32. R.C. Miller, Phys. Rev., *111*, 736 (1958).
33. V.M. Rudyak, A.Y. Kudzin, and T.V. Panchenko, Sov. Phys.—Solid State, *14*, 2112 (1973).
34. G.F. Bacon and R.S. Pease, Proc. Royal Soc. London, *A220*, 397 (1953) and *A230*, 359 (1955).
35. M. Tokunaga, Ferroelectrics, *1*, 195 (1970).
36. G. Busch, J. Helv. Phys. Acta, *11*, 269 (1938).
37. A. Von Arx and W. Bantle, J. Helv. Phys. Acta, *17*, 298 (1944).
38. W. Kanzig, "Ferroelectrics and Antiferroelectrics," in *Solid State Physics* edited by F. Seitz and D. Turnbull, Vol. 4 (Academic Press, New York 1957) pp. 1–197.
39. F. Jone and G. Shirane, *Ferroelectric Crystals*, (MacMillan, New York, 1962).
40. H.M. Barkla and D.M. Finlayson, Phil. Mag., *44*, 109 (1953).
41. V.L. Hablutzel, J. Helv. Phys. Acta, *12*, 489 (1939).
42. A.N. Holden and W.P. Mason, Phys. Rev., *57*, 54 (1940).
43. S. Hoshino, T. Mitsui, F. Jona, and R. Repinsky, Phys. Rev., *107*, 1255 (1957).
44. B. Jaffe, W.R. Cook, and H. Jaffe, *Piezoelectric Ceramics*, (Academic Press, New York, 1971).
45. H.M. Barnett, J. Appl. Phys., *33*, 1606 (1962).
46. C. Michel, J.M. Moreau, G.D. Achenbach, R. Gerson, and W.J. James, Solid State Commun., *7*, 865 (1969).
47. E. Sawaguchi, J. Phys. Soc. Japan, *8*, 615 (1953).
48. S. Roberts, J. Am. Ceram. Soc., *33*, 63 (1956), and Phys. Rev., *83*, 1078 (1951).
49. G. Shirane, E. Sawaguchi, and Y. Takagi, Phys. Rev., *84*, 476 (1951).
50. E. Sawaguchi and K. Kittaka, J. Phys. Soc. Japan, *7*, 336 (1952).
51. R. Redin, G. Marks, and C. Antoniak, J. Appl. Phys., *34*, 600 (1963).
52. G. Jonker, J. Am. Ceram. Soc., *55*, 57 (1972).
53. H. Jaffe and D. Berlincourt, Proc. IEEE., *53*, 1372 (1965).
54. K. Hardt, Ferroelectrics, *12*, 9 (1976).
55. L. Benguigui, Solid State Commun., *11*, 825 (1972).
56. K. Okazaki, Ferroelectrics, *35*, 173 (1981).
57. P. Thatcher, Appl. Opt., *16*, 3210 (1977).
58. R.C. Buchanan (Ed.), *Ceramic Materials for Electronics*, (Marcel Dekker, New York, 1986).
59. L.M. Levinson, *Electronic Ceramics*, (Marcel Dekker, New York, 1988).
60. G.H. Haceting and C.E. Land, J. Am. Ceram. Soc., *54*, 1 (1971).
61. R.H. Dungan, H.M. Barnett, and A.H. Stark, J. Am. Ceram. Soc., *45*, 226 (1962).
62. Y.L. Wang, W.Z. Yuan, G.H. He, S.W. Lin, Y.H. Ling, C.F. Qu, and B.G. Wang, Ferroelectrics, *49*, 169 (1983).
63. Y.L. Wang and K.C. Kao, in *Conference Record of the 2000 IEEE International Symposium on Electrical Insulation* (IEEE-DEIS, New York, 2000) pp. 58–61.
64. W.M. Prest, Jr. and D.J. Luca, J. Appl. Phys., *46*, 4136 (1975).
65. R. Hasegawa, Y. Takahashi, Y. Chatani, and H. Tadokoro, Polym. J., *3*, 600 (1972).
66. M.A. Bachmann and J.B. Lando, Macromolecules, *14*, 40 (1981).
67. A.J. Lovinger, Science, *220*, 1115 (1983).
68. J. Scheinbeim, C. Nakafukus, B.A. Newman, and K.D. Pae, J. Appl. Phys., *50*, 4399 (1979).
69. A.J. Lovinger, in *Developments in Crystalline Physics*, Vol. 1, edited by D.C. Bassett (Applied Sciences, London, 1982) Chapter 5.

70. T. Takahashi, M. Date, and E. Fukada, *Appl. Phys. Lett.*, **37**, 791 (1980).
71. A.J. Lovinger, G.T. Davis, T. Furukawa, and M.G. Broadhurst, *Macromolecules*, **15**, 323 and 329 (1982).
72. J.G. Bergman, Jr., J.H. McFee, and G.R. Crane, *Appl. Phys. Lett.*, **18**, 203 (1971).
73. A.F. Devonshire, Quarterly Supplement of *Phil. Mag.*, *Advances in Physics*, **3**, 85 (1954).
74. E.J. Huibregtse and D.R. Young, *Phys. Rev.*, **103**, 1705 (1956).
75. D. Meyerhofer, *Technical Report 121*, Laboratory for Insulation Research, Massachusetts Institute of Technology, 1957; also *Phys. Rev.*, **112**, 413 (1958).
76. C. Kittel, *Phys. Rev.*, **82**, 729 (1951).
77. L.E. Cross and B.J. Nicholson, *Phil. Mag.*, **46**, 453 (1955).
78. P.W. Forsbergh, "Piezoelectricity, Electrostriction and Ferroelectricity" in *Handbuch der Physik*, Vol. 17, (Springer-Verlag, Berlin, 1956) pp. 264–392.
79. G.J. Goldsmith and J.G. White, *J. Chem. Phys.*, **31**, 1175 (1959).
80. H.L. Stadler and P.J. Zachmanidis, *J. Appl. Phys.*, **34**, 3255 (1963).
81. E.A. Little, *Phys. Rev.*, **98**, 978 (1955).
82. J.C. Burfoot, *Ferroelectrics: An Introduction to the Physical Principles*, (Van Nostrand, Princeton, 1967).
83. J.A. Hooton and W.J. Merz, *Phys. Rev.*, **98**, 409 (1955).
84. R.C. Miller and A. Savage, *Phys. Rev.*, **112**, 755 (1958); *Phys. Rev. Lett.*, **2**, 294 (1959); also *J. Appl. Phys.* **30**, 808 (1959).
85. R.C. Miller and G. Weinreich, *Phys. Rev.*, **117**, 1460 (1960).
86. M. Hayashi, *J. Phys. Soc. Japan*, **33**, 617 and 739 (1972).
87. M. McQuarrie, *Bull. Am. Ceram. Soc.* **34**, 170 (1955).
88. G. Shirane and K. Sukuzi, *J. Phys. Soc. Japan* **7**, 333 (1952).
89. A. Glanc, V. Dvorak, V. Janovec, E. Rechzigel and V. Janouser, *Phys. Lett.* **7**, 106 (1966).
90. A.E. Feuersanger, in *Thin Film Dielectrics*, edited by F. Vratny (Electrochemical Society, New York, 1969), p. 209.
91. Y.Y. Tomashpolski, M.A. Sevostianov, M.V. Pentagova, L.A. Sorokina, and Y.N. Venevstsev, *Ferroelectrics* **7**, 257 (1974).
92. J.R. Slack and J.C. Burfoot, *J. Phys. C. Solid State*, **4**, 898 (1971).
93. J.L. Vossen and W. Kern (Eds.) *Thin Film Processes*, (Academic Press, New York, 1978).
94. K. Lichtenecker, *Phys. Z.* **10**, 1005 (1909) and **27**, 115 (1926).
95. K.W. Plessner and R. West "High Permittivity Ceramics for Capacitors," in *Progress in Dielectrics*, Vol. 2, edited by J.B. Birks and J.H. Schulman (Heywood, London, 1960) pp. 165–192.
96. A.J. Moulson and J.M. Herbert *Electroceramics: Materials, Properties and Applications*, (Chapman and Hall, London, 1997).
97. J.M. Herbert, *Ceramic Dielectrics and Capacitors*, (Gordon and Breach, London, 1985).
98. M.E. Lines and A.M. Glass, *Principles and Applications of Ferroelectrics and Related Materials*, (Clarendon, Oxford, 1977).
99. S. Nomura and S. Sawada, *J. Phys. Soc. Japan* **6**, 36 (1951); also *ibid.*, **10**, 108 (1955).
100. A. Glanc, Z. Malek, J. Mastner, M. Novak, and J. Strajblova, *J. Appl. Phys.* **35**, 1870 and 1875 (1964).
101. J. Fousek, *J. Appl. Phys.* **36**, 588 (1968).
102. R. Koontz, G. Blookhina, S. Gold and A. Krasnykh, *Annual Report of 1998 Conference On Electrical Insulation and Dielectric Phenomena* (IEEE Publication 98CH36257-1, New York) pp. 23–26.
103. F.W. Neilson, *Bull. Am. Phys. Soc.*, **2**, 302 (1957).
104. P.C. Lysne and C.M. Percival, *J. Appl. Phys.*, **46**, 1519 (1975).
105. W. Mock and W.H. Holt, *J. Appl. Phys.*, **49**, 5846 (1978).
106. J.R. Anderson, *Electrical Engineering*, **71**, 916 (1952).
107. J.F. Scott and C.A. Araujo, *Science*, **246**, 1400 (1989).
108. J.F. Scott, C.A. Araujo, H.B. Meadows, L.D. McMillan, and A. Schawabkeh, *J. Appl. Phys.*, **66**, 1444 (1989).
109. J.F. Scott, C.A. Araujo, and L.D. McMillan, *Ferroelectrics*, **116**, 107 (1991).
110. P.K. Larson, R. Cuppens, and C.A.C.N. Spierings, *Ferroelectrics*, **128**, 265 (1992).
111. L.K. Anderson, *Ferroelectrics*, **7**, 55 (1974); also **3**, 69 (1972).
112. S.Y. Wu, W.J. Takel, and M.H. Francombe, *Appl. Phys. Lett.* **22**, 26 (1973).
113. S.A. Keneman, A. Miller, and G.W. Taylor, *Ferroelectrics*, **3**, 131 (1972).
114. P.P. Peercy and C. Land, *IEEE Trans. Electron Devices*, **ED-28**, 756 (1981).
115. W.P. Mason, *Piezoelectric Crystals and their Applications in Ultrasonics*, (Van Nostrand, New York, 1950).

116. K.S. Van Dyke, *Proc. IRE.*, **16**, 742 (1928).
117. H.D. Megan, *Ferroelectricity in Crystals*, (Methuen, London, 1957).
118. D. Berlincourt, D. Curran, and H. Jaffee, in *Physical Acoustics*, edited by W.P. Mason (Academic Press, New York, 1964).
119. A.R. Von Hippel, *Molecular Science and Molecular Engineering*, (John Wiley, New York, 1959), p. 300.
120. J.M. Herbert, *Ferroelectric Transducers and Sensors*, (Gordon Breach, London, 1982).
121. J. Van Randerlaet and R. Setterington (Eds.), *Piezoelectric Ceramics*, (Mullard, London, 1974).
122. G. Shirane, K. Suzuki, and A. Takada, *J. Phys. Soc. Japan* **7**, 12 (1952).
123. G. Shirane and K. Suzuki, *J. Phys. Soc. Japan* **7**, 333 (1952).
124. B. Jaffe, R. Roth, and J. Marzullo, *J. Appl. Phys.*, **25**, 809 (1954).
125. G. Haertling and C. Land, *J. Am. Ceram. Soc.*, **54**, 1 (1971).
126. S. Liu, S. Pai, and J. Kyonka, *Ferroelectrics*, **22**, 689 (1978).
127. K. Hardt, *Ferroelectrics*, **12**, 9 (1976).
128. K. Okazaki, *Ferroelectrics*, **35**, 173 (1981).
129. M. Multani, S. Gokarn, R. Vijayaraghavan, and V. Polkar, *Ferroelectrics*, **37**, 652 (1981).
130. N. Murayama, T. Oikawa, T. Katto, and K. Nakamura, *J. Polym. Sci. Polym. Phys. Ed.* **13**, 1033 (1975).
131. N. Murayama and Hazhizume, *J. Polym. Sci. Polym. Phys. Ed.* **14**, 989 (1976).
132. G. Pfister, M. Abkowitz, and R.G. Crystal, *J. Appl. Phys.* **44**, 2064 (1973).
133. T. Furukawa, K. Ishida, and E. Fukada, *J. Appl. Phys.* **50**, 4904 (1979).
134. D. Liufu and K.C. Kao, *J. Vac. Sci. Technol.*, **A16**, 2360 (1998).
135. K. Uchino, *Piezoelectric Actuators and Ultrasonic Motors*, (Kluwer Academic, Boston, 1996).
136. K. Uchino, *Ferroelectrics*, **91**, 281 (1989).
137. K. Uchino, "Ceramic Actuators: Principles and Applications," (*MRS Bulletin*, April 1993), p. 42.
138. J.G. Smith and W.S. Choi, *IEEE Trans. Ultrason., Ferroelect., Freq. Contr.* **38**, 256 (1991).
139. Q.M. Wang, X.H. Du, B.M. Xu, and L.E. Cross, *IEEE Trans. Ultrason., Ferroelect., Freq. Contr.* **46**, 638 (1999).
140. P.J. Lock, *Appl. Phys. Lett.* **19**, 390 (1971).
141. E.T. Keve, K.L. Bye, R.W. Whipples, and A.D. Annis, *Ferroelectrics*, **3**, 39 (1971).
142. R.W. Whatmore, *Rep. Prog. Phys.* **49**, 1335 (1988).
143. C.N. Berglund and W.S. Baer, *Phys. Rev.* **157**, 358 (1967).
144. H. Heywang, *J. Mater. Sci.* **6**, 1214 (1971).
145. N. Hirose and H. Sasaki, *J. Am. Ceram. Soc.* **54**, 320 (1971).
146. S.B. Lang, *Sourcebook of Pyroelectricity*, (Gordon and Breach, London, 1974).
147. S.B. Lang and F. Steckel, *Rev. Sci. Instru.* **36**, 1817 (1965).
148. S.B. Lang, S.A. Shaw, L.H. Rice, and K.D. Timmerhaus, *Rev. Sci. Instru.* **40**, 274 (1969).
149. S.B. Lang, "Pyroelectricity: Its Phenomenology and Its Application to Temperature Measurements" in *Heat Transfer*, (Elsevier, Amsterdam, 1970), Section MT 1.6.
150. J.D. Zook and S.T. Liu, *Ferroelectrics*, **11**, 371 (1976).

5 Electrets

The science of electricity is that state in which every part of it requires experimental investigation; not merely for the discovery of new effects, but the development of the means by which the old effects are produced, and the consequent more accurate determination of the first principles of action of the most extraordinary and universal power in nature.

Without experiment, I am nothing.

Michael Faraday

5.1 Introductory Remarks

An electret can be considered a piece of dielectric material with the presence of quasi-permanent real charges on the surface or in the bulk of the material, or frozen-in aligned dipoles in the bulk. An electret behaves like a battery or acts as an electrical counterpart of a permanent magnet. A piece of poled ferroelectric material can also be an electret. The term *quasi-permanent* implies that the amount of charges stored in the material does not remain the same permanently, but decays very slowly depending on the situation, and the decay time is normally much longer than the time period over which the electret is in use.

In 1892, British scientist Oliver Heaviside was the first to introduce the term *electret* and to discuss its properties¹. In 1919, Japanese scientist Mototaro Eguchi was the first to fabricate an electret by melting equal parts of Carnauba wax (extracted from the Brazilian Carnauba palm tree) and resin with a little beeswax and then lowering the temperature to allow the liquid mixture to become solidified while being subjected to an electric field of about 10 kV cm^{-1} .^{2,3} The internal polarization of the electret discs prepared by this method persisted for many years. In fact, electrets prepared by Eguchi's method today are called the *thermo-electrets*, because they are formed mainly by a thermal process. Many materials can now be used to fabricate thermoelectrets, including organic materials such as ebonite, naphthalene, polymethyl-methacrylate, and many polymers, and inorganic materials such as sulfur, quartz, glasses, steatite, and some ceramics.

Later, other types of electrets were discovered using special methods of creating real charges or polarization. In 1937, Nadjakov discovered photoelectrets.^{4,5} He found that charge separation occurred when a sulfur layer was subjected simultaneously to a visible light illumination and an electric field stressing. This photoelectric phenomenon later led to the development of xerography.⁶ This phenomenon was further studied by Kallmann et al.^{7,8} and Fridkin et al.⁹

There are many other ways to form electrets. For example, electro-electrets are produced by applying a strong electric field across the dielectric material between two metallic electrodes, causing polarization, injection of charge carriers, or both. Electrets formed by carrier injection can also be achieved by corona discharges or electron beams on the surface of the dielectric materials. Some kinds of electrets can be produced without the use of an electric field, such as magnetoelectrets produced by heating the dielectric material (e.g., polymethyl-methacrylate) in a strong static magnetic field, and then cooling it slowly to the normal ambient temperature. This process makes the dielectric material become polarized due to the magnetic anisotropy of molecules, which tend to orient along with their dipole moments.¹⁰ Some electrets can be formed simply by a thermal process without involving either electric or magnetic fields. Such electrets are due mainly to charge separation through a phase transition on phase changes during solidification.¹¹ Some electrets can be formed by applying a mechanical pressure to the dielectric material. In this case, surface charges can be

acquired simply by contact electrification, deformation, or friction without the application of an electric field.¹²

In this chapter, we shall discuss the formation of electrets and their related properties with emphasis on the basic physical concepts needed to identify important material parameters and to deduce guidelines for optimizing some desirable properties and predicting possible new applications.

5.2 Formation of Electrets

The electrical charges that can be created and stored in a dielectric material to form an electret have two major types: monocharges (also called *real charges*) and dipolar charges, which are in fact aligned dipoles and can be induced in dipolar materials containing dipolar molecules or in ferroelectric materials. Real charges may appear on the material surface, called *surface charges*, or in the bulk to form so-called *space charges*.

Surface charges or space charges with the same polarity as that of the adjacent forming electrode are called *homocharges*, and those with the polarity opposite to that of the adjacent forming electrode are called *heterocharges*. Homocharges may appear as a result of charge carrier injection from the electrode into the dielectric material, followed by the capture of the injected carriers in traps near the injecting electrode. They may also appear due to the depositing of charge carriers resulting from the electric discharge in an air gap between the material surface and the forming electrode. Heterocharges are due mainly to the presence of aligned dipoles, to the displacement of the existing charge carriers (electrons or ions) in the material, or to the presence of residual impurities, which are apt to be ionized. Figure 5-1 depicts some common forms of electrets.

In this section, we shall describe some of the most important methods for forming electrets. These methods are based mainly on the manner of charging or polarizing the dielectric material and on the structure of the material, whether it is dipolar or nondipolar.

5.2.1. Thermo-Electrical Method

Basically, this method is similar to the one originally used by Eguchi.³ It involves the simultaneous application of an electric field and heat to a dielectric material for a predetermined period of time and the subsequent cooling to normal ambient temperature while the electric field is still applied. Figure 5-2 shows three general electrode arrangements for forming thermo-electrets.

One type has metallic electrodes vacuum-deposited on two surfaces of the dielectric material to form intimate contacts, as shown in Figure 5-2(a). Another type has only one surface with an intimate metallic contact; an air gap is between the other surface and a forming metallic electrode, as shown in Figure 5-2(b). A third type is without metallic contacts on both surfaces but with air gaps between the surfaces and the forming metallic electrodes, as shown in Figure 5-2(c).

In the case with a dielectric material between two intimate metallic electrodes, the dipolar molecules (dipoles) are randomly arranged, as shown in Figure 5-3(a), but they will actively orient under an electric field and at an elevated temperature. In general, the temperature is higher than the glass transition temperature T_g , at which the material becomes very amorphous, homogeneous, and elastic, so the dipoles can orient easily in the direction of the applied field, as shown in Figure 5-3(b). After the poling period, the material is gradually cooled down to the normal ambient temperature T_o while the electric field is still applied. Finally, the applied electric field is removed and the material forms a thermo-electret consisting mainly of aligned dipolar charges, as shown in Figure 5-3(c). The temperature-applied field time sequence for the poling process is shown in Figure 5-3(d), and the resulting polarization in Figure 5-3(e).

After the removal of the applied field, the total polarization gradually decreases to a quasi-steady level, which is mainly the remaining frozen-in dipolar polarization associated with the difference in relative permittivity between T_g and T_o . At T_g the induced polariza-

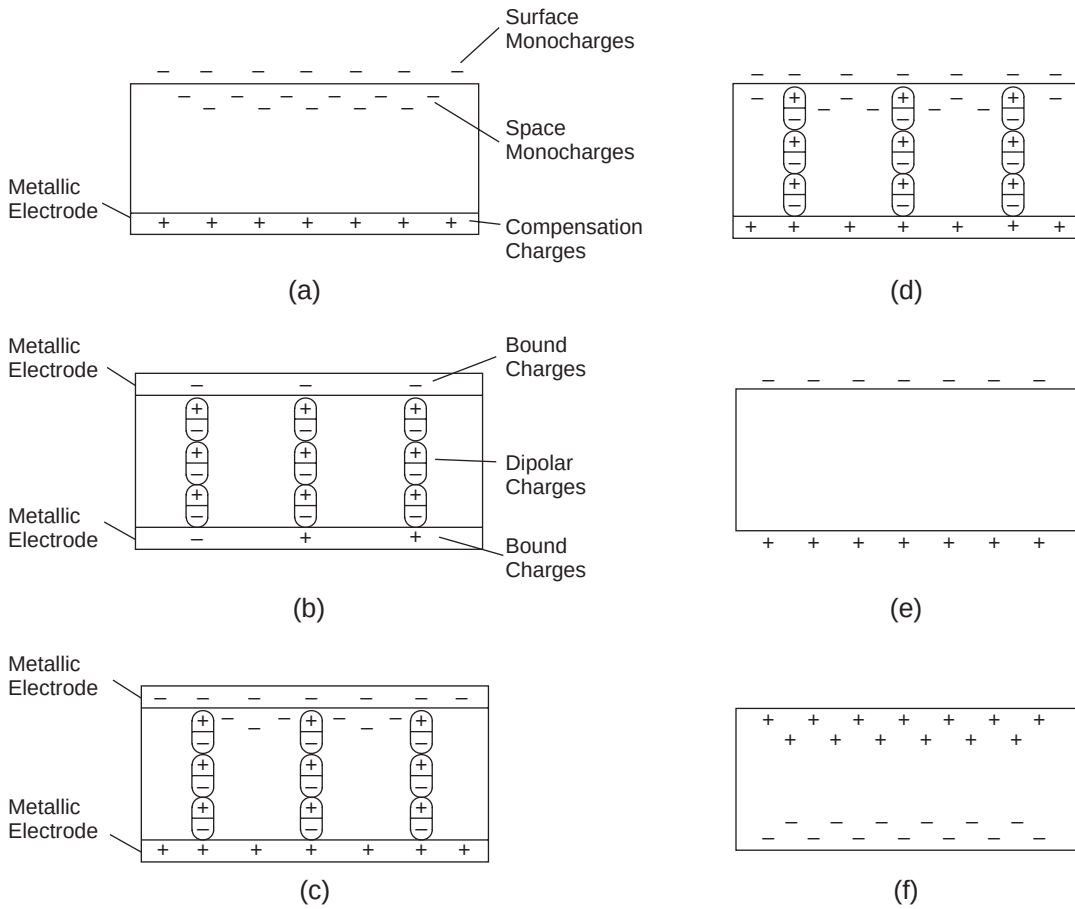


Figure 5-1 Schematic illustration of electrets involving (a) surface monocharges and space monocharges, (b) dipolar charges only, (c) dipolar charges and space monocharges, (d) surface monocharges, space monocharges, and dipolar charges, (e) deposited negative and positive surface charges, and (f) displacement of negative and positive real charges in the bulk.

tion is also a part of the total polarization under an electric field. As soon as the electric field is removed, the part due to induced polarization will gradually return to its randomized state. However, if the applied field is sufficiently high to cause carrier injection from the electrode to the dielectric material, but not high enough to cause internal discharge or breakdown, then the electret formed will consist of both frozen-in aligned dipolar charges and injected space monocharges captured in traps, as shown in Figure 5-1(c).

If the material is nondipolar, then we must produce real charge storage on the surface or in the bulk to form the electrets, as shown in

Figure 5-1(c) and (d). In doing so, we use the electrode arrangements with one or two air gaps, like those shown in Figure 5-2(b) and (c). When there is only one air gap, as shown in Figure 5-2(b), the system is basically a classical Maxwell–Wagner two-layer system. Assuming that the dielectric material is nonconductive with a relative permittivity ϵ_r , then the electric field in the air gap F_o and that in the dielectric material F_e are given by

$$F_o = \frac{\epsilon_r V}{d + \epsilon_r s} \quad (5-1)$$

$$F_e = \frac{F_o}{\epsilon_r} \quad (5-2)$$

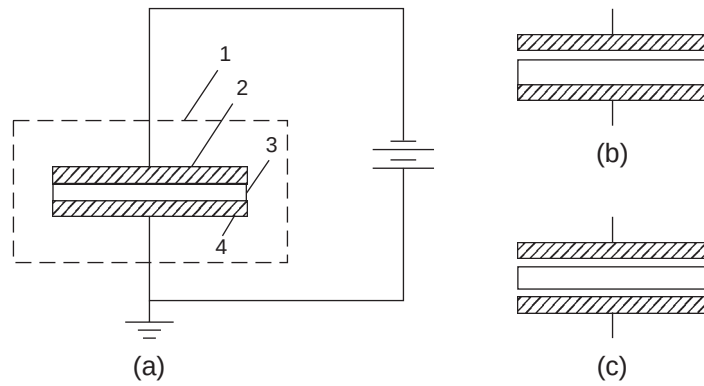


Figure 5-2 Three general electrode arrangements for forming thermo-electrets: (a) metallic electrodes deposited on both specimen surfaces, (b) a metallic electrode deposited on only one surface and the other surface bare, with an air gap between it and the forming metallic electrode, and (c) both surfaces bare with air gaps between bare surfaces and the forming electrodes. 1: Screening cage or heating chamber; 2: Upper metallic electrode; 3: Dielectric specimen; 4: Lower metallic electrode.

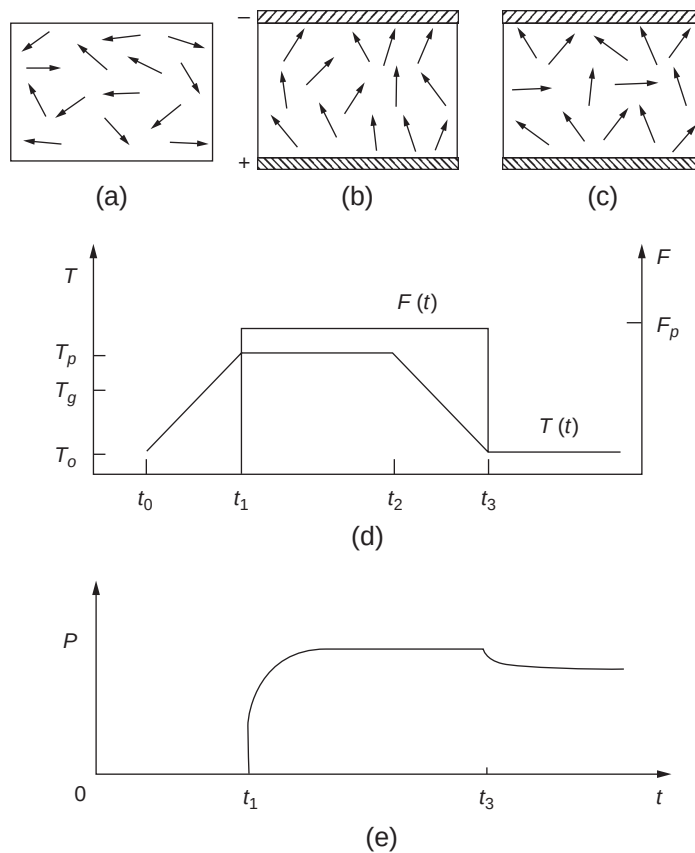


Figure 5-3 Thermo-electrical electret formation processes: (a) prior to the application of an electric field, (b) during the poling process at a temperature $T > T_g$ and a constant electric field, (c) alignment of dipoles in the electret after poling, (d) the temperature–time and electric field–time sequence in the poling process, and (e) the variation of polarization with time. T_p : poling temperature; T_g : glass transition temperature; T_o : normal ambient temperature; F_p : poling field.

where V is the applied voltage across the system and d and s are, respectively, the material specimen thickness and the air gap length.

For any applied voltage, the field in the air gap is E_γ times larger than that in the material. The field required to cause electrical discharge in the air gap is about 30 kV cm^{-1} , and generally the field required to cause carrier injection from the electrode to the dielectric material is much larger ($>100 \text{ kV cm}^{-1}$), depending on the potential barrier at the metal–dielectric interface. For this reason, it is easier and more convenient to use a system with one or two air gaps. When the applied voltage is large enough for F_o to reach the value to cause ionization in the air gap, real charges (electrons and ions) will be created and deposited on the material's bare surface; electrons also may penetrate into the bulk and be trapped if the applied field is sufficiently high. This process should also be carried out at an elevated temperature to ensure that charges will penetrate to be trapped in deep traps in the material near the surface, ensuring the thermal stability of the electret. It is important to remember that the applied field must not be removed until the material has been cooled down completely to the normal ambient temperature. For the electrets of $5\text{--}10 \mu\text{m}$ in thickness, the air gaps are usually of the order of $0.1\text{--}1.0 \text{ mm}$. This method is simple and the electrets formed by it are stable, but it is a slow process and the lateral charge distribution in the electret is not uniform.

When Eguchi made thermo-electrets in a Carnauba wax–beeswax mixture, the forming metallic electrodes were not in intimate contact with the material surfaces. There were tiny air gaps between the material surfaces and the adjacent forming electrodes. During the poling period with an electric field, the air gaps would break down, producing ions and electrons, which could be deposited as homocharges on the material surface. So, after the removal of the electric field and the forming electrodes, the charges on the electret surfaces were the sum of the heterocharges due to the frozen-in aligned dipolar charges and the homocharges due to the electrical discharge in the tiny air gaps during the poling period.^{13–15}

At the beginning, after the electret is formed, volume polarization is dominant. This is why

Eguchi observed the heterocharges appearing on the electret surface. But after a few days, he found that the polarity of the surface charges changed to the opposite of their original one, that is, homocharges appeared on the material surfaces. This was because the amount of dipolar charges decays with time, due to the continuous relaxation of the aligned dipoles to their randomized state. When such decay reaches a certain level, the deposited charges, which may have been trapped deep in the material just below the surface, may become dominant. This is why after a few days Eguchi observed homocharges appearing on the electret surfaces.

5.2.2 Liquid-Contact Method

With this method, the bottom electrode of the dielectric material is either a vacuum-deposited metallic film or a simple metal plate, but the top forming electrode is generally made of fabric wetted with a conductive liquid, such as ethyl alcohol or water.¹⁶ Under an applied electric field between these two electrodes, the charges will be transferred from the top wetted electrode to the dielectric material surface, due to the interaction of the electrostatic and molecular forces at the interface during the contacting period, as shown in Figure 5-4. Since the top wetted electrode can be moved on the material surface, this method can be used to transfer charges over a large area of the material surface. It is also easy to control the charge density deposited on the surface and to make the lateral charge distribution comparatively uniform. Note that a certain time is required for the transfer of charges, so it is suggested to

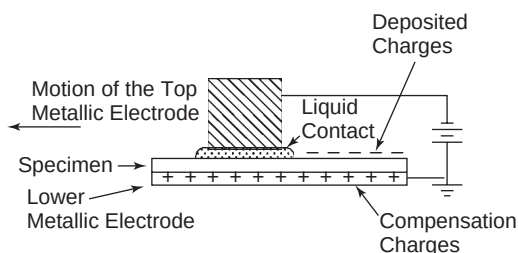


Figure 5-4 Schematic illustration of the liquid-contact charging method.

move the top electrode to one position and to leave it there for a short period (say, 10 to 20 minutes) to allow some time for the transferring process before moving it to the next position. The amount of charges or charge density on the surface depends on both the applied voltage and the poling time.

5.2.3 Corona Discharge Method

With this method, the bottom electrode is a vacuum-deposited metallic film on the material surface, and the top forming metallic electrode is usually made of a metallic wire (e.g., tungsten wire), with one end polished to a sharp point so that voltage of only a few kV between the two electrodes is sufficient to create a field near the point electrode, exceeding the breakdown strength of the air in a region of a few millimeters around the point electrode. This is shown in Figure 5-5. Under this condition, corona discharge will occur, and negative and positive ions will be formed in the region near the point electrode, depending on the surrounding medium and the polarity of the point electrode.

In air at one atmospheric pressure and humidity of 40–60%, negative ions such as CO_3^- and O_2^- are present when the point electrode is negative, and positive ions such as $(\text{H}_2\text{O})_n\text{H}^+$, NO^+ , and NO_3^+ are present when the point electrode is positive. For the negative point electrode, only negative ions can reach the dielectric material surface. These ions will then give up their attached electrons, becoming

neutralized, and return to the surrounding atmosphere, leaving a layer of negative trapped electrons just below the material surface. Similarly, when the point electrode is positive, positive ions, when reaching the material surface, will accept electrons from the material through the material–air interface, becoming neutralized, and return to the surrounding atmosphere, leaving a layer of positive trapped holes just below the material surface.

Since the electric field distribution around a single point electrode is very nonuniform, with the field decreasing gradually with radial distance from the point, the distribution of the deposited charges on the material surface is also very nonuniform. To improve this situation, a metal-wire grid can be placed between the point electrode and the material surface, as shown in Figure 5-5. The potential applied to the grid should be the same in polarity as, but lower in amplitude than, that of the point electrode. Some typical values are $V_1 = 8\text{ kV}$, $V_2 = 0.5\text{ kV}$, $d_1 = 4\text{ mm}$, and $d_2 = 3\text{ mm}$, the discharging time being about four minutes. With the grid, the distribution of the deposited surface charge is generally uniform if the discharging has reached such a level that the potential at the material surface approaches the potential of the grid.

The corona discharge forming process can be carried out at normal ambient temperature, but the deposited surface charges may decay easily because the charges on the surface layer are mainly resting in shallow traps, so thermal stability is not very good. To remedy this situation, it is desirable to carry out the forming process at an elevated temperature. This may cause more charges to be captured in deep traps. Furthermore, the heating chamber also serves as a screening cage to prevent any exterior interference. The high-temperature corona discharge forming process also plays an indirect role in the annealing process to enhance the thermal stability of the electrets formed. The advantage of using a point electrode over a planar electrode, as shown in Figure 5-2(b), is that the applied voltage required to cause electrical discharge near the point electrode is much lower than that for the planar electrode, thus

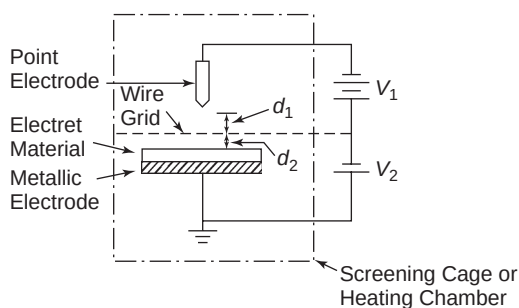


Figure 5-5 Schematic illustration of the corona discharge method.

reducing the possibility of causing damage to the material surface from the bombardment of charged particles. It is worth mentioning that, if the surface potential and the current flowing to the bottom electrode to form compensating charges there can be measured continuously, then the surface charge density can be estimated during the charging process.¹⁷

5.2.4 Electron-Beam Method

In order to produce real charges stored in the bulk of the material, we usually employ low-energy electron beams with energy of the order of 10–50 keV, so that the range of the penetrating electrons is smaller than the dielectric material thickness. When this method is used to inject electrons into the material, the electrons will soon be trapped inside, forming negative trapped space charges. The energy of the electron beam should be controlled according to the structure and thickness of the material specimens to be used for forming electrets.^{18,19} For example, the energy of an electron beam of 10–50 keV is suitable for Teflon of 1–20 μm in thickness. A typical electron beam system is shown schematically in Figure 5-6. Beam scanning is required to control the uniformity of the distribution of the injected electrons.

For a dielectric material specimen with a metallic electrode vacuum-deposited on the bottom surface and grounded, the energetic electrons from the beam, when striking the top bare surface, will generate electrons from the surface due to secondary emission, leaving a positively charged surface layer that is later neutralized by incoming slow electrons. The energetic primary electrons then penetrate into the bulk of the material and collide with molecules, creating electron-hole pairs, which may be quickly trapped but which will be occasionally released by thermal activation, contributing additional conductivity. This additional conductivity is generally referred to as *radiation-induced conductivity* (RIC).

The magnitude of RIC depends on the dose and the energy of the electron beam, but in most practical cases, RIC is much larger than the intrinsic conductivity of the material.²⁰ The original primary electrons, after one or more

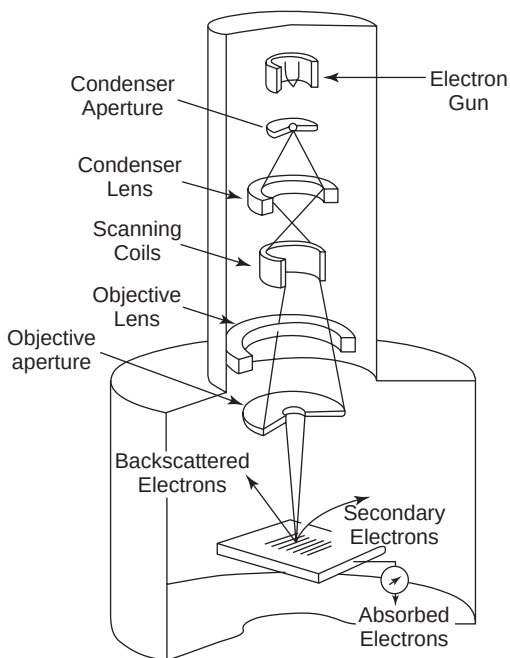


Figure 5-6 Schematic diagram of an electron-beam system.

collisions, lose most of their energy and are then trapped, forming negative charges located around the average range.^{20,21} The distribution of negative charges creates an internal field that tends to help the electrons to penetrate beyond the average range. In the RIC region, however, the electrons tend to move toward the interior of the material and the holes to move in the opposite direction. The motion of the electrons will gradually settle down to a thermal equilibrium state, resulting in a very stable distribution of negative charges and hence the formation of a negatively charged electret with high thermal stability. It has been found that for Teflon electrets formed by electron beam injection, the negative charge distribution is stable for a period of more than 10 years at normal ambient temperature (i.e., room temperature).²²

The electron-beam method has the advantage of easily controlling the density, location, and lateral distribution of the injected negative charges. This is why it is widely adopted for forming electrets for research purposes and for practical applications. It should be noted

that a similar ion beam technique, used for ion implantation in today's semiconductor technology, is not normally used for injecting ions into a dielectric material for forming an electret. This is simply because much higher energy is required to inject ions than electrons into the same range in the material, so this technique would cause a lot of damage to the material.

5.2.5 Electromagnetic Radiation Method

A dielectric material can be charged by the displacement of the charge carriers generated by various kinds of penetrating radiation, such as x-rays or ultraviolet or visible lights, under an externally applied electric field. The electrets formed by x-ray radiation are generally referred to as *radio-electrets*. When ultraviolet or visible light is used, the electrets formed are generally called *photo-electrets*.

The method is basically similar to that used by Vadjakov.^{4,5} The materials used for photo-electrets are generally photoconductive materials. A metallic electrode is usually deposited on one surface of the material specimen and a transparent electrode on the other surface, so that the light can illuminate the specimen through it. Sometimes it is desirable to have both electrodes transparent so that both the input light and the output light intensities can be measured, and hence, the photoelectron yield in the material can be determined. Under an electric field, the photogenerated carriers will be separated and will move toward the electrodes. These carriers may be trapped near the electrodes to create a space charge polarization. After the removal of the illumination and the electric field, this persistent polarization in the material forms a photo-electret, as shown in Figure 5-7. However, carrier trapping, detrapping, and retrapping occur continually during illumination. After the removal of the light, the charge distribution may become quasi-equilibrium, but the polarization still decays gradually, even in the dark. Because the electrets formed by the photo-electric method are generally less stable than those formed by

other methods described here, this method is not widely used.

The five methods described here are the major methods for forming electrets. Of course, they can be further improved by varying the performing ambients to suit the particular materials used for forming electrets. For example, for charging based on electrical discharge in air or other gas, we can change the pressure from the normal atmospheric pressure (760 torr) to a lower pressure (<100 torr) to produce electric discharge at a much lower voltage. The temperature during the forming process always plays an important role in the thermal stability of the electrets formed. An elevated temperature can provide a chance for the charges located in shallow traps to move to deep traps to stabilize the charge distribution, in other words, to improve the thermal stability of the electrets formed.

5.3 Charges, Electric Fields, and Currents in Electrets

In general, electrets for practical applications are of the configuration shown in Figure 5-8. Such an electret is usually made of a thin dielectric material in sheet form or thin-film form, with a metallic electrode deposited on the lower surface and the upper surface bare. Also, the upper floating planar metallic electrode is parallel to the bare surface with an air gap between them. The thickness of the electret is d and the air gap length is s . When the switch is on the *A* position, the specimen is being poled under a DC voltage V . Of course, we can use the thermo-electrical method described in Section 5.2.1. for the poling. If the poling voltage is sufficiently high to cause electrical discharge in the air gap, then after the poling period and the switch has been turned to the *B* position (open circuit), the electret formed will consist of quasi-permanent trapped real charges on the bare surface, to form the surface-trapped real charge density σ_r ; quasi-permanent trapped real charges in the bulk, to form the volume-trapped real charge density ρ_r ; and uniform

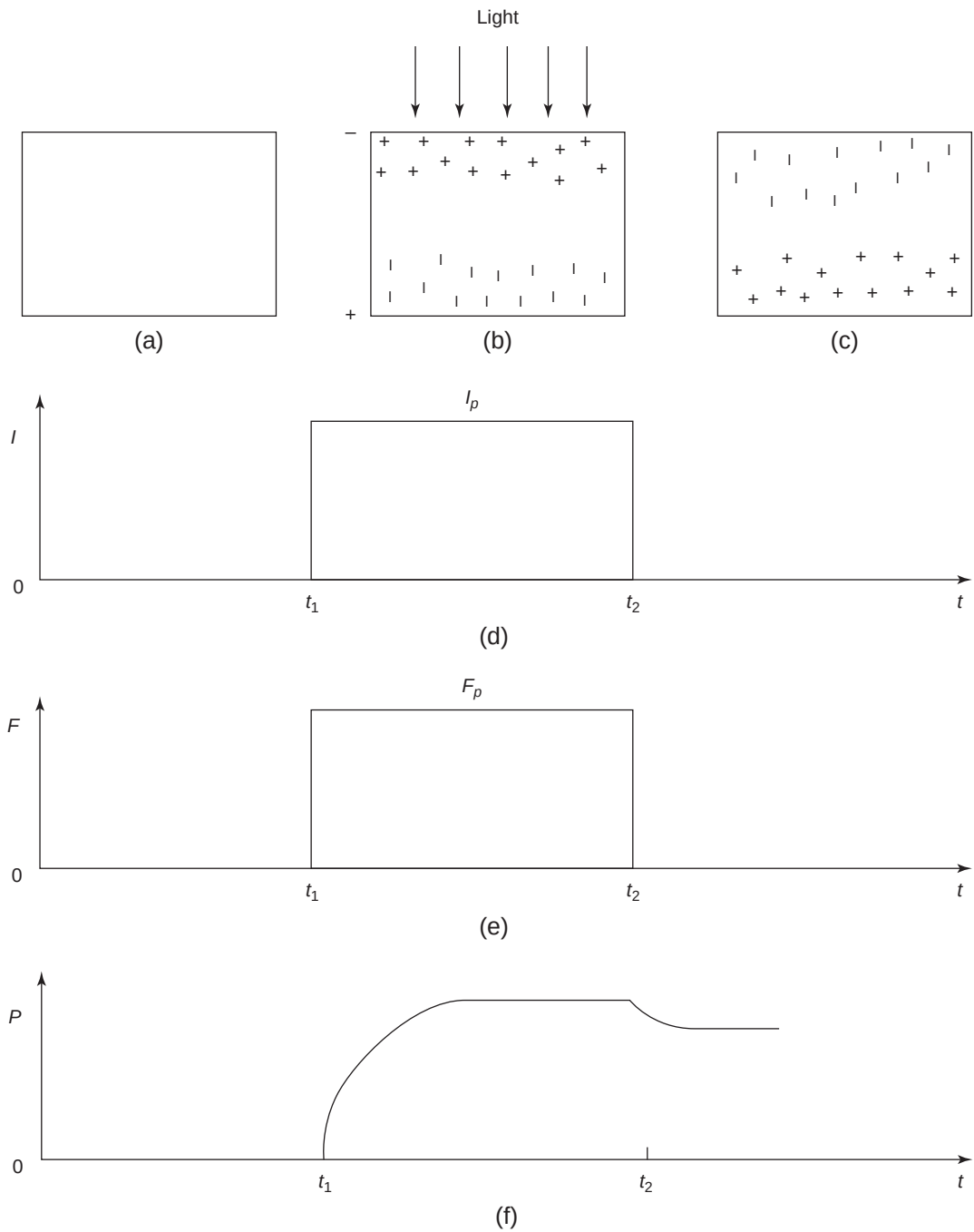


Figure 5-7 Photo-electret formation processes: (a) prior to the application of light and an electric field, (b) displacement of photo-induced charges under light intensity I_p and electric field F_p , (c) charge distribution after poling, (d) the light intensity I -time t sequence, (e) the electric field F -time t sequence, and (f) the variation of polarization with time. I_p : light intensity at poling; F_p : poling field.

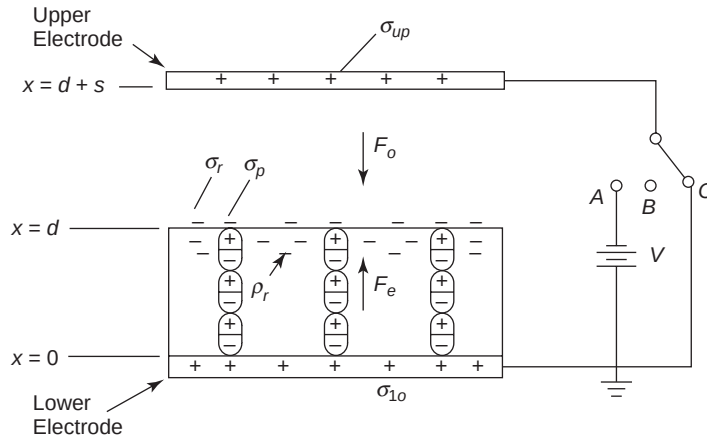


Figure 5-8 Arrangement of an electret and a floating upper electrode—the configuration most commonly used for applications. Switch on position A during poling period; switch on position B after being fully poled (open circuit); switch on position C under short-circuit condition.

polarization P , composed of the quasi-permanent frozen-in dipolar polarization P_p and the instantaneously induced polarization P_i , which follows the poling field F . Thus

$$P = P_p + P_i \quad (5-3)$$

where

$$P_i = (\epsilon_r - 1)\epsilon_0 F \quad (5-4)$$

in which ϵ_r is the dielectric constant of the electret material. P_p may consist of not only the frozen-in dipolar polarization, which is normally invariable with x , but also possibly a component due to charge displacement in the structural discontinuity. However, P would create a local volume charge with the local charge concentration given by

$$\rho_p = -dP/dx \quad (5-5)$$

The polarization P also creates a bound surface charge density on the bare surface, which is given by

$$\sigma_p = -P \quad (5-6)$$

Assuming that the lateral dimensions (i.e., the y and z directions perpendicular to the x direction) are much larger than d and s , and that the lateral distributions of σ_r , ρ_r and P are invariable with y and z , implying that they vary only with x ; then after the poling (charging)

process, the electret can be characterized by the following two charge equations:

$$\sigma(d) = \sigma_r(d) + \sigma_p(d) = \sigma_r - P \quad (5-7)$$

$$\rho(x) = \rho_r(x) + \rho_p(x) = \rho_r - \frac{dP}{dx} \quad (5-8)$$

In this section, we shall deal with two general cases.

Case I: The Electret Has Only Surface Charges

In this case we have

$$\begin{aligned} \rho(x) &= 0 \\ \sigma(d) &= \sigma_r - P \end{aligned} \quad (5-9)$$

When the switch is turned to the C position (short-circuit), the voltage between the two electrodes is zero. In this case, the compensation charges on both electrodes will rearrange in such a way that the two electrodes shield all charges stored in the electret. The compensation charges on the lower and the upper electrodes, σ_{lo} and σ_{up} , respectively, must satisfy the Gauss's law. Thus,

$$\sigma_{lo} + \sigma_{up} = -\sigma \quad (5-10)$$

Therefore, there is no current flowing through the short circuit because there is no potential between the two electrodes. So

$$F_{ed} = F_{os} \quad (5-11)$$

where F_e and F_o are, respectively, the electric fields inside the electret and in the air gap. Equation 5-10 can be rewritten in terms of F_e and F_o as

$$\epsilon_r \epsilon_o F_e + \epsilon_o F_o = -\sigma \quad (5-12)$$

From Equations 5-11 and 5-12, we obtain

$$F_e = -\frac{\sigma s}{(\epsilon_r s + d)\epsilon_o} \quad (5-13)$$

$$F_o = -\frac{\sigma d}{(\epsilon_r s + d)\epsilon_o} \quad (5-14)$$

and the voltage across the electret V_e and that across the air gap V_o as

$$V_e = -\frac{\sigma d}{\left(\epsilon_r + \frac{d}{s}\right)\epsilon_o} \quad (5-15)$$

$$V_o = -\frac{\sigma d}{\left(\epsilon_r + \frac{d}{s}\right)\epsilon_o} = V_e \quad (5-16)$$

From Equations 5-10, 5-13, and 5-14, we obtain the relations between σ_{io} or σ_{up} and σ , which are

$$\sigma_{io} = -\frac{\epsilon_r s \sigma}{\epsilon_r s + d} \quad (5-17)$$

$$\sigma_{up} = -\frac{d \sigma}{\epsilon_r s + d} \quad (5-18)$$

It can be seen that with this configuration, even d is constant. Varying the air gap length s will cause a change in both F_e and F_o and hence in compensation charges on the electrodes σ_{io} and σ_{up} . The change in σ_{io} and σ_{up} implies that there will be a current flow through the short circuit. If there is a load (e.g., resistor) connected between the two electrodes instead of a short circuit, then an electrical signal will appear across the load when the air gap s is changed by an external force. Many applications of electrets are based on this simple principle. We shall discuss this more in Section 5.9.

If we move the upper electrode to infinity (i.e., make $s \rightarrow \infty$), we have the fields in the electret and in the air gap as

$$\begin{aligned} F_e &= -\frac{\sigma}{\left(\epsilon_r + \frac{d}{s}\right)\epsilon_o} = -\frac{\sigma}{\epsilon_r \epsilon_o} \\ F_o &= -\frac{\sigma}{\left(\epsilon_r \frac{s}{d} + 1\right)\epsilon_o} = -\frac{\sigma}{\epsilon_r \epsilon_o} \frac{d}{s} \rightarrow 0 \end{aligned} \quad (5-19)$$

and the potential across the electret and air gap as

$$V_e = V_o = -\frac{\sigma d}{\epsilon_r \epsilon_o} \quad (5-20)$$

which is independent of s . This is equivalent to saying that $\sigma_{up} \rightarrow 0$, according to Equation 5-18. So, instead of moving the upper electrode to infinity, if we apply a voltage equal in magnitude but opposite in polarity to the potential V_o to compensate the potential across the air gap, then the voltage across the electret is still the same as that given in Equation 5-20. We can use this idea to determine the surface charge density. This is generally referred to as the *compensation method* and will be discussed in Section 5.4.

Case II: The Electret Has Both Surface Charges $\sigma(d)$ and Volume Charges $\rho(x)$

The convenient way to deal with this situation is to convert the volume charges per unit planar area to an equivalent surface charges per unit area σ_p . This means that the effect of $\rho(x)$ in the bulk is equivalent to that of σ_p on the bare surface. σ_p can be called the *projected* or the *equivalent surface charge density*, which is given by

$$\sigma_p = \frac{1}{d} \int_0^d x \rho(x) dx \quad (5-21)$$

Thus, the total surface charge density σ_T can be written as

$$\sigma_T = \sigma + \sigma_p = \sigma_r - P + \sigma_p \quad (5-22)$$

By replacing σ with σ_T , all equations from Equation 5-13 to Equation 5-20 can be used for the evaluation of F_e , F_o , V_e , V_o , σ_{io} , and σ_{up} .

However, to calculate σ_p , we must know the spatial distribution of $\rho(x)$, which is difficult to find theoretically. It can, however, be deter-

mined experimentally. This will be discussed in Section 5.6.

For practical applications, we always want V_e as large as possible to increase the sensitivity of the electret to respond to external signals. But the air gap s is a limiting factor, so s should be chosen so that V_o across the air gap is below the electrical breakdown voltage of the air gap. This directly limits the value of V_e . For example, with a Teflon electret of $20\text{ }\mu\text{m}$ in thickness (d) and the air gap of $40\text{ }\mu\text{m}$ in gap length (s), the breakdown voltage of the air gap at normal atmospheric pressure and temperature is 500 V for ps (gas pressure $\times$ air gap length) equal to $40\text{ }\mu\text{m} \times 760\text{ mmHg} = 3.04\text{ mm(Hg)} - \text{cm}$, based on Paschen's law.²³ This means that since $V_e = V_o$, the maximum value for V_e is 500 V. Normally, the allowable value for V_e in this case is about 200–300 V. Since $\sigma_T = \epsilon_r \epsilon_o V_e / d$, if we want a large V_e , we must make σ_T larger by producing more surface and volume charges in the electret. However, if σ_T is fixed, we can always adjust s to ensure that electrical breakdown will not occur in the air gap.

In electrets, there always exist mechanical forces induced by the electric fields on the material. As discussed in Electromechanical Effects in Chapter 2, the mechanical forces per unit area acting on both the upper and the lower electrodes and tending to pull them toward the electret surface are given by

$$\mathcal{F}_o = \frac{1}{2} \epsilon_o F_o^2 \quad (5-23)$$

pulling the upper electrode downward, and

$$\mathcal{F}_e = \frac{1}{2} \epsilon_r \epsilon_o F_e^2 \quad (5-24)$$

pulling the lower electrode upward.

All kinds of charges stored in electrets will give rise to a current flow, which may be due to a change of temperature, exposure to light, temporal variation of the electric field, decay of charges with time and so on. The parameters, P , σ_r , ρ , and F_e are not really time invariant. They are functions of time because they are always subjected to decay processes. In general, the current density in the electret can be written as

$$j_e(t) = \epsilon_r \epsilon_o \frac{\partial F_e}{\partial t} + \frac{\partial P}{\partial t} + j_g \quad (5-25)$$

where the first term on the right side of equation 5-25 is the displacement current due mainly to the variation of σ with time, the second term is due to the depolarization (dipolar relaxation), and the last term is due to intrinsic conductivity and the presence of thermally released trapped charge carriers (real charge decay processes).

5.4 Measurements of Total Surface Charge Density and Total Charges

The total charges in the electret per unit planar area are given by

$$Q_T = \sigma + \int_0^d \rho(x) dx \quad (5-26)$$

and the total surface charge density is given by

$$\sigma_T = \sigma + \frac{1}{d} \int_0^d x \rho(x) dx \quad (5-27)$$

There are several methods available to determine the total surface charge density σ_T , total charges per unit planar area Q_T , and the volume charges $\int_0^d \rho(x) dx$ per unit planar area. This section reviews some commonly used methods for these measurements.

5.4.1 Total Surface Charge Density

In this section we shall describe two methods for the measurement of the total surface charge density.

Compensation Method

In Section 5.3, we derived equations for the voltages across the electret and across the air gap for the electret with one surface metallized and one surface bare, and with a floating metal electrode above the bare surface and an air gap between them (see Figure 5-8 and Equations 5-15 and 5-16. When $s \rightarrow \infty$ we have

$$V_e = V_o = -\frac{\sigma_T d}{\epsilon_r \epsilon_o}$$

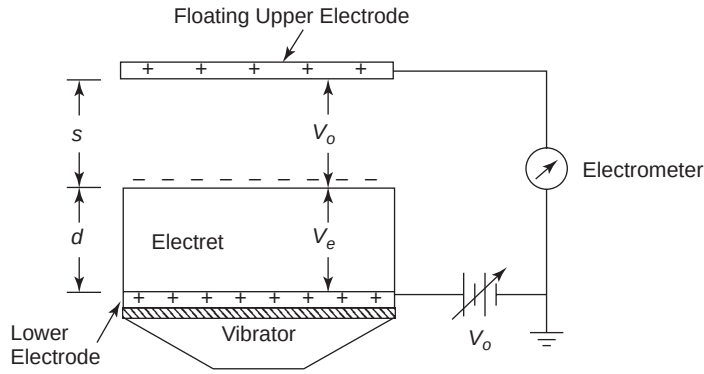


Figure 5-9 Schematic illustration of the compensation method for the measurement of total surface charge density on the charged electret.

and so

$$\sigma_T = -\frac{\epsilon_r \epsilon_0}{d} V_e \quad (5-28)$$

Thus, by moving the floating electrode a large distance from the bare surface of the electret so that $s \gg d$, σ_T can be measured simply by connecting a variable DC voltage supply between the lower electrode and ground, and an electrometer between the floating electrode and ground, as shown in Figure 5-9. Adjusting the applied DC voltage until it is equal to the electrometer reading, the electrometer reading is V_e , since the applied DC voltage has balanced out the voltage across the air gap V_o . This is why this method is called the compensation method. Knowing V_e , we can calculate σ_T from Equation 5-28. This method is simple, and it is independent of the air gap length s , so it does not involve the measurement of s .

The sensitivity of this method can be improved by attaching a mechanical vibrator to the lower electrode to cause the electret to vibrate, as shown in Figure 5-9. The vibrator can be a simple loudspeaker. The vibration will cause the variation of s , and hence the voltage reading in the electrometer has an AC component. However, if V_o can be completely balanced out by adjusting the compensation voltage precisely, the AC component should disappear.^{24,25}

Capacitive Probe Method

For the electret with one metallized surface and one bare surface, we can use the capacitive probe method to measure the total surface charge density σ_T .²⁶ With this method, a metal probe of area A is placed in parallel with the bare surface, with a relatively large air gap between them. The probe is connected to a large capacitor of capacitance C , much larger than the capacitance between the probe and the electrode, as shown in Figure 5-10.

A shutter shields the probe from the field of the electret. When the shutter is removed and the probe is exposed to the field from the charged electret, a charge of $-A\sigma_{up}$ will flow into the capacitor C creating a voltage across it. Thus

$$V = -\frac{A\sigma_{up}}{C} \quad (5-29)$$

The share of the induced charges on the lower electrode and the probe (i.e., the upper electrode) is similar to that under a short-circuit condition. From Equation 5-18,

$$\sigma_T = -(1 + \epsilon_r s/d)\sigma_{up} = (1 + \epsilon_r s/d)\frac{CV}{A} \quad (5-30)$$

The lateral charge distribution is usually different from the charge distribution in the bulk. If the probe can be made small enough, the lateral charge distribution can be determined by this capacitive probe method. This method can also

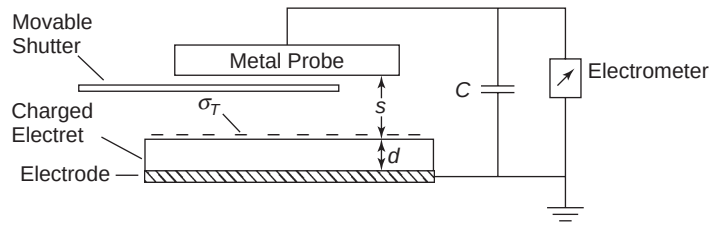


Figure 5-10 Schematic diagram showing the capacitive probe method for the measurement of the total surface charge density in a charged electret.

be used for the electrets with both surfaces bare (nonmetallized).

5.4.2 Total Charges per Unit Planar Area Q_T

In this subsection we shall describe two methods for the measurements of the total charge.

Electrostatic Induction or Faraday Pail Method

The Faraday pail method is a classic, basic electrostatic induction method for measuring total electrostatic charges in an electret with both surfaces bare (without metallic electrodes). To measure Q_T , we simply put the charged electret into a metallic Faraday pail, as shown in Figure 5-11. Based on the principle of electrostatic induction, the charges in the electret $-Q_T$ will induce $+Q_T$ on the inner surface and $-Q_T$ on the outer surface of the pail. If the collective capacitance of the system and the electrometer is C and the voltage measured by the electrometer is V , then

$$Q_T = CV \quad (5-31)$$

For electrets with both surfaces metallized, we must use the shaving technique to shave very thin layers of both metallic electrodes off the electret at low temperature to avoid any loss of the stored charges in the electrets. After the thin metallic layers have been removed, the same method can be used to measure Q_T .

Thermal Pulse Method

The total charges and the centroid of the charges can be measured by means of the

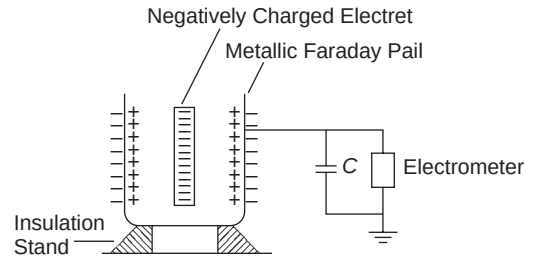


Figure 5-11 Experimental arrangement for the measurement of the total charges in a charged electret using a Faraday pail.

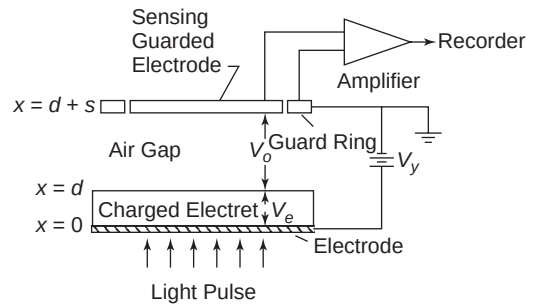


Figure 5-12 Schematic diagram showing the thermal pulse method for the measurement of the volume charge distribution in a charged electret.

thermal pulse method.²⁷⁻³⁰ The experimental arrangement for these measurements based on this method is shown schematically in Figure 5-12. The basic principle is to measure the change of the potential across the electret during diffusion of the thermal energy pulse generated by a short light pulse through the lower electrode at $x = 0$. Prior to the application of the thermal pulse, the voltmeter meas-

ures mainly the applied DC voltage V plus the voltage between the floating electrode and the lower electrode, which is, in fact, the difference between V_e and V_o . When a short light pulse illuminates the electret through the thin lower metallic electrode, the light energy will be absorbed by the electret material, generating a heat (thermal energy) pulse. The original potential at the bare surface is $V_e = \sigma_T d / \epsilon_r \epsilon_o$. The change of this potential by ΔV is due to the effect of the thermal pulse, which results in thermal expansion of the material (change in d) and variation of the dielectric constant (change in ϵ_r). Thus, by measuring the net change $\Delta V(t_1)$ taken at time t_1 immediately following the illumination, and $V(t_2)$ taken at time t_2 after the change has reached its thermal equilibrium, we can write

$$\frac{\sigma_T}{Q_T} = \frac{\Delta V(t_2)}{\Delta V(t_1)} \quad (5-32)$$

since $\Delta V(t_1)$ is related to σ_T due to the change of d and ϵ_r , while $\Delta V(t_2)$ is related to the diffusion of the volume charges from $x = 0$ to $x = d$. Therefore, to determine Q_T , we need two measurements: one to measure σ_T (from one of the methods given in Section 5.4.1) and the other to measure $\Delta V(t_1)$ and $\Delta V(t_2)$, using the thermal pulse method based on Equation 5-32.

The centroid of the volume charges $\int_o^d \rho(x) dx$ measured from the lower electrode $x = o$ is defined by

$$X_o = \frac{\int_o^d x \rho(x) dx}{\int_o^d \rho(x) dx} \quad (5-33)$$

If σ is much smaller than the volume charges $\int_o^d \rho(x) dx$, then σ can be ignored. Thus for this case, Equations 5-26 and 5-27 can be simplified to

$$\sigma_T = \frac{1}{d} \int_o^d x \rho(x) dx \quad (5-34)$$

$$Q_T = \int_o^d \rho(x) dx \quad (5-35)$$

From Equations 5-32 and 5-33, we obtain

$$\frac{X_o}{d} = \frac{\Delta V(t_2)}{\Delta V(t_1)} \quad (5-36)$$

So, using the thermal pulse method, we can also determine the centroid of the volume charges and the total volume charges $\int_o^d \rho(x) dx$ from Equation 5-36.

5.5 Charge Storage Involving Dipolar Charges

In general, all electrets consist of both dipolar and real charges. But in dipolar materials, such as semicrystalline polyvinylidene fluoride (PVDF), the dipolar polarization is dominant. In this case, the small contribution from the real charges, if any, can be ignored in order to simplify the analysis. In this section, we shall deal with cases involving only dipolar charges.

5.5.1 Basic Poling Processes

In Section 5.2.1, we discussed the thermo-electrical method of forming an electret by simultaneous application of an electric field and heat (at $T > T_g$). It should be noted that the average amount of the dipole moment per dipole aligning in the direction of the poling electric field F is less than u_o , which is the permanent dipole moment of each dipolar molecule in the dipolar material. The value of u_o is not affected by the field and the temperature. During the poling process, the polarization P increases with time because the amount of the dipole moments aligning in the direction of the field increases with time. Supposing that the material consists of only one type of dipoles, then the polarization P is governed by the Debye equation (see Electric Polarization and Relaxation in Time-Varying Electric Fields in Chapter 2)

$$\frac{dP}{dt} + \frac{P}{\tau_p} = \frac{\epsilon_o(\epsilon_{rs} - \epsilon_{r\infty})F}{\tau_p} \quad (5-37)$$

where ϵ_{rs} and $\epsilon_{r\infty}$ are, respectively, the static and optical dielectric constants and τ_p is the relaxation time of the dipoles, which is temperature dependent following an exponential relation

$$\tau_p = \tau_{p0} \exp(H/kT) \quad (5-38)$$

where H is the potential barrier height for the dipole orientation and the τ_{p0} is the average

time required by an excited molecule to turn from one equilibrium direction to the other.³¹

For the isothermal poling condition and the initial boundary condition $P(t = 0) = 0$ at $F = 0$, the solution of Equation 5-37 gives

$$P(t) = \epsilon_o(\epsilon_{rs} - \epsilon_{r\infty})F[1 - \exp(-t/\tau_p)] \quad (5-39)$$

When $P(t)$ reaches saturation, as shown in Figure 5-3, P becomes

$$P = \epsilon_o(\epsilon_{rs} - \epsilon_{r\infty})F \quad (5-40)$$

It should be noted that during the charging (poling) process, the dielectric constant ϵ_r is a function of temperature. But for the present analysis, we assume that the variation of ϵ_r with temperature is not appreciable and can be ignored. By doing so, the total charging current density can be written as

$$j_c(t) = dP(t)/dt + g(T)F \quad (5-41)$$

where $g(T)$ is the electrical conductivity of the material. Substitution of Equation 5-39 into Equation 5-41 yields

$$j_c(t) = \frac{P}{\tau_p} \exp\left[-\int_0^t \frac{1}{\tau_p} e^{-t/\tau_p} dt\right] + g(T)F \quad (5-42)$$

The first term on the right side of Equation 5-42 is the charging current, which is due to the orientation of the dipoles toward the direction of the poling field, and the second term is the conduction current due mainly to the motion of charge carriers thermally generated inside the material. The charging (polarization) current follows a transient process, giving rise to a peak if the temperature increases continuously, but the conduction current always increases with increasing temperature. So by measuring the $j_c - T$ characteristics at a constant applied electric field, we can separate the contribution of these two components to the total current j_c .

After the poling temperature has been cooled to the ambient temperature T_o and the applied electric field F has been removed, $P(t)$ will decay with time due to thermal relaxation at T_o . The decay process is governed by the following equation:

$$\frac{dP(t)}{dt} + \frac{P(t)}{\tau_p} = 0 \quad (5-43)$$

In this case, the initial boundary condition is at $t = 0$, $P(t = 0) = P(0) = P_o = \epsilon_o(\epsilon_{rs} - \epsilon_{r\infty})F$. Thus, the solution of Equation 5-43 is

$$P(t) = P(0)\exp(-t/\tau_p) \quad (5-44)$$

Since τ_p is a function of temperature, the rate of the decay depends on temperature: The lower the temperature, the slower the decay process. At a constant temperature, the aligned dipoles are constantly subjected to thermal relaxation, resulting in a continuous depolarization in the electret. So the stored charges can be said to be only quasi-permanent.

5.5.2 Relaxation Times of Dipoles and the Thermally Stimulated Discharge Current (TSDC) Technique

In many materials, and particularly in polymers, there exist several relaxation times, implying that there are several types of dipoles that may be associated with side groups of the molecular chains and their mutual interactions. These can be measured by means of the thermally stimulated discharge current (TSDC) technique. This is a very powerful technique for determining the distributions of the relaxation times of dipoles and the energy levels of carrier traps, and for studying the charge decay processes.^{32,33}

To measure the relaxation times of dipoles, we short-circuit the poled electret and measure the discharge current as a function of temperature at a constant rate of increase in temperature, as shown in Figure 5-13. To start with, we make the electret a fully poled electret, following the thermo-electrical poling process described in Section 5.2.1. After that, we measure the short-circuit discharge current j_d , that is, TSDC, while the electret is heated gradually by slowly increasing the temperature at a rate of $\beta = dt/dt$ (e.g., $\beta = 1^\circ\text{C min}^{-1}$). At $T = T_o$ during detention, $j_d = 0$, since the compensation charges on both electrodes are bound charges compensating the polarization charges in the material. When the temperature rises, $T > T_o$, the persistent polarization begins to decay, which means the average amount of dipole moments aligning along the direction of the poling field

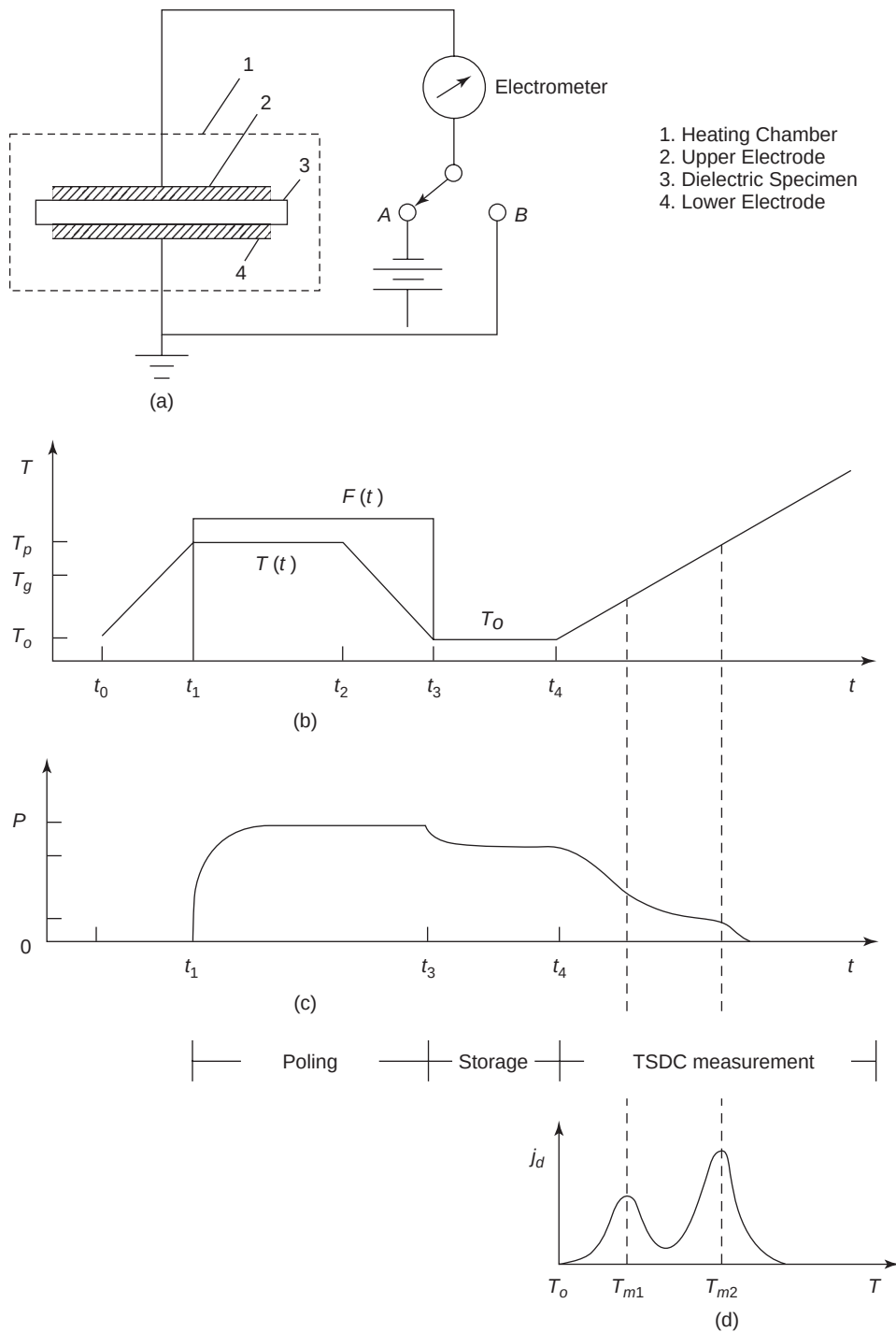


Figure 5-13 Schematic illustration of the experimental steps for the TSDC measurement: (a) arrangement for forming a thermo-electret—with switch on the A position, (b) the temperature-time and electric field-time sequence in the poling process, (c) the variation of polarization with time, and (d) the current j_d as a function of the temperature during TSDC measurement—with switch on the B position.

starts to decrease; hence, the amount of bound charges on the electrodes required for compensating the polarization charges also decreases. Some of the bound charges become released and free to contribute to the short-circuit current j_d . This discharge current is similar to the discharge current from a charged capacitor, which reveals the mechanism responsible for the polarization. From Equation 5-44, the discharging current density j_d is given by

$$j_d(t) = dP/dt = -P(t)/\tau_p \quad (5-45)$$

in which $P(t)$ cannot be expressed in the simple form given in Equation 5-44 because $P(t)$ during the TSDC process is also a function of temperature T and T is a function of t . For this reason, the appropriate expression for $P(t)$ is³³

$$P(t) = P(0) \exp \left[-\int_0^t \frac{dt}{\tau_p(t)} \right] \quad (5-46)$$

If the temperature is raised linearly with time at the rate β ($\beta = dT/dt$), Equation 5-45 can be written as

$$j_d(T) = -\frac{P(0)}{\tau_p(T)} \exp \left[-\frac{1}{\beta} \int_{T_0}^T \frac{dT}{\tau_p(T)} \right] \quad (5-47)$$

For a material having only one relaxation time, (i.e., only one type of dipoles), $P(0)$ can be expressed as

$$\begin{aligned} P(0) &= \epsilon_o [\epsilon_{rs}(T) - \epsilon_{r\infty}] F \\ &= \frac{N\mu_o^2 F}{3\epsilon_o kT} \end{aligned} \quad (5-48)$$

where N is the number of dipoles per unit volume. Since $\epsilon_{rs}(T)$ is temperature dependent, $P(0)$ is temperature dependent, but the temperature dependence of τ_p is much stronger than that of $P(0)$. To simplify matters, we will ignore the temperature dependence of $P(0)$ for the following analysis.

Since $\tau_p = \tau_{po} \exp(H/KT)$, the integral $\int_{T_0}^T \frac{dT}{\tau_p(T)}$ can be solved by partial integration, which leads to a solution in an asymptotic expansion.^{32,34} By expressing

$$\int_{T_0}^T \frac{dT}{\tau_{po} \exp(H/KT)} = \frac{H}{k\tau_{po}} \int_{T_0}^T e^{-H/KT} d\left(\frac{1}{H/KT}\right) \quad (5-49)$$

and letting $z = H/KT$, the integration would be in the following form:

$$\int e^{-z} d\left(\frac{1}{z}\right) = e^{-z} z^{-2} \left(1 - \frac{2!}{z} + \frac{3!}{z^2} - \frac{4!}{z^3} + \dots\right) \quad (5-50)$$

For most dielectric materials, $z = H/KT > 40$, and it is reasonable to truncate the series after the second term. However, this is still quite mathematically involved. So a reasonable approximation is needed in order to see easily the discharging process. An approximation in good agreement with experiments is given by

$$\begin{aligned} j_d(T) &= \\ C_1 \exp \left[-\frac{H}{kT} - \frac{C_2}{(H/kT)^2} \exp(-H/kT) \right] \end{aligned} \quad (5-51)$$

where C_1 and C_2 are adjustable constants.³⁵⁻³⁷ Equation 5-51 indicates that the thermally stimulated depolarization current starts out slowly, increases exponentially, and then reaches a peak value. After reaching the peak, it drops rapidly as the temperature is raised continuously, and then to zero when the arrangement of all dipoles becomes completely randomized. In Equation 5-51, the first term is dominant at low temperatures, that is, at $T < T_m$ where T_m is the temperature for the occurrence of the peak current. Thus, the slope of the $j_d - T$ curve, based on Equation 5-51, at low temperatures would yield the value of H as follows:

$$\frac{d[\ln j_d(T)]}{dT} = -\frac{H}{k} \quad (5-52)$$

By differentiating j_d with respect to T and setting it to zero, we would find the temperature T_m at which j_d becomes maximal. Under this condition, we have the following relation

$$\left. \frac{d\tau_p}{dT} \right|_{T=T_m} = -\frac{1}{\beta} \quad (5-53)$$

Substitution of Equation 5-38 into Equation 5-53 yields

$$\frac{kT_m^2}{\beta H \tau_p(T_m)} = 1 \quad (5-54)$$

Thus, from Equation 5-52 we can determine H , the potential barrier height that the dipoles must

surmount in order to proceed with the orientation. From Equation 5-54 we can determine τ_p at T_m , which can be found from the TSDC thermogram ($j_d - T$ curve).

In dipolar materials the TSDC peak may arise from a single relaxation with one relaxation time or from a distributed relaxation with several relaxation times. When TSDC thermograms are measured under various poling conditions, the shape and the temperature for the occurrence of the current peak do not depend on the poling conditions if there is one relaxation time, but they do if there are several relaxation times. This is because the fast- and slow-relaxing dipoles behave differently for different electret formation and storage times.^{32,33} A good example with a single relaxation can be found in ionic crystals, such as alkali halides, doped with divalent impurity ions, which coupled with their accompanying vacancies, will form the $I-V$ dipoles.³³

In many materials and particularly in polymers, however, there exist several relaxations, the most common ones being the γ , β , α , and ρ relaxations. The γ relaxation is associated with the motion of the dipoles within the side groups of the molecular chains, the β relaxation is due to the motion of the side groups themselves, and the α relaxation is caused by the joint motions of the side groups and the main chains. These three relaxations are directly related to the reorientation of dipoles, while the ρ relaxation is due mainly to the motions of real charges (space charges). Depending on the difference in activation energy ΔH between two types of dipoles, the two current peaks in the TSDC thermogram may merge to form one broad peak if ΔH is smaller than a few kT ; otherwise, two separate peaks will appear. For PMMA (polymethyl methacrylate), four current peaks associated with the four relaxations can be monitored clearly, as shown in Figure 5-14. If the material possesses several groups of different relaxations, then Equation 5-38 must be written as the sum of a number of relaxations

$$\tau_p = \sum_i \tau_{pi} = \sum_i \tau_{poi} \exp\left(\frac{H_i}{kT}\right) \quad (5-55)$$

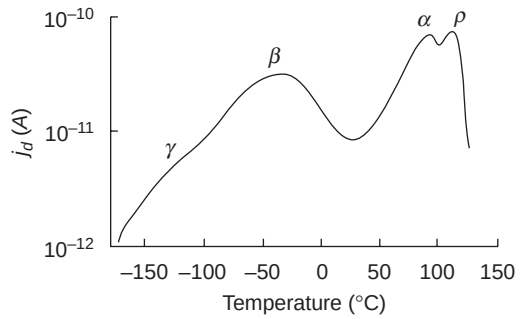


Figure 5-14 Thermally stimulated discharge current (TSDC) thermogram (recorded at 5°C min^{-1}) of polymethyl methacrylate (PMMA) electret, specimen area being 19.6 cm^2 and poled at 10 kV cm^{-1} at 127°C for 30 min.³³

where the subscript i specifies the i th group of dipolar relaxation. Each group will contribute to the polarization and the TSD current, so the total polarization or total TSD current is simply the algebraic sum of the contributions from all individual groups.

It is also important to note that Equation 5-38 holds only for relaxations involving the reorientation of small dipoles. It becomes invalid for relaxations involving large dipoles in macromolecular materials such as polymers—particularly when the temperature reaches the glass–rubber transition state in which the motion of the large dipoles involves configurational rearrangement of parts of the long main chains because the motion of such bulky dipoles requires some unoccupied space. In this case, we must use the WLF empirical relation (WLF stands for Williams, Landel, and Ferry)^{32,33} instead of Equation 5-38.

$$\tau_r(T) = \tau_g \exp[-C_a(T - T_g)/(C_b + T - T_g)] \quad (5-56)$$

for $T > T_g$

where τ_g is the relaxation time in the glass state, T_g is the glass transition temperature, and C_a and C_b are constants. By approximation, Equation 5-56 may be reduced to a form similar to Equation 5-38, as follows

$$\tau_r = \tau_w \exp[H_w/k(T - T_w)] \quad (5-57)$$

where $\tau_w = \tau_g \exp(-C_a)$, $H_w = C_a C_b k$, and $T_w = T_g - C_b$. This relation becomes invalid for $T < T_g$.

5.5.3 Spatial Distribution of Dipolar Polarization

If the poling field is uniform in the material specimen, the spatial distribution of the dipolar polarization is expected to be uniform. However, if volume space charges also exist inside the specimen, the charges will be separated: positive charges moving to the negative electrode and negative charges moving to the positive electrode, forming a space charge polarization. These heterocharges near the forming electrodes tend to lower the applied poling field near the center of the specimen, resulting in a smaller dipolar polarization in the center than that in the region near the electrodes, because dipolar polarization is field dependent. If the poling process also involves charge carrier injection, then the homocharges will cause the distribution of the field and hence the dipolar polarization will be nonuniform spatially in the specimen.

The spatial distribution of dipolar polarization can be determined simply by the sectioning method, which is discussed in Section 5.6.1. For dipolar polarization measurements, thin metallic electrodes must be deposited by fast evaporation of gold on both surfaces of each slice cut out from the electret to avoid heating the specimen. Then we can measure the discharging current using the TSDC technique. The polarization can then be evaluated by integration of the TSDC spectrum based on $P = \int_0^\infty J_d dt$. If the electret is cut into many slices, the distribution of $P(x)$ can be obtained by assembling the data from each slice.

Let us take a PMMA electret as an example. The electret was cut into three slices: one near the center and other two near the electrodes. Figure 5-15 shows the TSDC thermograms of the slices cut out from the PMMA electret. The data of these curves are from Tounhout.³² It can be seen that the polarization near the center of the electret is significantly smaller than near the two electrodes, indicating clearly that the PMMA electret has heterocharges near the electrodes that cause the lowering of the poling field and hence the polarization near the center, since polarization increases with increasing

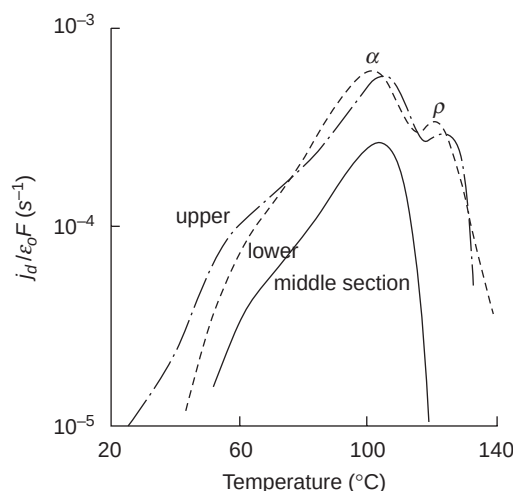


Figure 5-15 TSDC thermograms of three sectional slices of a PMMA electret showing the slices near the upper and lower electrodes having a higher polarization than the slice from the center. The specimen thickness was 4.8 mm and poled at 20 kV cm^{-1} at 140°C for 1.5 hours; j_d recorded at $0.5^\circ\text{C min}^{-1}$.³²

poling field. This implies that the electret has not only the dominant dipolar charges but also real space charges.

5.5.4 Isothermal Polarization Decay Processes

The retention of the dipolar charges depends on the structure of the material, which is directly related to τ_p and the operating and environmental conditions. In general, dielectric materials exhibiting good dipolar properties have a relatively high conductivity, due partly to their hygroscopic behavior. These materials, such as ceramic materials and some dipolar polymers, tend to absorb moisture from the surrounding environment. Dielectric materials with a low conductivity are usually not dipolar. The value of τ_p depends on the material parameters, such as the crystallinity and the arrangement of various dipolar groups (various types of dipoles), particularly for organic polymers, which are directly associated with the structure of the material³⁸; the dopants, which are chosen to increase the potential barrier height responsible for dipolar relaxation so as to improve the

retention time^{39,40}; crosslinking for polymers; hygroscopic behavior, and so forth. The effect of crystallinity on dipolar polarization is severe. For example, the β peak of the TSDC thermogram for polycarbonate (PC) decreases with increasing degree of crystallinity.

Isothermal depolarization in the electret results from the return of the partially aligned dipoles to the randomized state that they had in thermal equilibrium prior to the poling process. For materials involving only one type of dipolar relaxation, the decay process would be expected to follow Equation 5-44. But for materials involving several groups of different types of dipolar relaxations, the decay process would follow the following relation

$$P(t) = \sum_i P_i(t) = \sum_i P_{oi} \exp(-t/\tau_{pi}) \quad (5-58)$$

where τ_{pi} and P_{oi} are, respectively, the relaxation time and the preexponent factor for the i th type of dipoles. A typical isothermal polarization decay curve at room temperature is shown in Figure 5-16. The data are from Vander-schueren and Linkens.⁴¹ It can be seen that the curve does not follow a simple exponential function, indicating that polarization involves the contributions of several different dipolar groups. It is likely that those with small relaxation times decay faster, leaving those with large relaxation times to decay slowly.

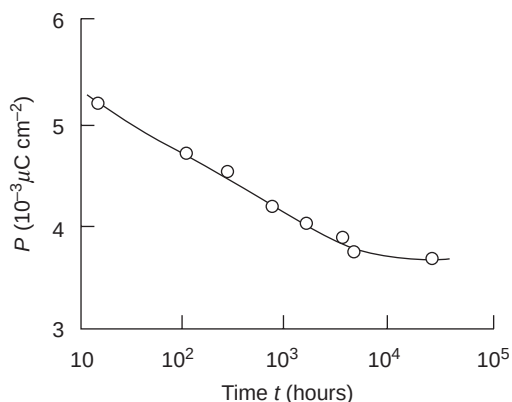


Figure 5-16 Isothermal decay of dipolar polarization in PMMA at room temperature. The electret was poled at 135°C.⁴¹

5.6 Charge Storage Involving Real Charges

In nondipolar materials, the charge storage is due mainly to the injected real charges; the contribution of dipolar charges, if any, can be ignored. All dielectric materials, such as polymers, consist of many deep and shallow traps. The quasi-permanent retention of real charges happens because these traps are capable of keeping the captured electrons or holes for a long period of time, creating trapped space charges (see Bulk-Limited Electrical Conduction in Chapter 7). The trapped charges may be located on the surface and in the bulk. The surface traps may be associated with chemical impurities, such as adsorption of foreign molecules, oxidation of the surface layer, structural discontinuity resulting from broken chains, etc. The bulk traps may be associated with material structure and chemical impurities. For most practical dielectric materials, the structure is either amorphous, polycrystalline, or partially crystalline. Therefore, local energy states are created in the energy band gap, forming distributed trapping energy levels.⁴² The origins of the surface traps and the bulk traps are still not fully understood. In this section, we shall discuss some commonly used methods for measuring spatial and energy level distributions of bulk traps.

5.6.1 Spatial Distributions of Trapped Real Charges

Sectioning and potential probe are the old, classic methods that have long been used to measure the spatial distribution of trapped real charges. Obviously, these methods are not suitable for thin electrets, although they have been used extensively on thick electrets. However, there are other methods for either thin or thick electrets. We shall review briefly some of the most commonly used methods.

Sectioning Method

For electrets with a relatively large thickness, the electret can be cut into several thin slices with a thin saw cutter at a low cutting speed

along planes perpendicular to the direction of the poling field, as shown in Fig 5-17. It is important to avoid loss of the stored charges and to prevent the cutter contact from charging during cutting, so the cutting should be performed at a low temperature.^{32,43,44} By putting each slice into a Faraday pail (see Section 5.4.2), we can determine the net trapped real charges in slice i of thickness d_i . Since the dipolar polarization within that slice has zero net charge, the total real trapped charges measured in slice i are

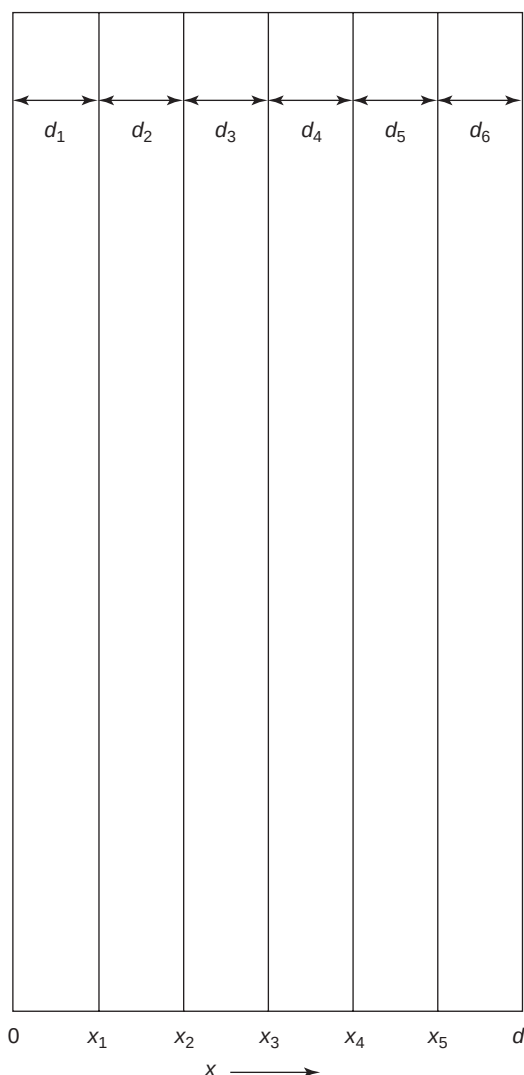


Figure 5-17 The sectioning of an electret into six slices.

$$q_i = \int_0^{d_i} \rho_r(x_i) dx \quad (5-59)$$

We can also assume that the distribution of the charges within the very thin slice is uniform. Thus, the trapped real charge density at $x = x_i$ can be expressed as

$$\rho_r(x_i) = q_i/d_i \quad (5-60)$$

All slices may not have the same thickness, but the spatial distribution can be easily obtained from Equation 5-60.

Instead of cutting the electret into slices, we can use the shaving technique. By shaving a thin layer off the electret step by step, measuring the charges on the shaved-off layers or the charges on the remainder of the electret, and also measuring the change of thickness of the electret, we can assemble the data to obtain the information of the spatial distribution of the trapped charges.⁴⁴

Electron Beam Sampling Method

When a charged electret with both surfaces metallized is irradiated with an electron beam through one electrode, as shown in Figure 5-18, the energetic electrons penetrating into the electret will create a conductive region, which may extend from the upper irradiated electrode to a certain depth. This conductive region acts as a virtual electrode, which is equivalent to a short circuit of this region. If the front of the conductive region is swept through the whole electret by gradually increasing the energy of the electron beam, then the charges originally stored in the electret will be gradually removed. This implies that the originally induced charges on the lower electrode will be gradually released to flow to capacitor C , and the amount of charges removed can then be measured with an electrometer. The change of the induced charges on the lower electrode q due to the front of the conductive region reaching x , will yield the charge density $\rho(x)$ at x following the relation

$$\rho(x) = \frac{d^2(qx)}{dx^2} \quad (5-61)$$

where q is the induced charges per unit area on the lower electrode and x is the thickness of the nonconductive region.^{45,46}

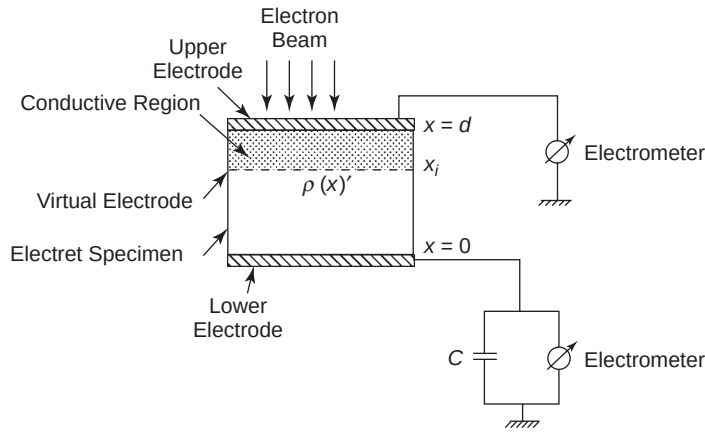


Figure 5-18 The experimental arrangement for the electron beam sampling method for measuring charge distribution in x direction.

Pulsed Electro-Acoustic Method

The pulsed electro-acoustic (PEA) method⁴⁷⁻⁴⁹ is now commonly used to measure the spatial distribution of space charges in dielectric materials, partly because it has the advantages of relatively simple experimental arrangement and comparatively easy interpretation of the experimental results. The general experimental arrangement for this method is shown in Figure 5-19(a).

The basic principle is as follows: When an electric field pulse is applied to the electret, it induces a perturbation force that causes a movement of the charges. This movement generates acoustic waves that propagate in the x direction in the material. A piezoelectric transducer (sensor) attached beneath the lower electrode converts the acoustic waves into electrical signals, which after amplification, can be monitored by a high-speed digital oscilloscope. The amplitude of the signals is related to the charge density, and the delay is related to the distance from the charges to the lower electrode. Thus, spatial distribution of the space charges can be measured. However, to use this method, the following conditions must be met:

1. The width of the electric field pulse Δt_p must be much smaller than the transit time of the acoustic waves across the electret $\Delta t_s = d/u_s$,

where u_s is the velocity of the acoustic wave (i.e., sound speed) in the material.

2. The acoustic impedance of the electret Z_s and that of the electrode Z_l must match each other. To make $Z_s = Z_l$, a piece of conductive material is usually placed between the electret and the electrode to satisfy the matching condition.

When the acoustic waves propagate through the electret and reach the sensor, the acoustic signal consists of three components: the induced charges on the lower electrode $\sigma(x=0)$, the induced charges on the upper electrode $\sigma(x=d)$, and the charges inside the electret $\rho(0 < x < d)$. This can be expressed as

$$P(t) = \frac{Z_l}{Z_s + Z_l} \left[\sigma(0)e_p(t) + \mu_s \int_0^t \rho(\tau)e_p(t-\tau)d\tau + \sigma(d)e_p\left(t - \frac{d}{\mu_s}\right) \right] \quad (5-62)$$

where $e_p(t)$ is the amplitude of the electric field pulse.⁴⁹

In order to achieve good and reliable results, a proper matching of the acoustic impedances is necessary to prevent any possible reflection of the acoustic waves at the interface between any two different materials. After being con-

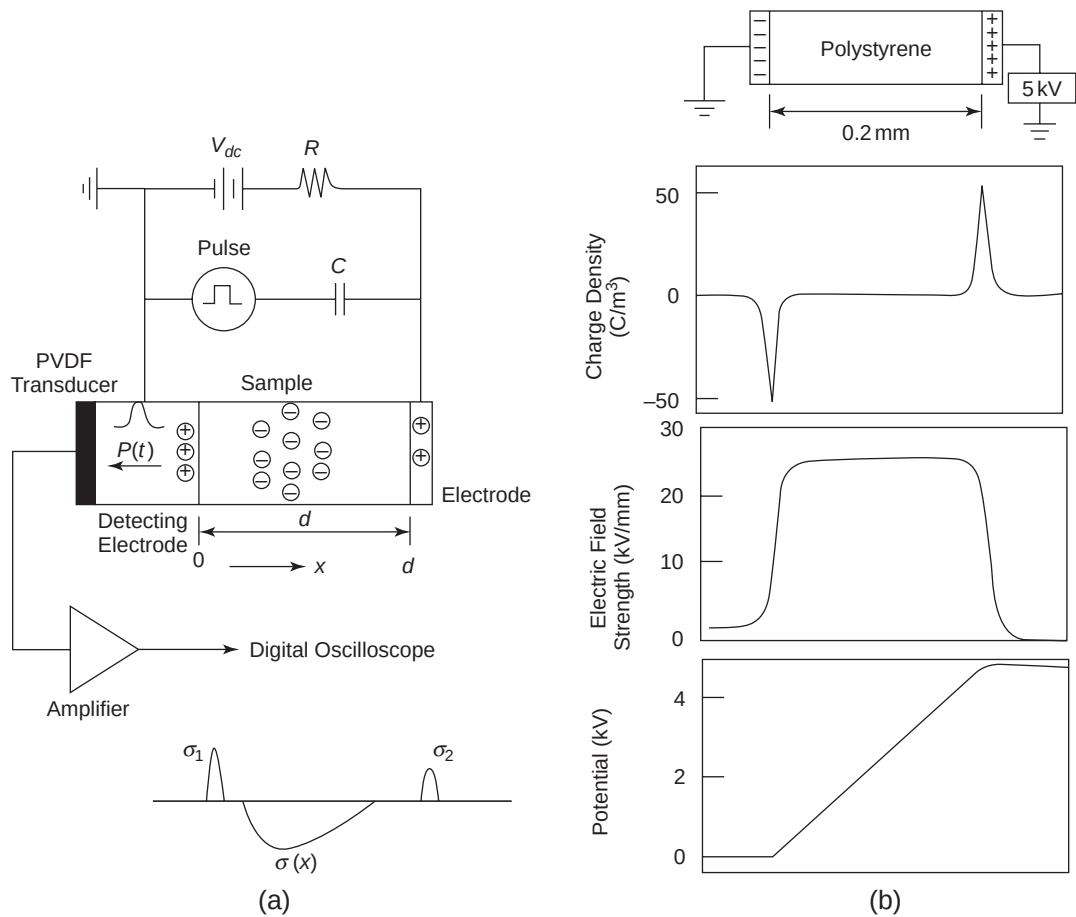


Figure 5-19 (a) Schematic diagram showing the pulsed electro-acoustic (PEA) method for the measurement of the spatial distribution of space charges in a charged dielectric solid; (b) an example of using polystyrene between two metallic electrodes for the calibration of charge density, electric field, and potential by the PEA method. The data are from Fukunaga.⁵³

verted to an electrical signal in the frequency domain, the acoustic wave signal is then changed by inverse Fourier transform to the space charge distribution in the time domain or in the position domain $\rho(x)$, where x is in fact $x = \mu_s \tau$ and τ is the time required for the acoustic wave signal to travel from the charges to the sensor.

The electrical signal must be calibrated. The usual way to calibrate the charge density is to use an uncharged dielectric material, such as polystyrene with two electrodes. For example, by applying 5 kV DC across a polystyrene specimen of 200 μm in thickness and then measur-

ing the charge distribution using the PEA system, we can observe only the charges on the two electrodes, because there are no internal space charges in the specimen. A typical result is shown in Figure 5-19(b).

In this case, the components of the induced charges on the electrodes can be calculated simply by $(\epsilon_r \epsilon_o V_{dc}/d) (\mu_s \Delta \tau)^{-1}$. For details of the PEA method, see references.⁴⁷⁻⁴⁹

Figure 5-20 shows typical spatial distribution of the space charges in a charged PMMA specimen of 500 μm in thickness with the two metallic electrodes short-circuited. Prior to the measurement, the specimen had been charged

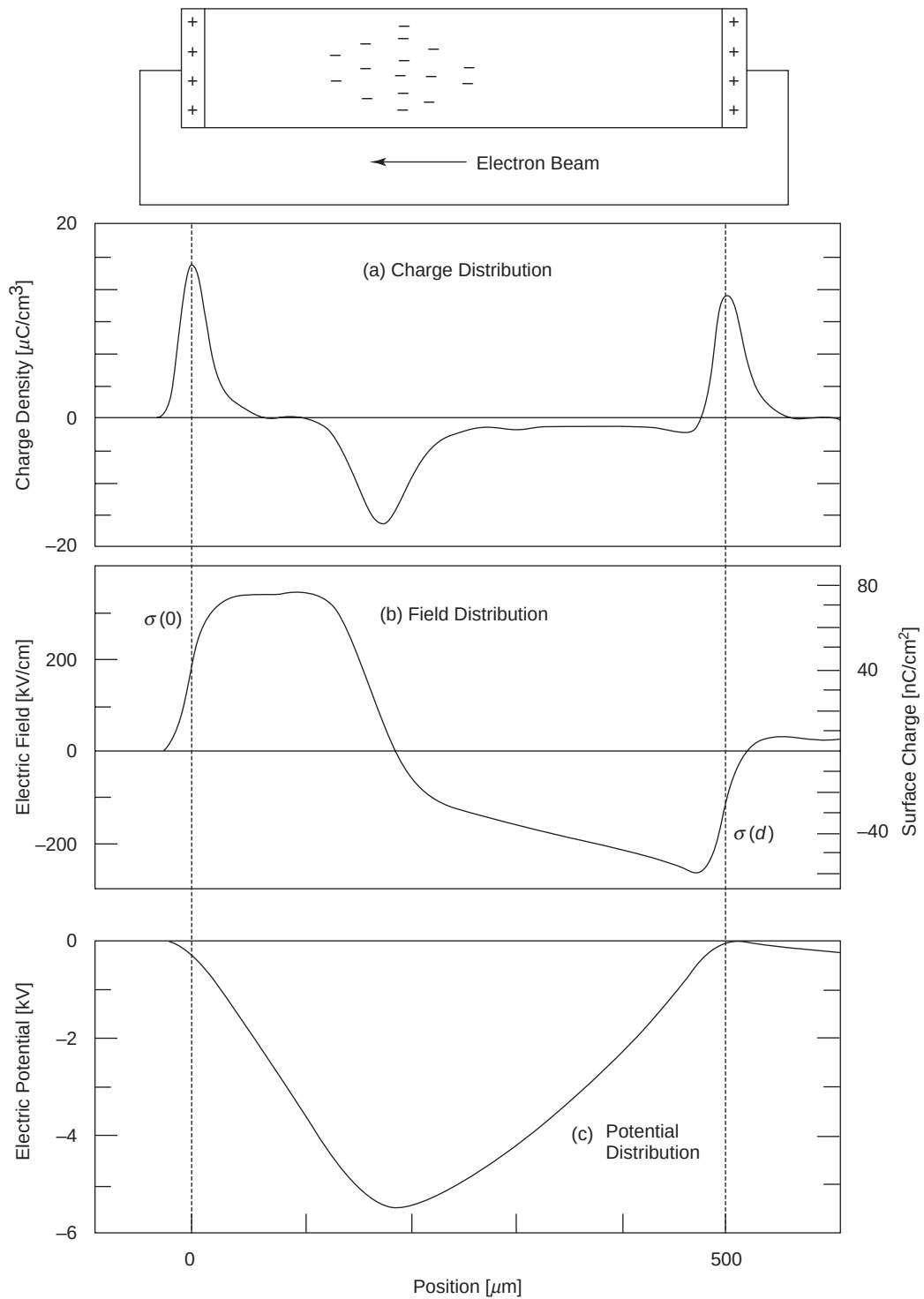


Figure 5-20 Spatial distribution of the trapped negative charge density, the electric field, and the electric potential in a PMMA specimen with negative space charges injected by an electron beam under the short-circuit condition. The data are from Takada.⁴⁹

with negative space charges (electrons) by an electron beam of 200 keV and $1 \times 10^{-9} \text{ A cm}^{-2}$ for three hours. The data are from Takada.⁴⁹

Other Methods

There are several other methods that can be used for the measurements of spatial distribution of space charges,⁵⁰⁻⁵³ such as pressure wave propagation (PWP), thermal step, laser-induced pressure pulse, and so on. Discussion of these methods is beyond the scope of this chapter. For further information, see the references cited previously.

5.6.2 Energy Distribution of Trapped Real Charges

Electrons (or holes) trapped in localized states in the energy band gap can be released and become free to move in the charged electret, provided that they are given sufficient energy to do so. Most electrets are made of amorphous or polycrystalline polymers with a high degree of structural disorder, implying that they have a high concentration of deep and shallow traps. In particular, deep traps are important for good electrets because they can hold the trapped charge carriers (electrons or holes) for a prolonged period of time. Suppose that electrons are the only type of monocharge trapped inside the electret; the trapped electrons during the heating process under short-circuit and TSDC conditions will gain enough energy to jump to the conduction band and drift by the internal field $F(x,t)$ inside the electret toward the electrodes. On their way, they may be retrapped and then detrapped again; ultimately, they will recombine with their compensating charges at the electrodes.

A simple approach to the analysis of TSDC processes is to assume that only one deep trap level acts at a given time. Here, we consider the case with electrons as the only type of real charges and electron traps situated at one energy level. In this case, the TSDC process for a charged electret under a short-circuit condition is governed by the following equations: The Poisson equation:

$$\frac{\partial F(x,t)}{\partial x} = -\frac{q[n(x,t) + n_t(x,t)]}{\epsilon_r \epsilon_0} \quad (5-63)$$

The continuity equation:

$$\frac{\partial n(x,t)}{\partial t} = \frac{\mu(T) \partial [n(x,t) F(x,t)]}{\partial x} - \frac{\partial n_t(x,t)}{\partial t} - \frac{n(x,t)}{\tau_\gamma} \quad (5-64)$$

The trapping kinetics equation:

$$\frac{\partial n_t(x,t)}{\partial t} = C_n n(x,t) [N_t - n_t(x,t)] - n_t v \exp\left[\frac{-\Delta E}{kT}\right] \quad (5-65)$$

where $n(x,t)$ and $u(T)$ are, respectively, the concentration and the average mobility of free electrons; $n_t(x,t)$ and N_t are, respectively, the concentrations of the trapped electrons (occupied traps) and the total traps (empty and occupied traps together); $F(x,t)$ is the internal field created by the space charges; τ_γ is the recombination lifetime of the thermally released charge carriers; C_n is the electron capture coefficient; v is the attempt-to-escape frequency; $\Delta E = E_c - E_t$ is the trap energy level measured from the conduction band edge; and E_c and E_t are, respectively, the energy levels of the conduction band edge and the traps. The first term on the right side of Equation 5-65 represents the retrapping and the second term the detrapping. C_n can be expressed as

$$C_n = \langle v \sigma_n \rangle = \frac{v}{N_c} \quad (5-66)$$

where σ_n and v are, respectively, the trap capture cross section and the mean thermal velocity of the electrons, and N_c is the effective density of states in the conduction band (see Physical Concepts of Carrier Trapping and Recombination in Chapter 7).

The thermally stimulated discharging current density is given by

$$j_d(t) = \epsilon_r \epsilon_0 \frac{\partial F(x,t)}{\partial t} + q \mu(T) n(x,t) F(x,t) \quad (5-67)$$

Using the short-circuit condition as the boundary condition, which gives

$$\int_0^d F(x,t)dx = 0 \quad (5-68)$$

then the integration of Equation 5-67, with respect to x , yields

$$j_d(t) = \frac{q\mu(T)}{d} \int_0^d n(x,t)F(x,t)dx \quad (5-69)$$

This equation cannot be solved analytically without resorting some approximations because $n(x,t)$ and $F(x,t)$ are governed by the three partial differential equations. For the simplest case of fast retrapping, each electron released from the trap will soon be retrapped, implying that the rate of detrapping is approximately equal to the rate of retrapping. Thus, we may set $\frac{\partial n_t(x,t)}{\partial t} \rightarrow 0$ and assume $N_t > n_t(x,t)$.

Then Equation 5-65 can be reduced to

$$n(x,t) = \frac{v \exp(-\Delta E/kT)}{C_n N_t} n_t(x,t) \quad (5-70)$$

Furthermore, we can also assume $n_t(x,t) > n(x,t)$. By substituting Equations 5-70 and 5-61 into Equation 5-69 and based on the rate of linear rise of the temperature in the TSDC process ($\beta = dT/dt$), we obtain

$$j_d(T) = \frac{\varepsilon_r \varepsilon_0 \mu(T)}{2d} \left[\frac{v \exp(-\Delta E/kT)}{C_n N_t} \right] \times [F^2(o,T) - F^2(d,T)] \quad (5-71)$$

The parameters $u(T)$, v , and C_n are functions of temperature. The fields at the electrode $x = 0$ and at the electrode $x = d$ are also functions of temperature. If we can find the temperature dependence of these parameters, the temperature T_m for the occurrence of the current peak can be easily obtained. However, Equation 5-71 indicates that if the centroid of the space charges at normal temperature T_o is located at x_o , which is smaller than $d/2$, then the field $F(0,T)$ is larger than $F(d,T)$ and there is an external current flowing through the short circuit, as shown in Figure 5-21. As temperature increases, the current flow will cause the centroid x_o to move toward $d/2$. When x_o reaches $d/2$, then the net current flow vanishes.

The TSDC process is used mainly to increase the rate of detrapping by increasing the temperature with time. The peak current occurs at $T = T_m$, corresponding to a peak value of $n(x, T_m)$ or $dn(x, T)/dT|_{T=T_m} = 0$ and the position of $x_o(T_m)$ located beyond $x_o(T_o)$ but below $d/2$.

Because of fast retrapping, recombination lifetime is very large, so both terms $dj_d(x,t)/dx$

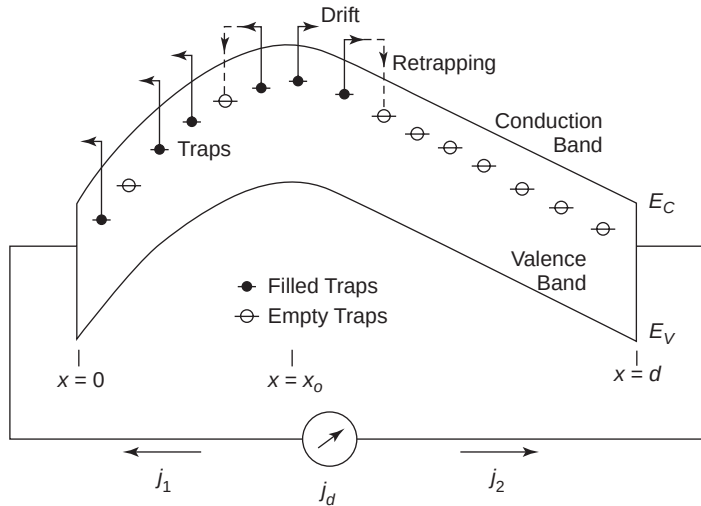


Figure 5-21 Energy band diagram during discharging from a charged electret under the short-circuit and TSDC conditions, with discharging electrons flowing in opposite directions.

and $n(x,t)/\tau_r$ can be neglected by approximation. By doing so, Equation 5-64 can be reduced to

$$\frac{\partial n(x,T)}{\partial T} = \frac{\partial n_i(x,T)}{\partial T} \quad (5-72)$$

Since $dn(x,T)/dT = 0$ at $T = T_m$, we can obtain ΔE directly from Equation 5-70:

$$\begin{aligned} \Delta E &= kT_m \ln \left[\frac{vn_i}{C_n N_i n} \right] \\ &= kT_m \ln \left[\frac{N_c}{n} \frac{n_i}{N_i} \right] \\ &= kT_m \ln(n_i/N_i) + (E_c - E_F) \end{aligned} \quad (5-73)$$

as n is given by

$$n = N_c \exp \left(-\frac{E_c - E_F}{kT} \right) \quad (5-74)$$

To use Equation 5-73, other experiments may be necessary for finding the value of N_i , $n_i(T_m)$, and $E_c - E_F(T_m)$. If the traps are created by the doped impurities, N_i may be easily determined, but if they are due to disorder in the structure, then it must be measured by a separate experiment. It is also necessary to know the trap occupancy at T_m in order to calculate $n_i(T_m)$. However, even for this simplest case, the evaluation of ΔE is quite involved. Obviously, different approximations based on different assumptions for different cases, such as fast retrapping, slow retrapping, quasi-equilibrium, and so on, would lead to different solutions of Equation 5-69. For more information about trap energy levels, see references 32, 33, 54–58 in this chapter and also Physical Concepts of Carrier Trapping and Recombination in Chapter 7.

5.6.3 Isothermal Real Charge Decay Processes

The density of the current available to flow out from the electret at a constant temperature T can be written as

$$j_e(t) = [g(T) + \mu_n(T)\rho_-(x,t) + \mu_p(T)\rho_+(x,t)]F(x,t) \quad (5-75)$$

where $g(T)$ is the intrinsic conductivity of the material, $\rho_-(x,t)$ and $\rho_+(x,t)$ are, respectively, the thermally released negative and positive

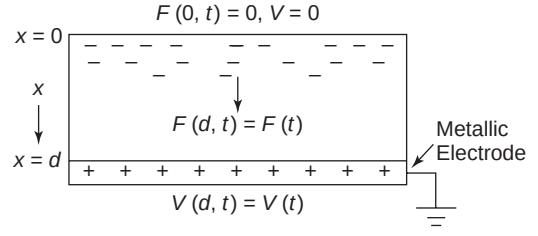


Figure 5-22 An electret with the stored negative charges drifting toward the metallic electrode and ground.

charge densities from traps; $u_n(T)$ and $u_p(T)$ are, respectively, the mobilities of the negative and positive charge carriers; and $F(x,t)$ is the internal field created by the space charges. Since the thermally released charge carriers are subjected to a detrapping and retrapping process, the charge carriers can be considered moving at the trap-modified mobilities. Here, we will consider the electret having only the negatively trapped charges; the dipolar charges (dipolar polarization) and the possible presence of some positive charges will be ignored.

The internal field created by the trapped negative charges $F(x,t)$ follows the Poisson equation:

$$\frac{dF(x,t)}{dx} = \frac{\rho}{\epsilon_r \epsilon_0} \quad (5-76)$$

The motion of the released charges also causes the internal field to change with time, which in turn gives rise to a displacement current, $\epsilon_r \epsilon_0 dF(x,t)/dt$. For the electret with a metallic electrode deposited on the lower surface and the upper surface bare, as shown in Figure 5-22, and with all stored charges assumed to be located initially on the bare surface at $t = 0$, then under the open-circuit condition, Equation 5-75 can be rewritten as

$$j_e(t) = \epsilon_r \epsilon_0 \frac{\partial F(x,t)}{\partial t} + \left[g(T) + \mu_n(T) \epsilon_r \epsilon_0 \frac{\partial F(x,t)}{\partial x} \right] F(x,t) = 0 \quad (5-77)$$

This nonlinear partial differential equation can be solved only under certain conditions. The basic boundary conditions are

$$\begin{aligned}
V(0,t) &= 0 \\
V(d,t) &= V(t) \\
F(0,t) &= 0 \\
F(d,t) &= F(t)
\end{aligned} \tag{5-78}$$

The charges are located at $x = 0$ at $t = 0$, so at $t = 0$ just prior to the movement of the released charges toward the lower electrode, we have

$$\begin{aligned}
V(0,0) &= 0 \\
V(d,0) &= V_0 \\
F(0,0) &= 0 \\
F(d,0) &= V_0/d
\end{aligned} \tag{5-79}$$

We also introduce the following parameters

The relaxation time of the dielectric material:

$$\tau = \varepsilon_r \varepsilon_0 / g \tag{5-80}$$

The transit time of the charge carrier traveling across the specimen:

$$t_t(t) = \frac{d^2}{\mu_n V(t)} \tag{5-81}$$

Obviously, the theoretical transit time under the initial condition (i.e., at $t = 0$) is

$$t_t(0) = \frac{d^2}{\mu_n V_0} \tag{5-82}$$

We can solve Equation 5-77 in two stages: stage 1 is for $t \leq t_{io}$ and stage 2 is for $t \geq t_{io}$, where t_{io} is the time for the charge carriers to reach $x = d$.

Stage 1: For $t \leq t_{io}$

Since $F(t)$ at $x = d$ depends on the arrival of the negative charges to the electrode, for $t \leq t_{io}$, the only charges contributing to $F(t)$ are those from the intrinsic conduction. There are no contributions from the stored charges because $\rho(x = d, t < t_{io}) = 0$. Thus, for this stage, Equation 5-77 is reduced to

$$\varepsilon_r \varepsilon_0 \frac{\partial F(x,t)}{\partial t} + g(T)F(x,t) = 0$$

or

$$\frac{\partial F(x,t)}{\partial t} + \frac{F(x,t)}{\tau} = 0 \tag{5-83}$$

The solution of Equation 5-83 for $x = d$ is

$$\begin{aligned}
F(d,t) &= F(t) = F(d,0)e^{-t/\tau} \\
&= \frac{V_0}{d} e^{-t/\tau}
\end{aligned} \tag{5-84}$$

Although $\rho(d, t < t_{io}) = 0$, $\rho(x < d, t < t_{io}) \neq 0$, and this will contribute to $V(t)$. Thus, by integrating Equation 5-77 with respect to x and substituting Equation 5-84 into it, we obtain

$$\frac{dV(t)}{dt} + \frac{g}{\varepsilon_r \varepsilon_0} V(t) + \frac{\mu_n}{2} \left[\frac{V_0}{d} e^{-2t/\tau} \right] = 0 \tag{5-85}$$

Solving Equation 5-85 for $V(t)$, we obtain^{59,60}

$$\frac{V(t)}{V_0} = \left[1 - \frac{1}{2} \left(\frac{\tau}{t_{io}} \right) (1 - e^{-t/\tau}) \right] e^{-t/\tau} \tag{5-86}$$

Because g is usually very small for good dielectric materials, $\tau \gg t$. Thus, for $t \geq t_{io}$ we can use the approximation

$$e^{-t/\tau} = 1 - t/\tau \tag{5-87}$$

So $V(t)/V_0$ can be approximately expressed as⁶¹

$$\begin{aligned}
\frac{V(t)}{V_0} &= \left[1 - \frac{1}{2} \frac{t}{t_{io}} \right] \left(1 - \frac{t}{\tau} \right) \\
&= 1 - \frac{t}{2t_{io}} \quad \text{for } t \leq t_{io}
\end{aligned} \tag{5-88}$$

Stage 2: For $t \geq t_{io}$

Once the charges can reach the electrode, we can assume those charges to be uniformly distributed spatially inside the electret. Thus, solving Equation 5-76 and using the boundary condition $F(0,t) = 0$, we obtain

$$F(x,t) = \frac{\rho(x,t)}{\varepsilon_r \varepsilon_0} x \tag{5-89}$$

By integrating Equation 5-89 with respect to x , we obtain the potential at $x = d$ as

$$V(t) = \frac{1}{2} F(d,t) d \tag{5-90}$$

Now, by integrating Equation 5-77 with respect to x and substituting Equation 5-90 into it, we obtain

$$\frac{dV(t)}{dt} + \frac{V(t)}{\tau} + \frac{2\mu_n}{d^2} V^2(t) = 0 \tag{5-91}$$

Solving Equation 5-91, we obtain^{59,60}

$$\frac{V(t)}{V_o} = \frac{1}{2} \frac{t_{io}}{\tau} \frac{e^{-t/\tau}}{(1 - e^{-t/\tau})} \quad (5-92)$$

We can use the same approximation given by Equation 5-87. Therefore, $V(t)/V_o$ can be approximately expressed as⁶¹

$$\frac{V(t)}{V_o} = \frac{1}{2} \frac{t_{io}}{t} \quad \text{for } t \geq t_{io} \quad (5-93)$$

It can be seen from Equations 5-88 and 5-93 that when g is negligibly small and τ is very large, $V(t)$ decays linearly with time for $t \leq t_{io}$ and then decays hyperbolically with time for $t \geq t_{io}$. At $t = t_{io}$, $V(t) = V_o/2$.

The effective surface charge density is proportional to the surface potential. So the ratio $V(t)/V_o$ is equal to the relative charge density $\sigma(t)/\sigma(0)$. The decay of real charges in the electrets can be caused by the internal mechanisms and the external environment. The previous analysis is mainly for internal cause. External cause can be associated with many factors that tend to accelerate the decay process, particularly for electrets with one unshielded bare surface. For example, when approaching the bare surface, ions and moisture in air will react with charges and molecules at and just below the surface, leading to the annihilation of the stored charges there. The decay due to internal cause depends on the structure of the material used for forming the electrets and the method of charge injection. Typical charge decay results for some polymer electrets are shown in Figure 5-23. The data are from Turnhout.³²

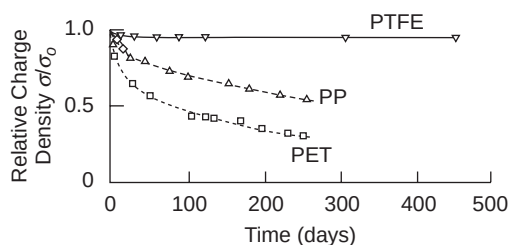


Figure 5-23 Some typical results of the decay of real charges in a dry atmosphere at room temperature for 50 μm polytetrafluorethylene (PTFE), 20 μm polypropylene (PP), and 25 μm polyethylene terephthalate (PET).

5.6.4 Distinction between Dipolar Charges and Real Charges

In nondipolar materials, only real space charges can be active because there are no permanent dipoles. But in dipolar materials, both dipolar and real charges may be present. Depending on the charging method and the conditions under which the charging process is performed, dipolar material can be charged mainly with real charges or mainly with dipolar charges. For example, if the material specimen is charged at low temperatures by electron-beam injection, then the charged specimen contains mainly real charges and practically no dipolar charges, even for dipolar materials, because the dipoles cannot reorient at low temperatures. On the other hand, if the material specimen with two intimate metallic contacts is charged by a thermo-electrical poling process at temperature $T > T_g$, then this charged specimen contains mainly dipolar charges because there is no injection of real charges into it.

Electrets containing mainly dipolar charges and those containing mainly real charges can be distinguished by some simple methods. The sectioning method described in Section 5.6.1 can be used for this purpose, because only the cut slices containing real charges show the induced charges in the Faraday pail. Another method is to superimpose a series of photocurrent pulses onto the isothermal discharging current under the short-circuit condition and possibly at a higher temperature to enhance the decay process.⁶² These photocurrent pulses can be produced inside the electret by illuminating the electret through a transparent electrode with a series of rectangular ultraviolet (UV) light pulses. If the decay is the real charge decay, the photogenerated carriers will interreact with the thermally released charges from the traps, and the photocurrent will change with time. If the decay is due to dipolar relaxation, the photocurrent will remain unchanged. An additional experiment measuring the dielectric constant and dielectric loss factor at low frequencies would also provide information about the contribution of dipolar charges on the TSDC peaks

when the results were compared with the TSDC thermogram.

Piezoelectric and pyroelectric effects are due mainly to dipolar polarization. When a charged electret with two deposited metallic electrodes is short-circuited, there is no external current flow. However, when this electret contracts due to an increase in the applied hydrostatic pressure or a decrease in temperature, there is a current flow, indicating that the electret contains mainly dipolar charges, because if the electret contained only space charges, there should be no response to the change of hydrostatic pressure or temperature (see Section 5.7).

If the electret contains both dipolar and real charges, then the TSDC thermogram may have peaks due to the release of trapped real charges and to depolarization. To identify these peaks, several methods can be used. One of them involves the measurements of both thermally stimulated conductivity (TSCo) and thermally stimulated polarization (TSP).⁶³ In the TSP process, the first step is to produce electron-hole pairs in the material by photoexcitation with UV light at a poling field and a low temperature. These electrons and holes will almost immediately be captured by traps. The second step is to remove the photoexcitation and also to raise the temperature linearly with time, so this step will produce only dipolar polarization. In this case, the TSCo measurement will reveal several peaks: One is the ρ peak due to the release of real charges from traps, and the other is due to dipolar depolarization. For information on other methods, see references.^{33,60}

5.7 Basic Effects of Electrets

The basic effects of electrets are based primarily on electrostatic induction. The external electric field from a charged electret with both of its surfaces bare, or with one surface bare and one surface metallized, will induce charges on a conductor or polarization in a dielectric material placed nearby. The configuration of the electrets with one surface metallized and the other surface bare is the one most commonly

used for practical applications, because it acts as a unipolarly charged unit and therefore creates an external electric field. The other configuration of the electrets, with both surfaces metallized, is usually used for electrets made of dipolar materials, so the stored charges are mainly dipolar polarization charges. Their effects are still based on electrostatic induction on the metallic electrodes. Applications using this configuration are based mainly on their piezoelectric and pyroelectric effects.

We analyzed the case of electrets with one surface metallized and the other surface bare in Section 5.3. The electret may consist of both dipolar and real charges, and the real charges may be distributed on the bare surface and in the bulk. But we can group all the dipolar and real charges into an equivalent total surface charge density σ_T as given by Equation 5-22. Suppose that such an electret configuration is situated with a distance to grounded surroundings much larger than the thickness of the electret d , as shown in Figure 5-24. The electric field inside the electret, F_e , and that outside the bare surface, F_o , are given by

$$\begin{aligned} F_e &= \frac{\sigma_T}{\epsilon_r \epsilon_o} \\ F_o &= 0 \end{aligned} \quad (5-94)$$

However, when an upper counter electrode is put in parallel with the electret at a distance s from the bare surface and connected through a load resistance R to the lower electrode and ground, then the induced charges on both electrodes σ_{up} and σ_{lo} will be shared in such a way that the potential across the electret, $V_e = F_e d$, is equal to that across the air gap, $V_o = F_o s$. Thus, in this case, the electric fields inside the electret and in the air gap are given by

$$\begin{aligned} F_e &= \frac{\sigma_T s}{(\epsilon_r S + d) \epsilon_o} \\ F_o &= \frac{\sigma_T d}{(\epsilon_r S + d) \epsilon_o} \end{aligned} \quad (5-95)$$

Obviously, both F_e and F_o are functions of s . In general, the electret thickness d is unchanged, but s can easily be changed. Once s is changed, V_e will not be equal to V_o , and there will be a

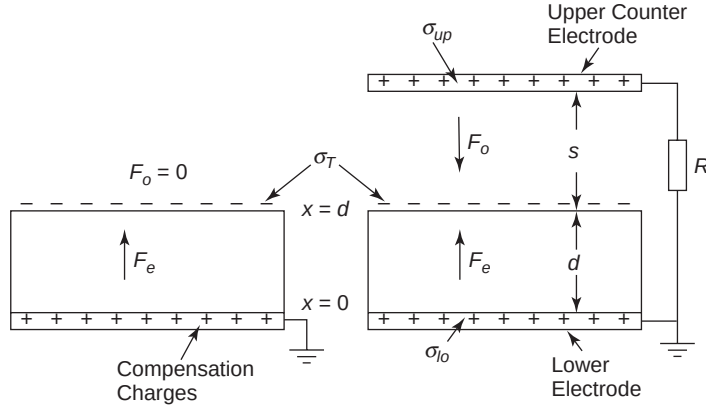


Figure 5-24 Charged electret with an equivalent total surface charge density σ_T on the bare surface: (a) with a lower electrode but without an upper counter electrode, and (b) with a lower electrode and an upper counter electrode.

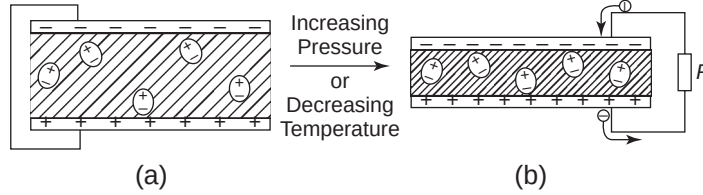


Figure 5-25 Electret with dipolar polarization charges exhibiting the flow of charges resulting from a change in thickness of the electret material caused by an increase in pressure or a decrease in temperature: (a) before the change of pressure or temperature and (b) during the change of pressure or temperature.

net potential across the load resistance R and hence a current i flowing from one electrode to the other, tending to maintain the balance in induced charges between the two electrodes. Thus, we can write

$$|V_o - V_e| = iR \quad (5-96)$$

and

$$i = \frac{d\sigma_{up}}{dt} = -\frac{d\sigma_{lo}}{dt} \quad (5-97)$$

Most practical applications are based on the effect of the field in the air gap F_o . Thus, by making either the electret or the counter electrode thin and flexible enough to function as a diaphragm or a membrane, there is no current flow when the diaphragm or membrane is at the normal resting condition. But there will be a current flow and hence a potential across R when the diaphragm or membrane is pressed by

hand or vibrated oscillatively by a compressed sound wave. Many applications, such as electret microphones, on and off keys on an electronic keyboard, etc., are based on this simple principle.

In general, electrets with both surfaces metallized are those containing mainly dipolar polarization charges. When the electrodes are electrically connected through a load resistance R , as shown in Figure 5-25, there is no current flow through the load since the induced charges on the electrodes compensate for the polarization charges. But when the electret is subjected to an increase in pressure or to a decrease in temperature, the thickness of the electret will decrease, resulting in the flow of a transient current due to the piezoelectric or pyroelectric effects in the material. The trend is reversible; when the electret is subjected to a decrease in pressure or an increase in temperature, the

thickness of the electret increases, giving rise to a current flow in the opposite direction. However, neither nondipolar electrets containing real charges only nor the real charges, if any, present in the dipolar electrets have piezoelectric or pyroelectric effects.

In general, when the dipolar electret is subjected to an increase in pressure or a decrease in temperature, the alignment of the dipoles to the original poling direction increases, increasing the induced charges on the electrodes. This is why, for the case shown in Figure 5-25(b), the electrons are flowing from the bottom electrode to the upper electrode. This effect can be considered due to thermal agitation. Although the so-called frozen-in dipoles have their mean dipole moment in the original poling direction,

there is always thermal motion, whose magnitude increases with increasing temperature, tending to alter the original alignment. Thus, an increase in temperature would cause a decrease in the mean moment of the dipoles in the original poling direction, and vice versa.

A similar argument can be used to explain the piezoelectric effect. As the applied pressure increases, the density of the material increases and the thermal motion of the dipoles becomes more sluggish, leading to an increase in the mean moment of the dipoles in the original poling direction, as illustrated in Figure 5-26. The configuration of dipolar electrets with both surfaces metallized is widely used in electro-mechanical conversion devices and various sensors.⁶⁴

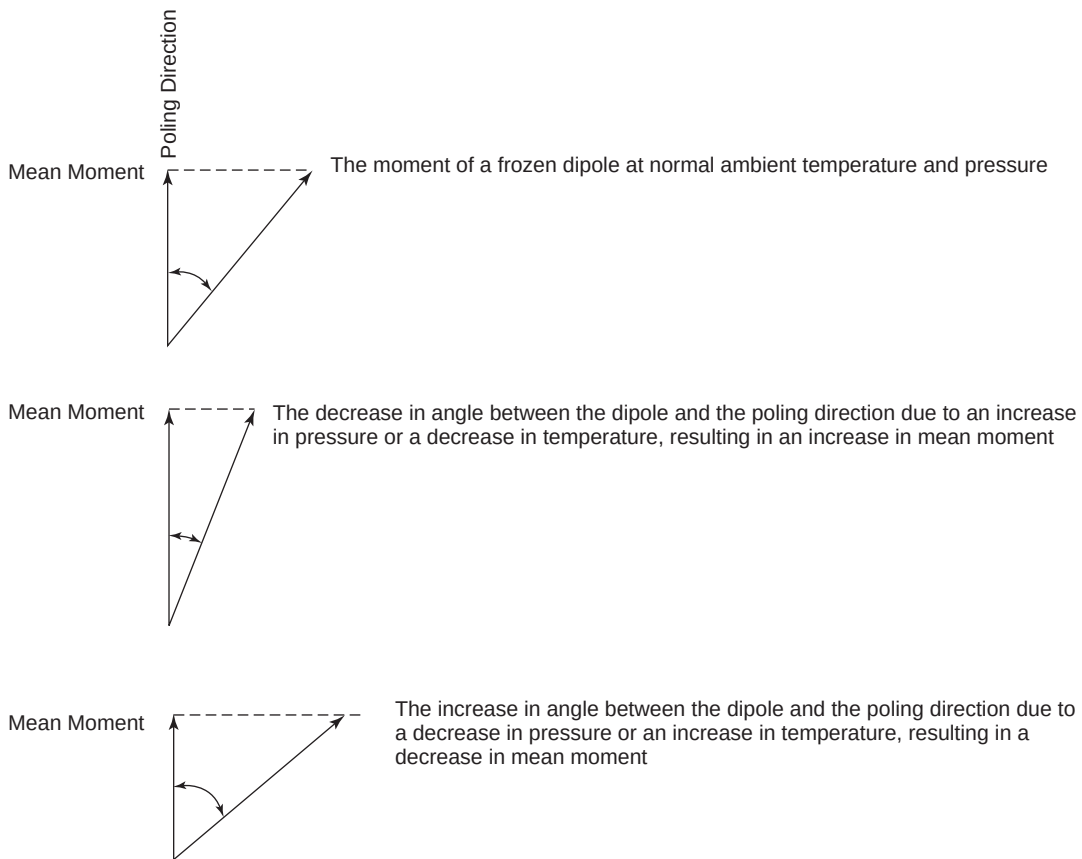


Figure 5-26 The change of the mean moment of a librating dipole due to an increase in pressure or a decrease in temperature, or vice versa.

5.8 Materials for Electrets

In this section we shall describe some properties of the materials commonly used for electrets.

5.8.1 General Remarks

To select suitable materials for forming electrets for certain applications, the following items must be considered:

- The type of charges: dipolar or real charges
- The electric field, temperature, and time required for the poling processes
- The thermal stability of the charged electrets attained
- Plate, sheet, or flexible thin-film form
- The operating conditions of the electrets produced

Materials for electrets can be classified into two major categories: inorganic materials and organic polymers. Inorganic materials have a high melting point and are hard, sometimes brittle, particularly in thin-film form. Organic polymers have a low melting point and are soft and flexible; they are comparably easy to fabricate into flexible thin films. Inorganic materials, such as titanate ceramics (BaTiO_3 , PZT, etc.) and metallic oxides (Al_2O_3 , SiO_2 , etc.) can be used for electrets. However, the suitability of these materials is still under investigation. But organic polymers, such as wax-based materials, PTFE, PTFE-FEP, PVDF, and PET have been widely used to form electrets for various applications.

On the basis of their composition, polymers can be classified into two major categories: fluorocarbon polymers and nonfluorinated polymers. Such commonly used polymers as PTFE, PTFE-FEP, and PVDF belong to the fluorocarbon family. Polymers such as polyethylene (PE), polypropylene (PP), PMMA, and PET, belong to the nonfluorinated family. In general, fluorocarbon polymers have a very stable structure and hence contain deep carrier traps, implying that they can hold trapped charges for a long period of time. For example, PTFE-FEP is a nondipo-

lar material and slightly p-type, so this material is most suitable for negative real-charge storage, while PVDF is a dipolar and piezoelectric material and has a very stable structure, so it is more suitable for dipolar charge storage. The nonfluorinated polymers are generally n-type, so they are more suitable for positive real charge storage. These materials are not as good as the fluorinated materials in terms of persistency and ability to hold the trapped charges for a long period of time. However, nonfluorinated polymers can be improved through the incorporation of suitable impurities.

Polymers can also be classified into two categories: nondipolar and dipolar. For example, PTFE, PTFE-FEP, and PE are nondipolar, while PVDF, and PMMA are dipolar. At normal temperature and pressure, the so-called nondipolar and dipolar materials do not exhibit dipolar behavior, simply because the dipoles cannot orient in solid state at low temperatures. Dipolar properties appear only after poling at $T \geq T_g$. In general, nondipolar or weak dipolar materials are used to form electrets with mainly real charge storage, and dipolar materials for forming electrets with mainly dipolar polarization charge storage.

5.8.2 Physical Properties

Polymers are the most commonly used materials for electrets because they can be produced easily in flexible thin-film form and have good dielectric properties. Polymers are organic compounds formed by large molecules called *macromolecules*. A polymer consists of many macromolecules, each of which is formed by a large number of smaller structural units called *monomers*, usually of the order (10^3 – 10^5). The monomer is a building block, and it possesses two or more bonding sites so that it can bond with other monomers to form a link in the long polymer chain. Macromolecules are in the form of random coils, but under the influence of external forces, such as shearing or stretching, the polymer chains can undergo crystallization at a temperature above the glass transition temperature T_g . Crystallization results in rigidity, strength, thermal stability, and resistance

against dissociation, while polymers in the amorphous phase possess softness, elasticity, and solubility.

By controlling the degree of crystallinity, we can tailor to a certain extent the physical properties of polymers to suit special applications. Also, a polymer of a given chemical composition can be manufactured into products with widely different properties by controlling the degree of crosslinking, which is commonly achieved by chemical means to make the atoms of the backbone in the two crossing chains joined by primary chemical bonds, which are mostly covalent in nature. Crosslinking may also be formed by physical bonds, such as hydrogen bonds, but these bonds are much weaker than the chemical ones. A low degree of crosslinking may impart elastic memory to the polymer when it is being subjected to mechanical deformation, but it recovers its original form after the removal of the stress. The higher the degree of crosslinking, the harder the polymer and the less it is affected by solvents. However, when the degree of crosslinking becomes too high, the polymer may completely lose its elastic properties and

become hard, even brittle, and also resistant to softening by heating.

The macromolecules in polymers are kept together mainly by van der Waals or London forces, or hydrogen bonds. Although these interaction forces are much weaker than those responsible for binding the monomers to form the macromolecules, they play a decisive role in determining the physical and chemical properties of the polymers.

A polymer melts when the temperature reaches its melting point T_m , that is, when the vibrational energy of the macromolecules exceeds the bonding energy between them. Conversely, when a molten material is cooled, it becomes solidified, implying that the attractive force between the molecules overcomes the kinetic energy of their thermal motion. In general, the melt of a crystal is always accompanied by a change in phase; for amorphous materials, such a change in phase does not occur. However, there are three states in which an amorphous polymer may be set, depending on the temperature. These three states are the viscofluid state, the rubbery state, and the glassy state, as shown in Figure 5-27. The

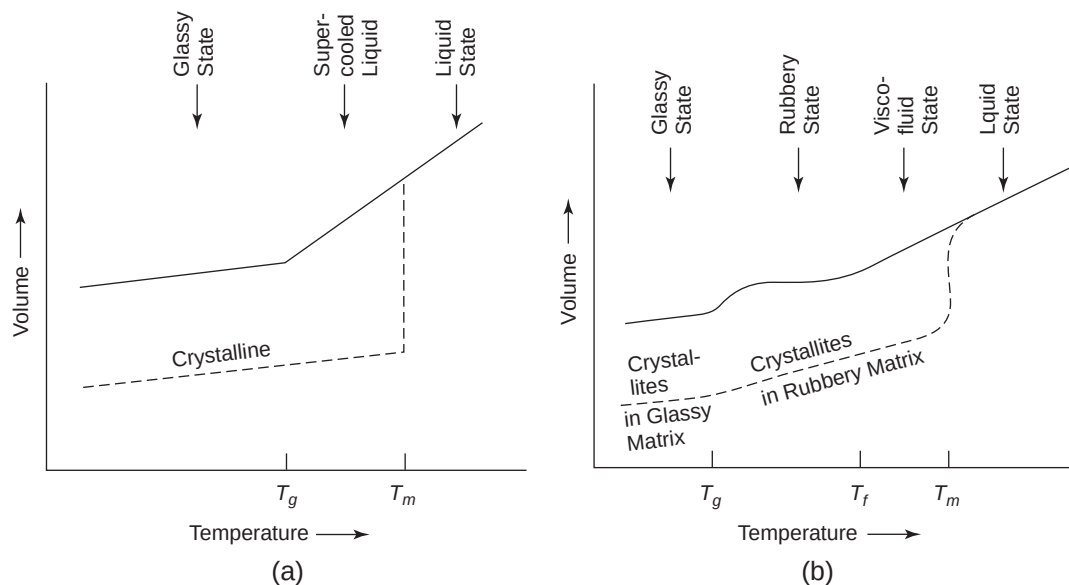


Figure 5-27 The volume-temperature characteristics in different states for (a) a typical low molecular-weight polymer and (b) a typical high molecular-weight polymer.

change from one state to another is not accompanied by a change in phase; the change is due mainly to a relaxation process. Each relaxation process requires a certain time to reach its thermodynamic equilibrium state, which is generally referred to as the *relaxation time* of the process. As the temperature is lowered, the relaxation time corresponding to a particular process will increase. All transition temperatures, T_g , T_f , and T_m , depend on molecule weight, chain branching, and the presence of impurities or additives in the polymer.

The properties and the structures of some commonly used electret materials, mainly polymers, are given in Tables 5-1 and 5-2. Following are brief descriptions of the important

properties of some polymers most commonly used for electrets.

In general, for nondipolar materials we would like the conductivity and the carrier mobilities as small as possible, while for dipolar materials we would also like low conductivity and carrier mobilities, as well as a large piezoelectric constant and a large pyroelectric coefficient, in order to attain good electret properties. PTFE, with the trade name Teflon, is one of the most important nonpolar organic polymers, which has a high melting point ($T_m = 327^\circ\text{C}$) and a high heat deflection temperature ($T_d = 121^\circ\text{C}$). PTFE is a tough, flexible polymer that retains its ductility at extremely low temperatures (-269°C). This

Table 5-1 Some properties of some commonly used electret materials at room temperature. ϵ_r = dielectric constant, g = electrical conductivity, T_m = melting point. T_g = glass transition temperature, d_{31} = piezoelectric constant, p = pyroelectric coefficient, and ρ = density.

Material Name	Abbreviation	Class and Structure	ϵ_r (60 Hz)	g (Ωcm) ⁻¹	T_m ($^\circ\text{C}$)	T_g ($^\circ\text{C}$)	ρ g cm ⁻³	d_{31} (cm V ⁻¹)	p (C cm ⁻² K ⁻¹)
Silicon Dioxide	Quartz (SiO ₂)	Piezoelectric Crystalline	3.78	$<10^{-14}$	1600	—	2.2	2×10^{-10}	—
Polytetrafluoroethylene	PTFE (Teflon)	Nonpolar Crystalline & Amorphous Regions	2.20	$<10^{-18}$	327	127	2.15–2.20	—	—
Tetrafluoroethylene–Hexafluoropropylene Copolymer	PTFE-FEP	Nonpolar Crystalline & Amorphous Regions	2.20	$<10^{-18}$	270–280	130	2.14–2.17	—	—
Tetrafluoroethylene–Perfluoromethoxyethylene Copolymer	PTFE-PFA	Nonpolar Crystalline & Amorphous Regions	2.06	$<10^{-16}$	302–310	150	2.15–2.16	—	—
Polyethylene Terephthalate	PET (Mylar)	Crystalline & Amorphous Regions	3.30	$<10^{-18}$	250–265	88	1.37–1.39	—	—
Polyvinylidene Fluoride	PVDF (Kynar)	Piezoelectric Crystalline & Amorphous Regions	13.00	$<10^{-14}$	170	–40	1.78	2×10^{-9}	4×10^{-9}
Triglycine Sulfate	TGS	Crystalline	43.00	$<10^{-15}$	—	—	1.69	5×10^{-9}	3×10^{-8}
Polyvinyl Chloride	PVC	Crystalline & Amorphous Regions	10.00	$<10^{-16}$	220	80	1.40	7×10^{-11}	1×10^{-10}

Table 5-2 Chemical structures of some commonly used polymers.

Polymer	Abbreviation	Structure	W	X	Y	Z
Polyethylene	PE		H	H		
Polypropylene	PP		H	CH ₃		
Polyvinyl Chloride	PVC		H	Cl		
Polyvinylidene Chloride	PVDC		Cl	Cl		
Polyvinyl Fluoride	PVF (Tedlar)		H	F		
Polyvinylidene Fluoride	PVDF (Kynar)		F	F		
Polystyrene	PS		C ₆ H ₅	H		
Polyethylmethacrylate	PEMA		CH ₃	COOC ₂ H ₅		
Polymethylmethacrylate	PMMA		CH ₃	COOCH ₃		
Polytetrafluoroethylene	PTFE (Teflon)		F	F	F	F
Polytetrafluoroethylene–hexafluoropropylene Copolymer	PTFE-FEP		F	CF ₃	F	F
Polytetrafluoroethylene–perfluoromethoxyethylene Copolymer	PTFE-PFA		F	O—CF ₃	F	F
Polyethylene Terephthalate	PET (Mylar)					

may be due to its strong C—F bond, which is about 393 kJ/mol, much higher than the C—C, C—Cl, and C—H bonds of 335, 276, and 261 kJ/mol, respectively. In PTFE, the effect of *F* is even stronger than that of *C*. Compared to other polymers, PTFE has a longer C—C chain, resulting in a high molecular weight of the order of 10⁵–10⁷. PTFE has compression and tensile strength of the order of 25 MPa, elongation of 200%, and low water absorption. After a PTFE specimen has been immersed in water at 23°C for 24 hours, the change of its density is less than 10^{−4}. PTFE retains its superior chemical resistance against corrosives and solvents, and its good strength, toughness, and wear resistance over a wide range of temperatures (from 260°C to −250°C). PTFE films can easily be bonded by adhesives to surfaces of other materials.⁶⁵

The electrical properties of PTFE-FEP are very similar to those of PTFE, but it has better mechanical properties. Its elongation is 350%. For example, the electret films made of the copolymer with 11% hexafluoropropylene and 89% tetrafluoroethylene can store a high density of negative charges for a long period of time because of its high concentration of deep electron traps. PTFE-FEP copolymer films are now widely used to form electrets for various applications because of their superior properties.⁶⁶

PVDF, with the trade name Kynar, is a polymer with a high degree of crystallinity, a glass transition temperature *T_g* = −40°C, a heat deflection temperature *T_d* = 80°C, and a melting point between 158°C and 197°C, depending on the degree of crystallinity. PVDF's piezoelectric constant *d*₃₁ and pyroelectric coefficient *p*

are closely related to the poling-induced polarization P_s . In general, d_{31} , p , and P_s are mutually proportional and increase with increasing poling electric field F_p , poling temperature T_p , and poling time t_p .^{67,68} The dependence of F_p and t_p tends to approach a saturation at high F_p and t_p , but this trend depends on T_p . The dipolar relaxation time is about 10^9 seconds. PVDF has also been found to be ferroelectric.⁶⁹ Its piezoelectric and pyroelectric effects are reversible. Electrets formed by PVDF films with both surfaces metallized poled with an electric field of $500\text{--}800\text{ kV cm}^{-1}$ at $90\text{--}110^\circ\text{C}$ for one hour following the thermo-electrical process have been widely used for various applications, because PVDF has a relatively high piezoelectric constant for a wide range of frequencies; high resistance to most acids, alkalis, salts, and solvents; and good mechanical strength.

The method of combining two or more homopolymers to form a copolymer has been used to produce a material with properties suitable for certain operating conditions. For example, by adjusting the composition of each of the homopolymers trifluoroethylene (TrFe), tetrafluoroethylene (TFE), and vinylidene fluoride (VDF), it is possible to produce a copolymer with different electret properties. This method has generated great interest among researchers and technologists.

In general, the properties of electret materials can be improved or modified by either chemical or physical means. A polymer may be very good in certain properties but very weak in others. This problem may be solved by mixing two polymers: one good in certain properties and the other good in other properties. For example, PVDF has strong piezoelectric properties, but its dielectric constant is relatively small. If the dielectric behavior of the amorphous region of the PVDF can be improved, the problem can be solved. To do this, we first must find a suitable polymer that has good dielectric properties and is also dissolvable in PVDF, so that the two polymers can be mixed uniformly to form a good copolymer. In this case, PEMA or PMMA would be a good candidate. For PVDF mixed with PMMA, the latter plays the important role of improving the

dielectric behavior of the amorphous region, leaving PVDF to remain as the crystalline region responsible for the piezoelectric behavior.

For electrets with one surface or both surfaces bare, the surface is easily contaminated by the surrounding medium. If the bare surface is treated with a chemical solution (e.g., sodium naphthalene), both the storage of charges in the electret and its stability will be improved significantly.⁷⁰ In general, the absorption of smaller molecules, such as water, depends on the structure of the polymers. Nonpolar polymers, such as PE, PP, PS, and PTFE, absorb nonpolar gases and liquids but are largely immune to water and ethanol. On the other hand, polar polymers, such as biphenol polycarbonate (PC) and polyurethane (PU), readily absorb water. However, the bare surfaces of the electrets exposed to air can easily be contaminated by various impurities, which may be hydrophilic, tending to absorb moisture from the surrounding medium. This may cause an increase in the surface conductivity and hence a fast rate of the decay of stored charges. To improve this situation, a suitable chemical treatment on the surface, such as a thin coating of silicone, may be used to convert the hydrophilic behavior to hydrophobic behavior. For example, chemically treating the surface of SiO_2 with hexamethyl disilazane (HMDS) significantly improves the stability of the SiO_2 electret formed. In short, the mechanical properties, charge storage ability, and thermal stability can be modified or improved to suit any particular application by chemical means, or by physical means through mechanical, heat, light, or radiation treatments.⁶⁶ New materials and new treatments are continuously being developed. For more information, see references 71–74, just a few of many publications.

Composite materials made from a filler—particles, flakes or fibers—embedded in a matrix formed by polymers or glasses, taking advantage of particular properties of each component, may be used to form electrets. A *composite* means that the material consists of two or more distinct phases. For example, when a piezoelectric ceramic and a polymer are com-

bined to form ferroelectric composite electrets, these component materials allow greater flexibility to tailor the properties, whether electrical, thermal, mechanical, or other, to suit a particular application.^{73,75–78}

5.9 Applications of Electrets

The most common applications of electrets are transducers and sensors. Because any external force resulting from mechanical compression, heat, sound waves, electricity, light, or radiation would interact with the stored charges in the electrets, a wide variety of applications, including novel devices in the medical and biological areas, have been put forward. This section briefly describes the basic principles of some typical applications. From these few examples, it can be imagined that all the different applications of electrets are based mainly on the simple, basic effects discussed in Section 5.7.

5.9.1 Electret Microphones

The earliest electret transducers were electret microphones, first developed in Japan in 1928.^{79,80} In the early days, electrets were made of wax-based materials, which were poor in thermal stability. Since 1962, however, many other materials have been developed for forming electrets with good thermal stability, such as PTFE, PTFE-FEP, and PVDF. The

materials with better electrical and mechanical properties are those containing halocarbons, such as PTFE-FEP, and chlorotetrafluoroethylene (CTFE). These materials have been widely used to form electrets because they can be made into very thin, flexible polymer films.

The basic principle of electret microphones is shown in Figure 5-28. The system consists of a negatively charged electret: One surface is coated with metallic film and the other, bare surface, is a diaphragm. A metal plate is separated from the diaphragm by an air gap. Without a compressed sound wave, the voltage across the electret V_e and that across the air gap V_o are equal, and they are

$$\begin{aligned} |V_e| &= F_e d \\ &= |V_o| = F_o s_o = \frac{\sigma_T ds_o}{(\epsilon_r s_o + d)\epsilon_o} \end{aligned} \quad (5-98)$$

where σ_T is the total surface charge density (see Section 5.3); d is the thickness of the electret, which can be assumed to be unchanged under any external perturbation; and s_o is the air gap spacing. If the diaphragm is made to oscillate due to a compressed sound wave, then the air gap spacing will vary according to

$$s = s_o + a \sin \omega t \quad (5-99)$$

where s_o is the original gap spacing, ω is the frequency of the oscillation, and a is the amplitude of the variation. The induced charge on the upper electrode, from Equation 5-18, is given by

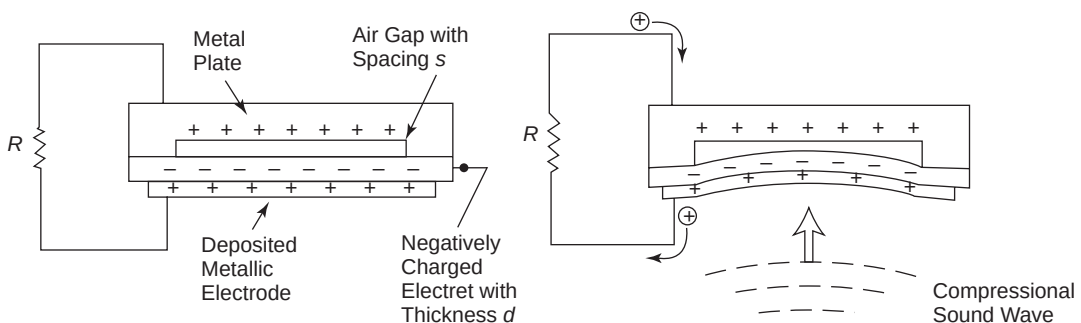


Figure 5-28 Schematic diagrams illustrating the basic principle of an electret microphone. A current flow occurs only when there is a compressional sound wave.

$$\sigma_{up} = \frac{d\sigma_T}{\epsilon_r(s_o + a \sin \omega t) + d} \quad (5-100)$$

Since σ_{up} varies with time, there will be a current i flowing through the load resistance R , which is

$$i = A \frac{d\sigma_{up}}{dt} = \frac{Ada\epsilon_r\omega\sigma_T}{[\epsilon_r(s_o + a \sin \omega t) + d]^2} \cos \omega t \quad (5-101)$$

where A is the area of the electret. Since $a \ll s_o$, the potential across R can be written as

$$V_R = iR = \frac{Ada\epsilon_r\omega\sigma_T R}{(\epsilon_r s_o + d)^2} \cos \omega t \quad (5-102)$$

This voltage may be amplified in the conventional way, so the microphone can be made to work without any external voltage supply. Typical electret microphones are those made of PTFE-FEP films of 500–1000 Å in thickness with a stored charge density of the order of 10–20 nC cm⁻² and an air gap of about 10–20 μm.

5.9.2 Electromechanical Transducers

In fact, electro-acoustic transducers, such as microphones and earphones, are indirectly electromechanical transducers because the mechanical forces, in this case, come from sound waves. Thus, the same principle is applied to the applications in phonograph cartridges, touch or key switches, impact detectors, etc.

Since the discovery of strong piezoelectric and pyroelectric effects in PVDF, this material has been widely used for electrets in electro-mechanical transducers, particularly in high frequency loudspeakers and headphones, hydrophones, medical microphones for fetal phonocardiography, and heart-rate and blood-pressure monitors.^{68,81,82} In comparison with conventional piezoelectric and pyroelectric crystals, polymers offer the following advantages:

- They are flexible and tough.
- They can be made very thin with a large area.
- They have a low mechanical impedance and hence exhibit good acoustic coupling to water and biological systems.

However, their high mechanical and dielectric losses sometimes limit their applications under certain situations.

5.9.3 Pyroelectric Detectors

Several polymers, such as PVDF and PVF, possess pyroelectric properties.^{64,83} The pyroelectric coefficient of PVDF is not high compared to many inorganic materials, such as BaTiO₃ and TGS, as shown in Table 4.5 (Chapter 4). This coefficient is, however, relatively high compared to other polymers. In addition, PVDF has excellent mechanical properties and low thermal conductivity, and it can be made into flexible thin films. Thus, PVDF has been widely used for pyroelectric detectors, such as the one shown in Figure 5-29.

The PVDF electret is a very thin sheet of film that heats up quickly. The metallic face electrode deposited on the upper surface must be good in absorption of radiation energy so that heat radiation, such as infrared radiation striking it will be quickly converted to heat and transmitted to the PVDF electret, causing a change in the induced charges on the upper and bottom electrodes and giving rise to a current flow i and a signal voltage $V_R = iR$ across the load resistance R , based on the principle given in Section 5.7. The bottom electrode is usually a metal block bonded by adhesive to the bottom surface of the electret, serving both as an electrode and as a good heat sink to avoid the build-up of the heat in the electret.

The detector does not respond to continuous radiation. It responds to a change in received

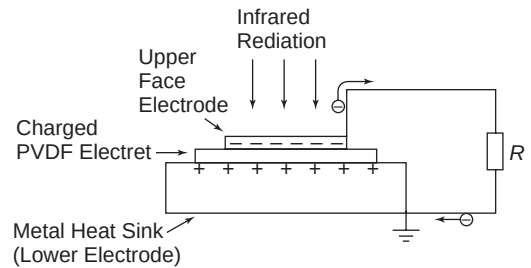


Figure 5-29 Schematic diagram showing the basic principle of a PVDF pyroelectric detector.

radiation by giving a voltage blip when the radiation changes. Obviously, the blip signal will be contaminated by noise, so the blip should be as large and as long-lasting as possible. However, this kind of simple detector has been used in fire detection, burglar alarm systems, vidicon, etc. Pyroelectric electrets have also been used in photopyroelectric spectroscopy (PPES) for photothermal measurements of thermal properties of solids, such as thermal diffusivity and conductivity, or optical properties such as absorption, transmission, and nonradiative de-excitation efficiency.^{84–87} Pyroelectric electrets have also been used as pyroelectric calorimeters for measurements of the specific heats of thin films,⁸⁸ and as a titration for characterizing ligand-macromolecules and macromolecule–macromolecule reactions.⁸⁹

5.9.4 Other Applications

There are a number of other applications for electrets. In this section, only a few are given as examples. Electrets have been used as air filters. For this use, the electret filters are usually made of polypropylene film, which is first stretched to a thickness of about 8–10 μm , then charged by a corona discharge technique, and finally fibrillated into tiny fibers. These fibers are then made to spread into a broad web (due to their repulsive forces) to form an air filter.⁹⁴ The fibers capture particles from the air due to the induction of the dipoles in them.

Electrets can also be used as radiation dosimeters. The basic principle of radiation dosimetry is based on the interaction between radiation and the stored charges in the electret.⁹⁵ A systematic study of such interactions⁹⁶ and a detailed description of various electret dosimeters^{97,98} have been reported. In general, radiation effects, such as the decay of electret charges due to radiation or the generation of radiation-induced conductivity (RIC) in the electret, are employed for the measurement of radiation doses.⁹⁹

Real charge or polarization charge storage has been found in many biomaterials. This implies that these materials can behave as electrets capable of performing certain functions

for biomedical applications, as well as for fundamental studies of biophysical phenomena. Electret effects have also been found in some biopolymers, such as protein.^{74,100,101} In fact, Carnauba wax, the material originally used by Eguchi to verify the electret effects,^{2,3} is a material of biological origin. The methods for producing electret effects in biomaterials are similar to those given in Section 5.2. However, in using such electret effects, we must understand the mechanisms for the storage of real charges or polarization charges and their importance to biophysical phenomena and biomedical applications. This area is still at an early stage, but because of its potential for biomedical applications, new research and development results will likely emerge rapidly in the future.

Regarding the applications of polymer electrets in the biological and biomedical areas, we mention here only a few selected topics to demonstrate the potential of electret applications in these areas. Teflon electrets placed in contact with bones would cause accelerated growth of callus, necessary to heal fractures, which may be associated with negative electret charges.¹⁰² Electret bandages placed on skin incisions would considerably improve the tensile strength of the wound over a period of time and thus accelerate the healing process.¹⁰³ It is important to find blood-compatible biomaterials for use inside a human body (implants), which do not cause blood coagulation (thrombus formation). It has been found that blood platelets are electrically charged and that a negatively charged surface inhibits blood coagulation.¹⁰⁰ Teflon has a good biological compatibility, so negatively charged Teflon electrets may be used for human heart or arterial surgery.^{102–104}

References

1. O. Heaviside, *Electrical Papers*, (Chelsea, New York, 1892), p. 488.
2. M. Eguchi, *Proc. Phys. Math. Soc. Japan*, **1**, 326 (1919), **2**, 169 (1920), and **3**, 107 (1923).
3. M. Eguchi, *Phil. Mag.*, **49**, 178 (1925).

4. G. Nadjakov, C.R. Acad. Sci., 204, 1865 (1937).
5. G. Nadjakov, Phys. Z., 39, 226 (1938).
6. J. H. Dressauer and H. F. Clark (Ed.), *Xerography and Related Processes*, (Focal Press, London, 1965).
7. H. P. Kallmann and B. Rosenberg, Phys. Rev., 97, 1865 (1955).
8. J. R. Freemann, H. P. Kallmann, and M. Silver, Rev. Modern Phys., 33, 553 (1961).
9. V. M. Fridkin and I.S. Zheludev, *Photoelectrets and the Electrophotographic Processes*, (Consultants Bureau, New York, 1961).
10. I. Bunget and M. Popescu, *Physics of Solid Dielectrics*, (Elsevier, Amsterdam, 1984).
11. P. Eyerer, J. Appl. Polymer Sci., 15, 12 (1971).
12. H. Kawai, Jpn. J. Appl. Phys., 8, 975 (1969).
13. B. Gross, J. Chem. Phys., 17, 866 (1949).
14. B. Gross and L. F. Dennard, Phys. Rev., 67, 253 (1945).
15. M. M. Perlman and J. L. Meunier, J. Appl. Phys., 36, 420 (1965).
16. R. W. Chudleigh, J. Appl. Phys., 47, 4475 (1976).
17. R. A. Moreno and B. Gross, J. Appl. Phys., 47, 3397 (1974).
18. B. Gross, J. Polym. Sci., 27, 135 (1958).
19. T. Matsukawa, R. Shimizu, K. Karada, and T. Kato, J. Appl. Phys., 45, 733 (1974).
20. G. M. Sessler, in *Electrical Properties of Polymers*, edited by D. A. Seanor, (Academic Press, New York, 1932), p. 245.
21. B. Gross, G. M. Sessler, and J. E. West, J. Appl. Phys., 45, 2841 (1974).
22. G. M. Sessler and J. E. West, J. Electrostatics, 1, 111 (1975).
23. E. Kuffel and M. Abdullah, *High Voltage Engineering*, (Pergamon, Oxford, 1970).
24. C. W. Reedyk and M. M. Perlman, J. Electrochem. Soc., 115, 49 (1968).
25. G. M. Sessler and J. E. West, J. Electrochem. Soc., 115, 836 (1968).
26. G. M. Sessler and J. E. West, Rev. Sci. Instr., 42, 15 (1971).
27. R. E. Collins, Rev. Sci. Instr., 48, 83 (1977).
28. R. E. Collins, J. Appl. Phys., 47, 4804 (1976).
29. R. E. Collins, Ferroelectrics, 33, 65 (1981).
30. R. E. Collins, J. Appl. Phys., 51, 2973 (1980).
31. H. Frohlich, *Theory of Dielectrics*, (Clarendon, Oxford, 1958) p. 80.
32. J. Van Turnhout, *Thermally Stimulated Discharge of Polymer Electrets*, (Elsevier, Amsterdam 1975).
33. J. Van Turnhout, *Thermally Stimulated Discharge of Electrets*, in "Electrets," 2nd edition, edited by G. M. Sessler (Springer-Verlag, New York, 1987), p. 81.
34. M. Abrowitz and J. A. Stegun, *Handbook of Mathematical Functions*, (Dover, New York, 1965).
35. J. Van Turnhout, Polymer J., 2, 173 (1971).
36. T. A. T. Cowell and J. Woods, British J. Appl. Phys., 18, 1045 (1967).
37. S. H. Carr, in *Electrical Properties of Polymers*, edited by D. A. Seanor, (Academic Press, New York, 1982) p. 215.
38. G. R. Davies, in *Physics of Dielectric Solids*, Institute of Physics Conference Series No. 58 (Institute of Physics, Bristol and London, 1980), p. 50.
39. C. Lacabanne and D. Chatain, Macromol. Chem., 179, 2765 (1978).
40. V. K. Jain, C. L. Gupta, R. K. Jain, and R. C. Tyagi, Thin Solid Films, 48, 1975 (1978).
41. J. Vanderschueren and A. Linkens, J. Polym. Sci. Phys., 16, 223 (1978).
42. N. F. Mott and E. A. Davis, *Electronic Processes in Non-Crystalline Materials*, (Clarendon Press, Oxford, 1971).
43. B. Gross, in *Static Electrification*, (Institute of Physics, London, 1971), p. 33.
44. D. K. Walker and O. Jefimenko, in *Electrets, Charge Storage and Transport in Dielectrics*, edited by M. M. Perlman, (Electrochemical Society, Princeton, 1971), p. 455.
45. G. M. Sessler, J. E. West, D. A. Berkley, and G. Morgenstern, Phys. Rev. Lett., 38, 368 (1977).
46. D. W. Tong, in *Conference Record of the 1980 IEEE International Symposium on Electrical Insulation*, (IEEE Service Center, Piscataway, N.J., 1980), p. 179.
47. T. Maeno, T. Fusibe, T. Takada, and C. M. Cooke, IEEE Trans. Electr. Insul., EI-23, 433 (1988).
48. Y. Li, M. Yasuda, and T. Takada, IEEE Trans. Diel. & Elect. Insul., DEI-1, 188 (1994).
49. T. Takada, IEEE Trans. Diel. & Elect. Insul., DEI-6, 519 (1999).
50. Y. Li and T. Takada, IEEE Electrical Insulation Magazine, 10, 704 (1994).
51. P. Laurence, G. Dreyfus, and J. Lewiner, Phys. Rev. Lett., 38, 46 (1977).

52. G. M. Sessler, J. E. West, R. Gerhard-Multhaupt, and H. von Seggern, *IEEE Trans. Nucl. Sci.*, **NS-29**, 1644 (1982).
53. K. Fukunaga, *IEEE Electrical Insulation Magazine*, **15**, 6 (1999).
54. A. G. Milnes, *Deep Impurities in Semiconductors*, (Wiley, New York, 1973).
55. R. A. Creswell and M. M. Perlman, *J. Appl. Phys.*, **41**, 2365 (1970).
56. M. M. Perlman, *J. Electrochem. Soc.*, **119**, 892 (1972).
57. R. Chen, *J. Mat. Sci.*, **11**, 1521 (1976).
58. R. J. Fleming, *IEEE Trans. Elect. Insulation*, **EI-24**, 523 (1989).
59. K. K. Kanazawa, I. P. Batra, and H. J. Wintle, *J. Appl. Phys.*, **43**, 719 (1972).
60. G. M. Sessler, "Physical Principles of Electrets," in *Electrets*, 2nd edition, edited by G. M. Sessler, (Springer-Verlag, New York, 1980), p. 13.
61. A. Reiser, M. W. B. Lock, and J. Knight, *Trans. Faraday Soc.*, **65**, 2168 (1969).
62. K. Tahira and K. C. Kao, *J. Phys. D: Appl. Phys.*, **18**, 2247 (1985).
63. S. W. S. McKeever and D. M. Hughes, *J. Phys. D: Appl. Phys.*, **8**, 1520 (1975).
64. M. G. Broadhurst and G. T. Davis, "Piezoelectric and Pyroelectric Properties," in *Electrets*, 2nd edition, edited by G. M. Sessler (Springer-Verlag, New York, 1987) p. 285.
65. R. B. Seymour and C. E. Carraher, *Structure-Property Relationships in Polymers*, (Plenum Press, New York, 1984).
66. D. P. de Nemours, *Teflon FEP Fluorocarbon Films, Electrical Properties*, Technical Information Bull., **T-4E**, (1989).
67. N. Murayama and H. Hashizume, *J. Polym. Sci. Polym. Phys. Ed.*, **14**, 989 (1976).
68. Y. Wada, "Piezoelectricity and Pyroelectricity," in *Electronic Properties of Polymers*, edited by J. Mort and G. Pfister, (Wiley, New York, 1982) p. 109.
69. T. Furakawa, M. Date, and E. Fukada, *J. Appl. Phys.*, **51**, 1135 (1980).
70. S. Haridoss and M. M. Perlman, *J. Appl. Phys.*, **55**, 1332 (1984).
71. R. Kressmann, G. M. Sessler, and P. Gunther, *IEEE Trans. Diel. & Elect. Insul.*, **DEI-3**, 607 (1996).
72. G. Eberie, H. Schmidt, and W. Eisenmenger, *IEEE Trans. Diel. & Elect. Insul.*, **DEI-3**, 624 (1996).
73. C. J. Dias and D. K. Das-Gupta, *IEEE Trans. Diel. & Elect. Insul.*, **DEI-3**, 706 (1996).
74. S. B. Lang, *IEEE Trans. Diel. & Elect. Insul.*, **DEI-7**, 466 (2000).
75. Y. Ohara, M. Miyayama, K. Koumoto, and H. Yanagida, *Sensors and Actuators, A-36*, 121 (1993).
76. Y. Wang, W. Zhong, and P. Zhang, *J. Appl. Phys.*, **74**, 512 (1993).
77. H. Zewdie and F. Brouers, *J. Appl. Phys.*, **68**, 713 (1990).
78. H. L. W. Chan and L. L. Guy, "Piezoelectric Ceramic/polymer Composites for High Frequency Applications" in *Ferroelectric Polymers and Ceramic-Polymer Composites*, edited by D. K. Das Gupta, (Trans-Tech. Pub., Switzerland, 1994) p. 275.
79. G. M. Sessler and J. E. West, "Applications," in *Electrets* 2nd edition, edited by G. M. Sessler, (Springer-Verlag, New York, 1987), p. 347.
80. S. Mascarenhas and A. A. de Carvalho, *IEEE Trans. Elect. Insul.*, **EI-27**, 835 (1992).
81. E. Fukada, *IEEE Trans. Elect. Insul.*, **EI-27**, 813 (1992).
82. S. Bauer-Gogonea and R. Gerhard-Multhaupt, *IEEE Trans. Diel. & Elect. Insul.*, **DEI-3**, 677 (1996).
83. S. Nishikawa and D. Nukijama, *Proc. Imp. Acad. (Tokyo)*, **4**, 290 (1928).
84. A. Gemant, *Phil. Mag.*, **20**, 929 (1935).
85. M. Latour, O. Guelorget, and P. V. Murphy, *Charge Storage, Charge Transport and Electrostatics with their Applications*, edited by Y. Wada, (Elsevier, Amsterdam, 1979), p. 175.
86. K. Kobayashi and T. Yasuda, *Ferroelectrics*, **32**, 181 (1981).
87. J. C. Hicks, T. E. Jones, and J. C. Logan, *J. Appl. Phys.*, **49**, 6092 (1978).
88. S. Bauer and S. B. Lang, *IEEE Trans. Diel. & Elect. Insul.*, **DEI-3**, 647 (1996).
89. H. J. Coufal, *Appl. Phys. Lett.*, **44**, 59 (1984).
90. H. J. Coufal, *Thin Solid Films*, **193-194**, 905 (1990).
91. H. J. Coufal and A. Mandelis, *Ferroelectrics*, **118**, 379 (1991).
92. S. Bauer and B. Ploss, *IEEE Trans. Elect. Insul.*, **EI-27**, 861 (1992).
93. E. K. Merabet, H. K. Yuen, W. A. Grote, and K. L. Deppermann, *J. Therm. Anal.*, **42**, 895 (1994).
94. C. N. Davis, *Air Filtration*, (Academic Press, New York, 1973).
95. B. Gross, "Radiation-Induced Charge Storage and Polarization Effects," in *Electrets*, 2nd edition, edited by G. M. Sessler, (Springer-Verlag, New York, 1987), p. 217.
96. G. W. Fabel and H. K. Henisch, *Phys. Status Solidi, A-6*, 535 (1971).

97. A. G. Holmes-Siodle, Nucl. Instrum. Methods, *121*, 169 (1974).
98. L. Adams and A. G. Holmes-Siedle, IEEE Trans. Nucl. Sci., *NS-25*, 1607 (1978).
99. H. Bauser and W. Ronge, Health Phys., *34*, 97 (1978).
100. S. Mascarenhas, "Bioelectrets: Electrets in Biomaterials and Biopolymers," in *Electrets*, 2nd edition, edited by G. M. Sessler, (Springer-Verlag, New York, 1987), p. 321.
101. B. Lipinski (Editor), *Electronic Conduction and Mechanoelectrical Transduction in Biological Materials*, (Marcel Dekker, New York, 1982).
102. E. Fukada, T. Takamatsu, and I. Yasuda, Jpn J. Appl. Phys., *14*, 2079 (1975).
103. J. J. Konikoff and J. E. West, *Annual Report of 1978 IEEE Conference on Electrical Insulation and Dielectric Phenomena*, (The IEEE Dielectrics and Electrical Insulation Society, New York, 1978), p. 304.
104. P. V. Murphy and S. Merchant, in *Electrets, Charge Storage and Transport in Dielectrics*, edited by M. M. Perlman, (Electrochemical Society, Princeton, 1973), p. 627.

6 Charge Carrier Injection from Electrical Contacts

When you can measure what you are speaking about, and express it in number, you know something about it; but when you cannot measure it, when you cannot express it in number, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your own thoughts, advanced to the state of science.

William Thomson, Lord Kelvin

An electrical contact is generally referred to as a contact between a metal and a nonmetallic material, which may be an insulator or a semiconductor. Its function is either to enable or to block carrier injection. Contacts such as metal–electrolyte contacts, electrolyte–insulator, or electrolyte–semiconductor contacts are also common electrical contacts. Electrical contacts are heterojunctions but normally exclude contacts between two different semiconductors, between two different metals, or between a semiconductor and an insulator. This chapter deals only with contacts between metal and nonmetallic materials.

6.1 Concepts of Electrical Contacts and Potential Barriers

When two materials with different Fermi levels are brought into contact, free carriers will flow from one material into the other until an equilibrium condition is established, that is, until the Fermi levels of both materials are aligned (meaning that the Fermi levels for electrons in both materials are equal at the contact). This net carrier flow will set up a positive space charge on one side and a negative space charge on the other side of the interface, forming an electric double layer. This double layer is generally referred to as the *potential barrier*, and the potential across it is called the *contact potential*. The function of this double layer is to set up an electric field to stop any further net flow of free carriers from one material to the other.

Thermodynamically, the flow of free carriers in both directions always exists, but it would be very small and equal in quantity, maintaining a statistically zero net flow under a thermal equilibrium condition.

The Fermi level E_F is sometimes called the *electrochemical potential* or simply the *chemical potential*. The Fermi level can be considered a reference level. A state of energy equal to E_F will have the same probability ($f = 1/2$) for being occupied or vacant. This implies that the probability for a state at the level ΔE above E_F to be occupied is equal to that at the level ΔE below E_F to be vacant.

In this section, we shall discuss the features of various types of potential barriers formed by some ideal contacts. Note that although the energy band diagrams might be used to analyze and to predict some consequences, the surface states, created by contaminating impurities, crystallographic defects, and lattice mismatching (which unavoidably exist in the interface), greatly affect the electrical performance of a contact. However, the physics of contacts, including the effects of surface states, is not yet fully understood; the science and technology of producing a desired electrical contact for a certain application (e.g., ohmic contact to an insulator) is not yet fully explored.

Since the wave functions or orbitals of the outermost electrons of atoms or molecules overlap to some extent in a solid (because of interatomic or intermolecular bonding), the charge carriers injected from a metallic electrode into a solid become delocalized inside the

solid. Charge carriers produced by thermal or optical excitation inside a solid usually do not alter the electrical neutrality of the solid as a whole. (They may cause a net space charge of one sign in a local domain and a net space charge of opposite sign in the other domain due to the difference in mobility and diffusion constant between electrons and holes. The solid as a whole, however, can always be assumed to be electrically neutral.) This is because such excitations produce either an equal number of electrons and holes, or an equal number of one type of free carrier (mobile) and the other type of bound charge (nonmobile). But charge carriers injected from a contact will produce a net space charge in the solid; this net space charge produces the so-called *space-charge limited* (SCL) current under an applied electric field.

6.1.1 Electrical Contacts, Work Functions, and Contact Potentials

The nature of a contact is a complex affair. No single crystal free of any kind of imperfection has been found in this world. There always exist structural imperfections (e.g., dislocations in a crystal lattice) and chemical imperfections (e.g., impurities already existing internally in the material or due to external contamination) even on a cleavage surface. Imperfections in structure are normally accompanied by imperfections in geometric shape; imperfections due to impurities always produce protuberances and depressions on the surface. It is very difficult to avoid impurity contamination, even for a surface cleaved in a vacuum chamber, because a vacuum of 10^{-12} torr still contains about 3×10^4 particles per cm^3 . However, a surface with protuberances gently undulating over it may be considered a mathematically smooth surface, while that with protuberances existing as irregular steep and jagged asperities is considered a rough surface.

When two surfaces are brought into contact, some parts of the surfaces may not be in contact and some parts may be in real contact, with mechanical actions and reactions between the surfaces.¹ A contact that is prevented from being intimate by an extraneous body (contam-

inating particle) is referred to as an *impeded contact*, because in such a contact the mechanical forces are transferred from one surface to the other through the intermediary of this extraneous body. In the absence of extraneous bodies, an intimate contact may be formed by long-range molecular forces with a finite gap between the two surfaces (close contact), or by much stronger short-range molecular forces (true contact). For mathematical analyses, we always assume that the contact is a true intimate contact and the gap between the surfaces is so small that it can be considered transparent to quantum-mechanical tunneling of electrons. However, a practical contact is not so ideal; imperfections on the surfaces, coupled with the nonhomogeneity of the solid, lead to the concepts of filamentary carrier injection, which will be discussed in Filamentary Charge-Carrier Injection in Solids in Chapter 7.

The simplest contact between a metal and a nonmetallic material is the contact between a metal and a vacuum. When two metallic plates are placed in parallel in a vacuum with a small separation, the current flow is negligibly small if the applied voltage across the two plates is small. This is not because there are no free electrons in the metal, nor is it because any electrons present in the vacuum are not mobile in the vacuum. Rather, it is because the electrons in the metal must surmount a potential barrier before they can leave the metal and enter the vacuum. This potential barrier between the highest energy level of the electrons in the metal (termed the *Fermi level* of the metal) and the lowest energy level of the electrons in the vacuum (termed the *vacuum level*) is called the *work function* of the metal, ϕ_m , as shown in Figure 6-1(a). It is given by

$$\phi_m = \zeta - E_{Fm} \quad (6-1)$$

in which ζ is the difference in potential energy of the electrons between the inside and the outside of the metal, and depends on the structure of the crystal and the condition of the surface. The higher the cohesive energy of the metal, the higher the work function, but, on the other hand, the work function may be appreciably altered by the presence of an absorbed or

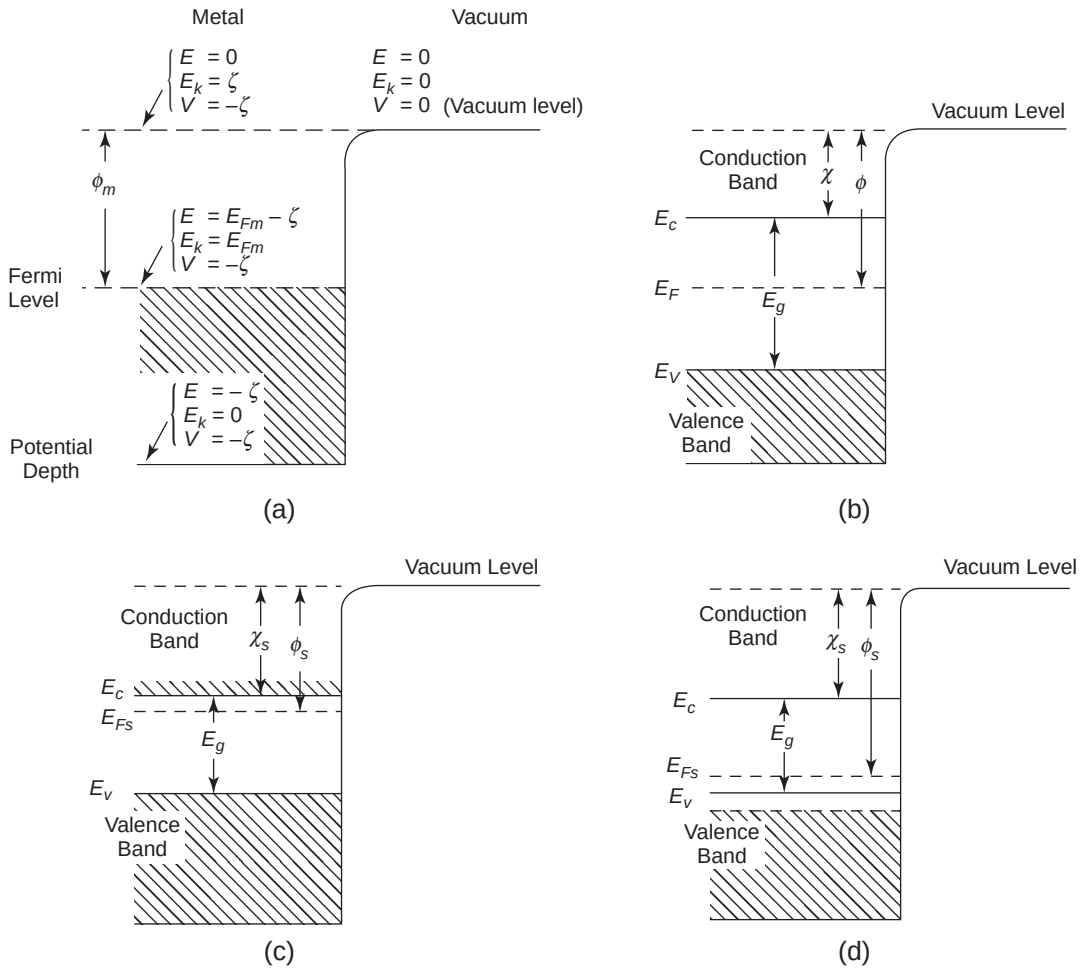


Figure 6-1 Energy band diagrams showing the work function of (a) metal, ϕ_m ; (b) dielectric ϕ ; (c) n-type semiconductor, ϕ_s ; and (d) p-type semiconductor, ϕ_s .

adsorbed layer of foreign atoms or molecules on the surface.²⁻⁴ Thus, the work function consists of two parts: the energy of binding the electron and the energy required to move the electron through an electrostatic double layer at the surface. This implies that ζ must depend partly on the structure of the surface and partly on the dipole moment of such a double layer.

A metal as a whole is electrically neutral, but at its surface facing a vacuum as a discontinuity, the electron distribution may be unsymmetrical with respect to the ion cores, resulting in the formation of a double layer in a way similar to that due to a net charge flow. If σ_s is

the charge per unit area, ϵ_0 the permittivity in vacuum, and t the thickness of the layer, then the dipole moment per unit area is $\sigma_s t$ and the potential is $\sigma_s t / \epsilon_0$. This potential may be positive or negative outwardly, depending on the double layer with positive or negative charge on the outside.

For clean metal surfaces, the magnitude of the dipole moment of such an intrinsic double layer is different for different orientations of the crystal plane, and the potential is of the order of 1/2 to 1 volt; for alkali metals, it is less than half a volt. For contaminated metal surfaces, however, the adsorbed layer of foreign atoms,

either neutral or ionized, may greatly modify the surface potential barrier and may then raise or lower the work function of the metal by more than 2 eV. For example, atoms of an electronegative gas such as oxygen adsorbed on the surface will capture electrons from the metal and form a layer of negative ions. This layer will in turn induce a layer of positive image charge in the metal. A double layer formed in this way, with the negative potential or negative charge outward, always tends to raise the work function. Conversely, cesium, barium or thorium atoms may give up their outermost electrons to the metal before they become adsorbed on the metal surface (such as tungsten) to form a double layer with the positive charge outward. Such a double layer always tends to lower the work function. The adsorbed layer may be electrically neutral, but it would still be polarized by the field at the surface of the metal to form a double layer. The effect of the layer formed by neutral atoms with or without permanent dipoles is smaller than that formed by ionized atoms.

The work function of a metal surface is therefore mainly determined by the top few layers of atoms, and not by the metal as a whole. Because the properties of the double layer are temperature dependent, the work function would also be expected to be temperature dependent.

Even if there is no contamination layer on the surface, the work function of a polycrystalline metal surface may vary from place to place. The domains of different work functions are patches. Thus, the average work function of a surface can be expressed as

$$\langle \phi_m \rangle = \sum_i a_i \phi_{mi} / \sum_i a_i \quad (6-2)$$

where ϕ_{mi} is the work function of patch i of area a_i . The potential outside the metal surface containing patches of different double layers is not constant but varies in the same manner as the work function. The resulting field between different patches is called the *patch field*. The presence of patches on the surface affects the measured value of the work function and the electron emission.⁵

The following sections describe the methods for determining the work function of a metal.

Contact Potential Measurements

Contact potential between two materials is defined as the difference in their work functions. If the work function of one material is known and the surface of this material is used as a standard surface, then the unknown ϕ_m of the other material can be determined by measuring the contact potential between the surfaces of these two materials using the Kelvin method.⁶ The whole system containing the two materials for contact potential measurement must be kept completely isothermal; otherwise, errors would be introduced due to the thermoelectric effect if a pair of surfaces were at different temperatures.

Photoelectric Emission Method

The photoelectric emission yield is a function of temperature. Fowler was the first to derive an expression for photoelectric emission yield as a function of work function, photon energy, and temperature,⁷ which is given by

$$J = AT^2 f\left(\frac{h\nu - \phi_m}{kT}\right) \quad (6-3)$$

or

$$\ell n \frac{J}{T^2} = B + F(x)$$

where $B = \ell n A$ and $F(x) = \ell n f\left(\frac{h\nu - \phi_m}{kT}\right)$. By plotting J/T^2 as a function of $h\nu/kT$, known as the *Fowler plot*, we can compare the theoretical Fowler plot based on Equation 6-3 with the experimental Fowler plot. A vertical shift of the data of fit $F(x)$ enables the determination of B , while a horizontal shift enables the determination of ϕ_m/kT . This is one of the most accurate ways to determine the work functions of metals, because in most cases the theoretical and experimental curves fit very well.

Thermionic Emission Method

This method is based on the common plot of Richardson lines following the equation

$$J = A(1-r)T^2 \exp[-\phi_m/kT] \quad (6-4)$$

or

$$\ell n \frac{J}{T^2} = \ell n A(1-r) - \frac{\phi_m}{kT}$$

The slope of the Richardson line would be $-\phi_m/k$, enabling the determination of ϕ_m . However, ϕ_m is most likely temperature dependent even for a clean metal surface,⁸ because interactions between the electrons in the solid have not been taken into account to derive Equation 6-4. If ϕ_m is a function of T , then the apparent work function ϕ_m^* is given by

$$\phi_m^* = \phi_m - T \frac{d\phi_m}{dT} \quad (6-5)$$

Although these three methods can be used to estimate the work functions of the metals, some discrepancies between measured ϕ_m are expected because of the uncertainties in measuring ϕ_m thermionically, and particularly because the measurement involves very high temperatures. Surface preparation for photoelectric emission measurements is extremely important, because patchiness of the surface means that ϕ_m is not constant over it, and this may spoil the fit of the Fowler plot. Again, the accuracy of the contact potential method depends so much on the choice of the standard surface of known ϕ_m and the surrounding ambient. However, the Fowler plot satisfactorily explains the temperature dependence of photoelectric emission, so that ϕ_m determined by the photoemission method may be considered as the true ϕ_m at 0 K.

In general, all the fundamental principles outlined above for metals can also be applied to nonmetallic materials. In nonmetallic materials, the Fermi level is always located within the energy band gap, except for degenerate extrinsic semiconductors. The work function for these materials is defined by

$$\phi = \chi + (E_c - E_F) \quad (6-6)$$

Figure 6-1 shows the definition of ϕ for metals, insulators, and n-type and p-type extrinsic semiconductors. The definition of ϕ for insulators is the same as for intrinsic semiconductors,

except that the energy band gap for the latter is normally much smaller. The conditions for electron emission from an insulator or from a semiconductor differ appreciably from those for metals. There are no electrons at the Fermi level E_F and in the forbidden gap, where there are no quantum states, so the electrons that may be removed from the interior of a nonmetallic material must be either in the conduction band, in the valence band, or in the impurity levels. In a metal, the minimum energy required to be imparted to an electron to remove it from the metal at $T = 0$ into a vacuum is the energy difference between the Fermi level and the potential energy given by Equation 6-1. At $T > 0$, the Fermi-Dirac distribution function becomes smeared out over a distance of the order kT , that is, over only a small fraction of an electron volt.

In an insulator or a semiconductor, the situation is different; only a very small portion of electrons in the conduction band requires the minimum excess energy (of the order of χ , the electron affinity) to leave the material to the vacuum level. Electrons at the impurity levels require higher energy, and those in the valence band require much higher energy to do so. After an electron has left the material, the remaining electrons in the material restore their statistical distribution. If an electron in the conduction band receives an energy greater than χ and leaves the material, its place is immediately taken by an electron either from the impurity levels or from the valence band. Since electron distribution inside the material before and after the emission of an electron is determined by the energy levels of these electrons with respect to the Fermi level E_F , the free energy required for an electron emitted from the material to the vacuum level is the work function defined by Equation 6-6. The work function of a material depends on the location of E_F , which is a function of temperature, impurity concentration, external pressure, etc.

However, in insulators and semiconductors, the electron affinity χ is an important quantity defined as the energy required for an electron to be removed from the bottom edge of the conduction band at the surface to a point in the vacuum just outside the material. In insulators

or high-resistivity nondegenerate semiconductors, χ can be determined by measuring the threshold wavelength for photoemission corresponding to the energy separation between the electron energy in the vacuum just outside the surface and the highest energy level that the electrons occupy in the material, that is, $\chi + E_g$. The yields associated with the threshold wavelengths corresponding to the transitions from the conduction band, from impurity and trapping levels, and from surface states are usually very small. However, the following conditions must be satisfied in order to determine χ accurately using this method:

- The materials must have a high resistivity and be nondegenerate so that the photoelectric threshold is insensitive to the height of the potential barrier. This means that the band bending over the escape depth of the emitted electrons must be negligibly small.
- The concentration of surface states must be very small so that the photoelectric yield from the valence band, rather than from the surface states, is predominant.

Contact potential is defined as the potential difference created between two dissimilar materials when they are brought into intimate contact. It is basically equal to the difference in the work functions of the two materials. Taking a metal–n-type semiconductor contact with $\phi_m > \phi_s$ as an example, the contact potential is given by

$$V_d = \frac{1}{q}(\phi_m - \phi_s) = \frac{1}{q}[\phi_m - \chi_s - (E_c - E_F)] \quad (6-7)$$

Since $E_c - E_F$ is sensitive to temperature and impurity (particularly donor) concentration, V_d depends strongly on temperature and impurity concentration. Figure 6-2(a) shows that before contact, the electrons in both the isolated metal and the semiconductor experience a short-range binding force exerted by the lattice of their own crystal (metal or semiconductor). This implies that the electrons must surmount a very steep potential barrier in order to leave the solid and go into the vacuum. If we now allowed the elec-

trons to flow from one solid to the other by making an electrical contact between the two solids (e.g., by connecting a metal wire to the back faces of the solids, leaving the front faces not in contact), as shown in Figure 6-2(b), there will be a net flow of electrons from the n-type semiconductor to the metal. This is because $\phi_s < \phi_m$, so a space charge will build up near the two surfaces to hinder the electron flow. The net flow will cease (although the electron flow from both solids always exist due to thermal excitation) when the space charge density builds to such a level that the Fermi levels in the metal and in the n-type semiconductor are aligned to the same height. This means that a thermal equilibrium has been reached between these two solids. The space charge will establish an electrostatic field and hence a potential difference between the metal and the semiconductor bulks, which is generally referred to as the *contact potential* and sometimes called the *diffusion potential*, because in the space charge region, the carrier movement is due mainly to a diffusion process, since the thickness of this region is normally very large in comparison with the mean free path of the carriers. Under the condition shown in Figure 6-2(b), in which the separation between the front surfaces of the metal and the semiconductor is still large, the major portion of the contact potential is the potential difference across the vacuum gap; only a small portion is due to the small space charge inside the semiconductor. When the separation between these two solid front surfaces decreases, the positive space charge region in the n-type semiconductor will extend much farther into the bulk. Electrons are driven farther away from the surface due to the proximity effect, so the portion of the contact potential across the vacuum gap decreases and the portion across the space charge region in the n-type semiconductor increases.

When the separation reduces to the value of the order of interatomic distance and the contact becomes an intimate contact, practically the whole contact potential will be across the space charge region, the portion across the vacuum gap (atomic scale) being negligibly small, as shown in Figure 6-2(c). In general,

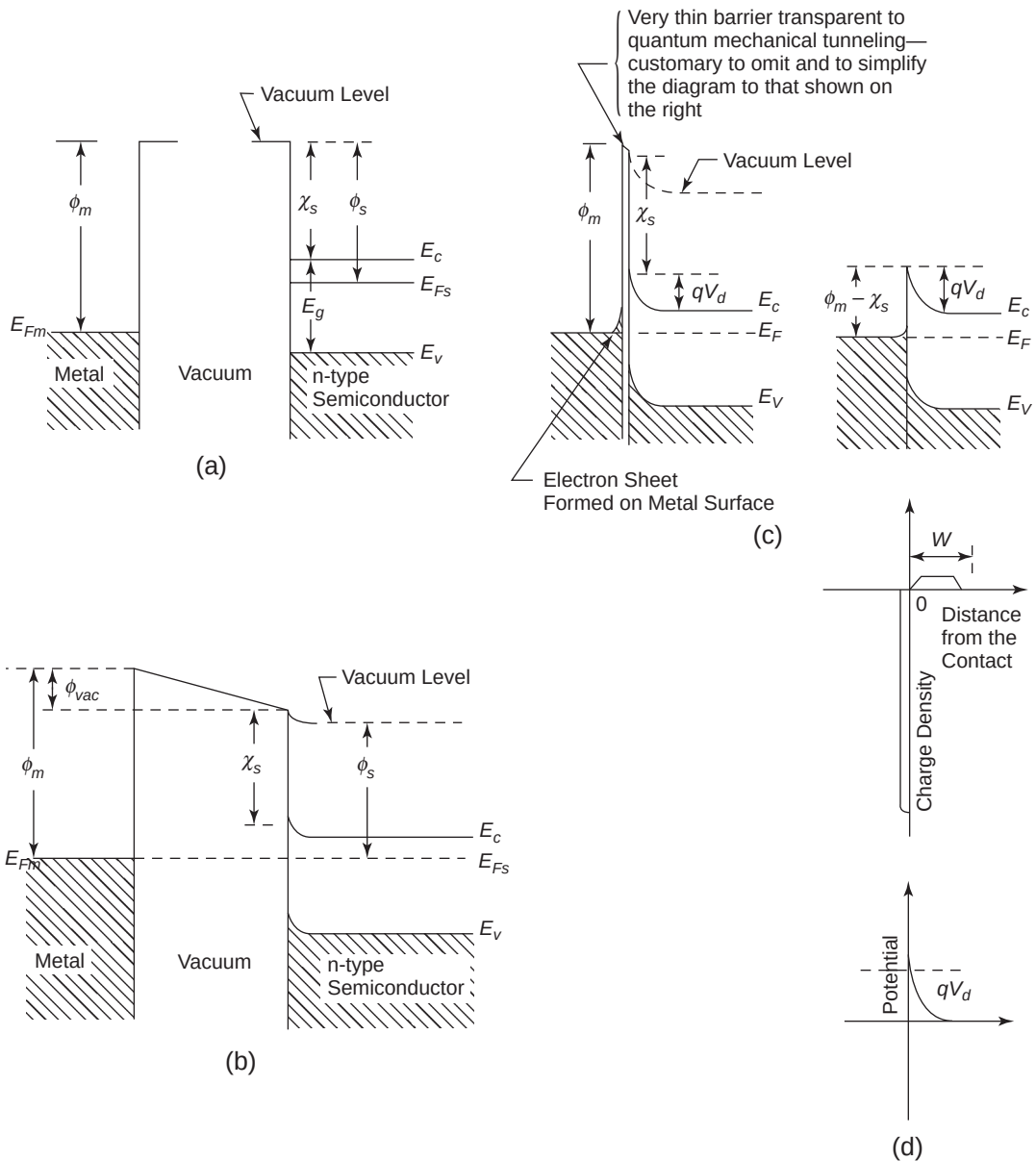


Figure 6-2 Energy level diagrams for a contact between a metal and an n-type semiconductor for $\phi_m > \phi_s$ and without surface states: (a) before contact, (b) in thermal equilibrium, (c) in intimate contact, and (d) charge density distribution and contact potential.

this extremely thin barrier between the two surfaces is omitted in the energy level diagram because it is transparent quantum-mechanically to electron tunneling. Because the free carrier density in the semiconductor is much smaller

than that in the metal, the space charge region extends much farther into the bulk of the semiconductor, as shown in Figure 6-2(c). The space charge region in the metal side is very thin and can be thought of as an electric charge sheet at

the surface, containing charge carriers equal in quantity but opposite in sign to those in the semiconductor. This also implies that a potential barrier is naturally accompanied by a double layer.

The width of the double layer or the space charge region is the width of the potential barrier denoted by W in Figure 6-2(d). The electrons at the bottom of the conduction band of the semiconductor must have an energy equal to or larger than the height of the potential barrier qV_d before they can leave the semiconductor and pass into the metal side. Similarly, the electrons at the Fermi level of the metal must have an energy equal to or larger than the height of the potential barrier $\phi_m - \chi_s$ before they can be injected from the metal into the semiconductor. Potential barriers of this type, which are formed by a space charge double layer and whose width and height are dependent on applied voltage, are generally referred to as *Schottky barriers*. If the height of a potential barrier is of the order of or smaller than the thermal energy kT , or if the width of a potential barrier is of the order of or smaller than the wavelength of a conduction electron or hole, this means that the carriers may quantum-mechanically tunnel through, and the barrier does not perform effectively as a barrier.

Before closing this section, we would like to clarify the concept of the vacuum level shown in Figure 6-2. The vacuum level serves as a reference for the potential energy of the electrons at a given position. In an isolated homogeneous material, it corresponds to the potential energy of an electron at a point where the attractive force between the electron and the surface of this material is negligible. In fact, the vacuum level does not have any absolute meaning. It represents only the relative energy of electrons at rest located just outside the various regions of the material at a position not influenced by the attraction force from the surface. However, the potential energy ϕ_{vac} in the vacuum level, due to the difference in work function between two different materials—see Figure 6-2(b)—will create a built-in field in the vacuum gap

$$F_m = \frac{1}{q} \frac{\partial \phi_{\text{vac}}}{\partial x} \quad (6-8)$$

even in the absence of an applied voltage. Thus, under such a condition, the electrons at the surface would experience a force in the x direction due to F_{in} .

6.1.2 Types of Electrical Contacts

To define different types of electrical contacts, we will choose a metal-insulator-metal (MIM) system, assuming that the insulator is intrinsic or contains a low concentration of traps and that the two metallic electrodes are identical, unless otherwise stated. Before an intimate contact is made, we assume that the work function of the metal ϕ_m and that of the insulator ϕ are not equal. Therefore, after they are brought into contact, charge transfer between the electrode and the insulator will prevail until the Fermi levels of the electrode and the insulator are aligned to the same height. Depending on the values of ϕ_m , ϕ , and other conditions, there are many types of electrical contacts, which are discussed in the following sections.

Neutral Contacts

The word *neutral* implies that the regions adjacent to the contact on both sides are neutral electrically. To satisfy the condition of electrical neutrality, no space charge will exist and no band bending will be present within the insulator, so both the conduction and the valence band edges will be flat right up to the interface. A condition like this is sometimes referred to as the *flat band condition*. The possibilities for neutral contacts are

- When $\phi_m = \phi$, the contact is neutral, as shown in Figure 6-3(a), because when the metal and the insulators are brought into contact, the probability that the electrons will flow from the metal to the insulator is equal to the probability that the electrons will flow in the reverse direction. Thus, there is no net flow and hence no space charge formed near the interface.
- When $\phi_m \neq \phi$ (either $\phi_m > \phi$ or $\phi_m < \phi$) at low temperatures, or with an electron trapping level at a distance sufficiently above E_F (or a hole trapping level below E_F) in wide-bandgap insulators, the contact can be

neutral because the trapped space charge will be too small under such conditions to cause significant band bending.⁹ A neutral contact is defined as one in which the carrier concentration at the contact is equal to that in the bulk of the insulator.

- When the work functions of electrode 1, electrode 2, and the insulator follow $\phi_{m1} < \phi < \phi_{m2}$, the contact can still be neutral, as for the case of wide-bandgap insulators at low temperatures. However, as discussed in Section 6.1.1, there will be a potential difference across the insulator given by

$$\begin{aligned} V_{12} &= \frac{1}{q} [(\phi_{m2} - \chi) - (\phi_{m1} - \chi)] \\ &= \frac{1}{q} (\phi_{m2} - \phi_{m1}) \end{aligned} \quad (6-9)$$

as shown in Figure 6-3(b), and a built-in field in the insulator is given by

$$E_{12} = -\frac{dV_{12}}{dx} \approx -\frac{V_{12}}{d} \quad (6-10)$$

which can be very large for insulating thin films for which d is small.

The zero-bias built-in field existing within the insulator is caused by the charge transfer between the electrodes: electrode 1 with a lower work function transfers electrons to electrode 2. The amount of charges transferred is

$$Q = \frac{(\phi_{m2} - \phi_{m1})A\epsilon_r\epsilon_o}{qd} \quad (6-11)$$

We now return to Figure 6-3(a). If a DC voltage V is applied between the two electrodes, and electrode 1 (the cathode) can supply a maximum electron density n_o through a thermionic emission process to maintain the current flow in the insulator, then current recorded at electrode 2 (the anode) is given by

$$J = qn_o\mu\frac{V}{d} = qn_o\mu F \quad (6-12)$$

Equation 6-12 follows Ohm's law, and the contact is ohmic if the following three conditions are met:

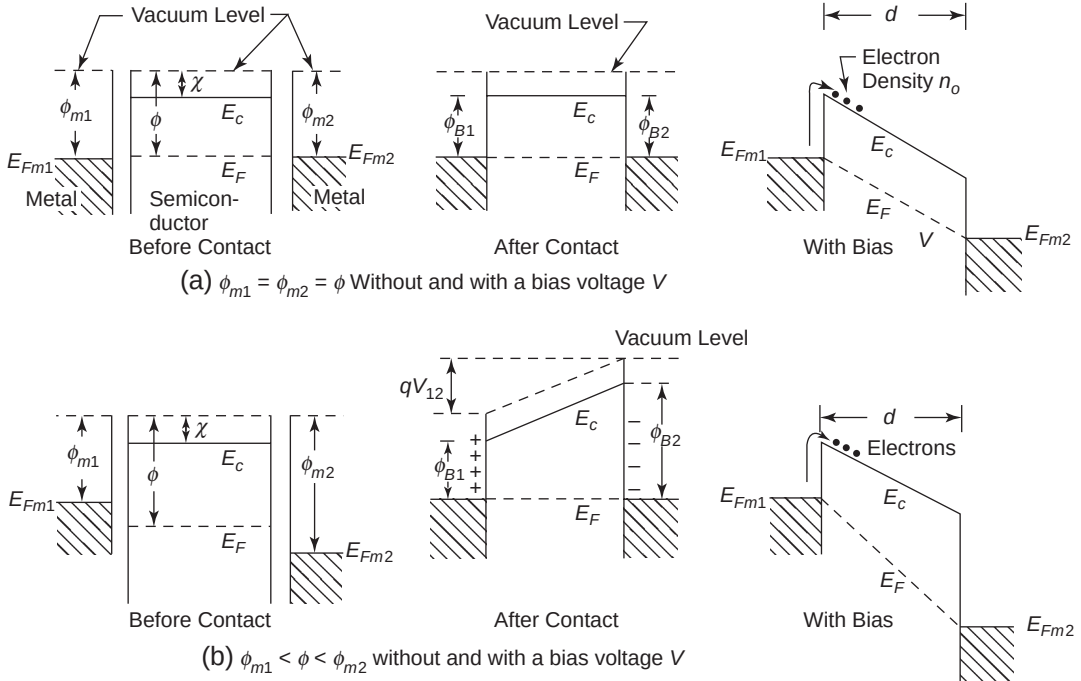


Figure 6-3 Energy level diagrams for a neutral contact between a metal electrode and an insulator in MIM systems. $\phi_{B1} = \phi_{m1} - \chi$, $\phi_{B2} = \phi_{m2} - \chi$.

- There is no band bending, so F is constant throughout the insulator for a given V .
- μ is independent of F . This requires that the current cannot be large enough to cause the change of μ with F through the Joule heating effect.
- The current drawn through the insulator is less than the saturated thermionic emission current from the cathode.

J is proportional to F until J is equal to the saturated thermionic emission current, for which F becomes F_o . When F is increased beyond F_o , the thermionic emission current is no longer capable of replacing the electrons drawn out at the anode. Under such a condition, the contact ceases to be ohmic and tends to become blocking, and the conduction becomes electrode-limited.

Blocking Contacts

The Schottky barrier, shown in Figure 6-2, is formed by an electron blocking contact for which $\phi_m > \phi_s$. The condition for a contact to be blocking, seen by electrons from the metal, is $\phi_m > \phi_s$ for a metal–n-type semiconductor junction, or $\phi_m > \phi$ for a metal–intrinsic semiconductor (or metal–insulator) junction. Under such a condition, electrons will flow from the semiconductor to the metal, leaving a positive space charge region (called a *depletion region*) in the semiconductor, as shown in Figure 6-4. W is the width of the depletion region, and ϕ_B is the height of the potential barrier that an electron in the metal must surmount in order to pass into the semiconductor. Such a contact is sometimes referred to as a *rectifying contact*, because under forward bias electrons can flow easily from the semiconductor to the metal, while under reverse bias the flow of electrons from the metal is limited by the electrons available over the Schottky barrier, the density of which is much smaller than that in the bulk of the semiconductor. So, a blocking contact can be defined as one that creates a depletion region extended from the interface to the inside of the semiconductor. With this contact, the thermionic emission from the metal tends to

be saturated. This is why, from the current-injection point of view, such a contact is called the *blocking contact* and why the conduction is electrode-limited under reverse bias. Electron emission from a metal across the blocking contact may be due either to a thermionic process or to a high-field tunneling process. These will be discussed in Section 6.2.

The condition for a contact to be blocking, seen by holes from the metal side (or by electrons from the opposite side), is $\phi_m < \phi_s$ for a metal–p-type semiconductor junction, or $\phi_m < \phi$ for a metal–intrinsic semiconductor (or metal–insulator) junction. Such a contact will block the hole emission from the metal, as shown in Figure 6-4.

Ohmic Contacts

An ohmic contact between a metal and a semiconductor is defined as one with a negligibly small impedance compared to the series impedance of the bulk of the semiconductor. This implies that the free carrier density at and in the vicinity of the contact is much greater than that in the bulk of the semiconductor (e.g., thermally generated carriers in the bulk), so the contact may act as a reservoir of carriers. An ohmic contact can also be defined as one that creates an accumulation extended from the interface to the inside of the semiconductor. Unfortunately, the term *ohmic* is not appropriate insofar as the current–voltage relationship is not linear. With ohmic contacts, the current–voltage relationship is nonlinear and depends on many factors. This will be discussed in detail in Chapter 7 in High-Field Effects and Bulk-Limited Electrical Conduction Involving One Type of Carriers. In general, conduction is ohmic at low fields if the metal does not inject carriers, and becomes nonlinear or nonohmic when the carrier injection from the electrode or the space charge effect becomes predominant.

There are two ways of making ohmic contacts:

Method 1: Choose metals of low work functions so that $\phi_m < \phi_s$ (for metal–n-type semiconductor junctions) or $\phi_m < \phi$ (for metal–intrinsic semiconductor or

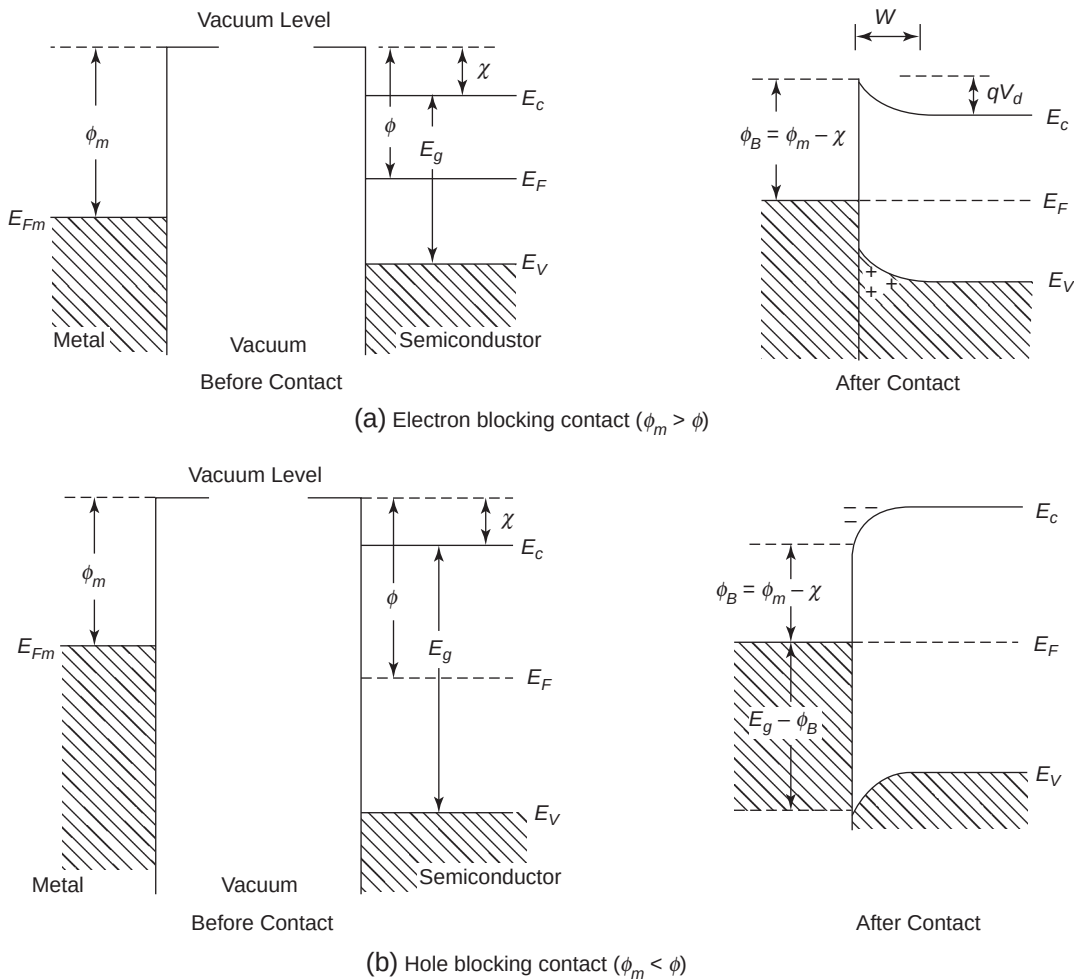


Figure 6-4 Energy level diagrams for a blocking contact between a metal and an intrinsic semiconductor (or an insulator).

metal–insulator junctions) for electron injection, or choose metals of high work functions so that $\phi_m > \phi_s$ (for metal–p-type semiconductor junctions) or $\phi_m > \phi$ (for metal–intrinsic semiconductor or metal–insulator junctions) for hole injection. This lowers the potential barrier for efficient thermionic emission to make the free carrier density higher (or the impedance smaller) at the contact than that in the bulk of the semiconductor.

Method 2: Dope the semiconductor surface heavily near the contact to make the poten-

tial barrier thin enough for efficient quantum-mechanical tunneling. In general, the resistivity of most insulators is very large, so the electrical contact impedance is normally considered negligibly small compared to the resistance of the insulator specimen.

For a case without surface states, the ohmic contact for an n-type semiconductor and the ohmic contact for a p-type semiconductor are shown schematically in Figure 6-5, based on method 1 just described. The work functions of the metals or alloys commonly used for electrical contacts are within the range of 3 to 5 eV.

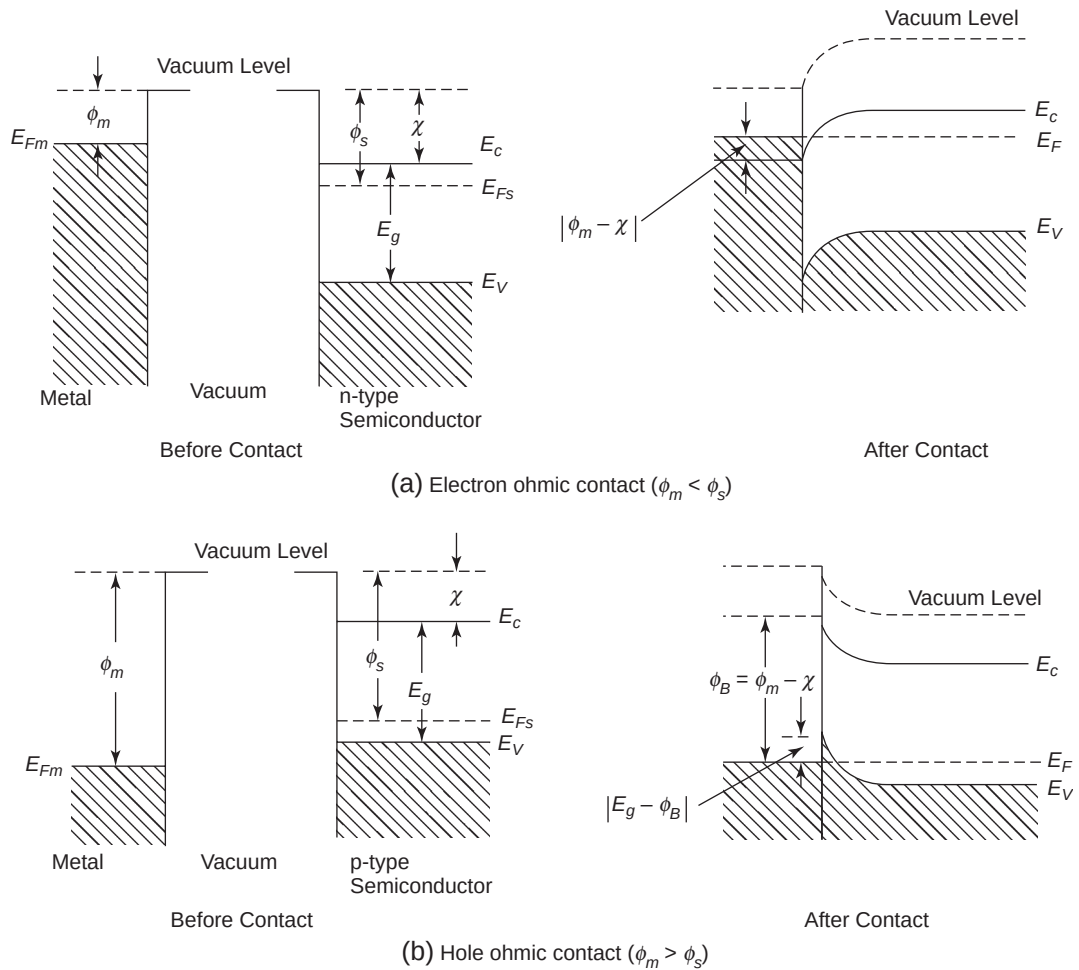


Figure 6-5 Energy level diagrams for an ohmic contact between a metal and an extrinsic semiconductor: (a) n-type, (b) p-type.

It is usually difficult to find a suitable metal with $\phi_m < \phi_s$ for the ohmic contact to n-type semiconductors, or with $\phi_m > \phi_s$ for the ohmic contact to p-type semiconductors. Furthermore, most semiconductors for electronic devices are covalently bonded, so they have surface states. For these reasons, it is not possible to produce an ohmic contact simply based on method 1. Even though it may be possible for other semiconductors, it is still not practical, because the behavior of such ohmic contacts is not reproducible.

The method for obtaining good, reliable ohmic contacts for most semiconductor devices

is based on method 2, that is, producing a very thin layer heavily doped with dopants by either diffusion or ion implantation techniques in order to make this layer become degenerate. Such a layer is called the n^+ layer for n-type semiconductors and the p^+ layer for p-type semiconductors. After this layer has been produced, any metal or alloy can be deposited on the surface of this layer to form a good ohmic contact.

Let us take the ohmic contact for n-type semiconductors as an example. The n^+ layer provides a narrow barrier width for electrons to tunnel quantum-mechanically from the metal

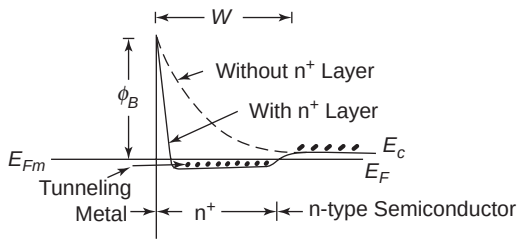


Figure 6-6 The formation of an ohmic contact to n-type semiconductors by means of a very thin n^+ layer.

electrode to the conduction band of the semiconductor, as shown in Figure 6-6.

It is important to understand the behavior of the normal ohmic contact between a metal electrode and an insulator (or an intrinsic semiconductor). Figure 6-7 shows the energy level diagram for an ohmic contact without surface states. The distributions of charge carriers and potential in the insulator are governed by Poisson's equation

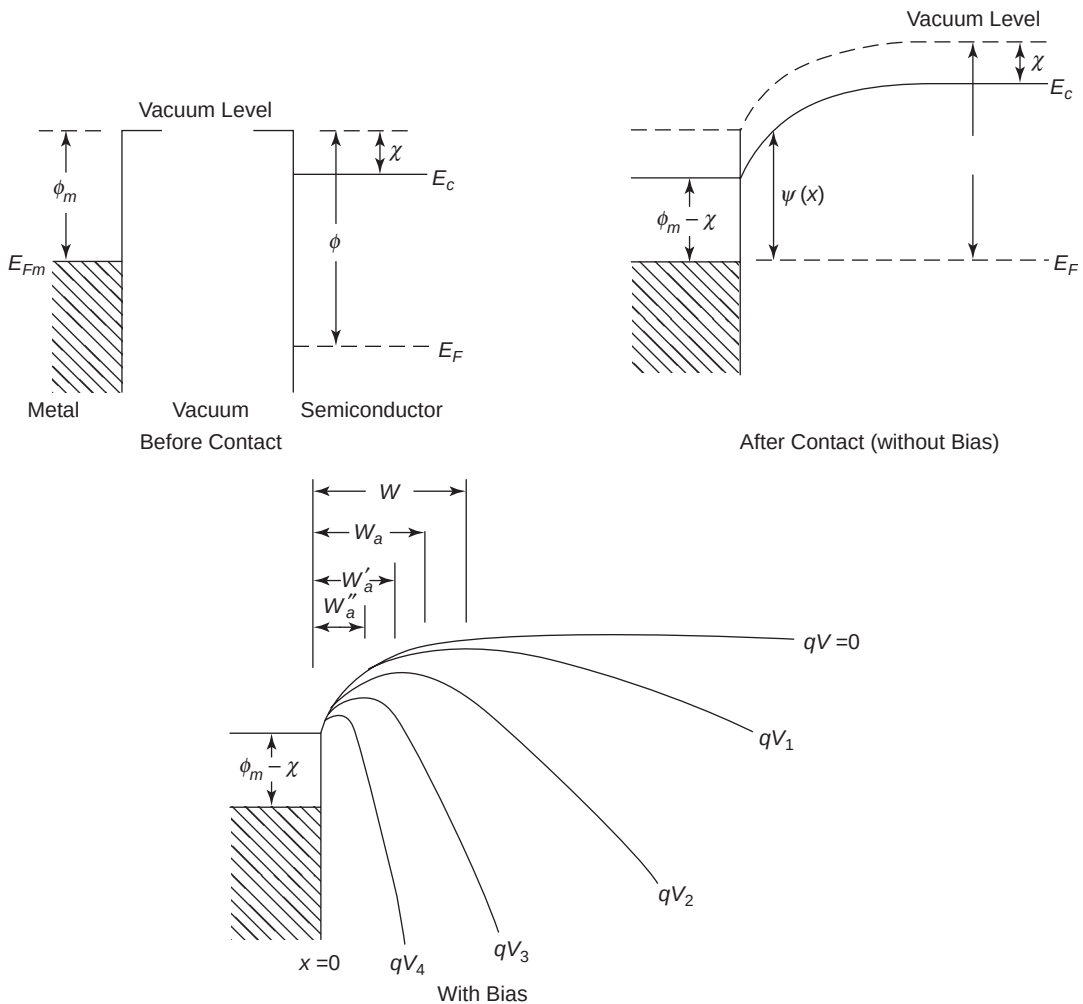


Figure 6-7 Energy level diagrams for an ohmic contact between a metal and an intrinsic semiconductor (or an insulator). $\phi_m < \phi$ and applied voltages $V_4 > V_3 > V_2 > V_1 > 0$.

$$\frac{dF}{dx} = \frac{qn}{\epsilon} \quad (6-13)$$

and the current flow equation

$$J = q\mu nF - qD \frac{dn}{dx} = 0 \quad (6-14)$$

since there is no external applied field. Using the boundary condition (see Figure 6-7)

$$\frac{\psi(x)}{q} - \frac{\phi_m - \chi}{q} = - \int_0^x F dx \quad (6-15)$$

and the Einstein relation

$$\frac{D}{\mu} = \frac{kT}{q} \quad (6-16)$$

we obtain

$$n(x) = n_s \exp[-(\psi - \phi_m + \chi)/kT] \quad (6-17)$$

where n_s is the electron density at the contact $x = 0$ (where $\psi = \phi_m - \chi$). Thus, Equation 6-13 can be written as

$$\frac{d^2\psi}{dx^2} = - \frac{q^2 n_s}{\epsilon} \exp[-(\psi - \phi_m + \chi)/kT] \quad (6-18)$$

using the boundary condition $\frac{d\psi}{dx} = 0$ when $\psi = \phi - \chi$ the solution of Equation 6-18 yields

$$\left(\frac{d\psi}{dx}\right)^2 = \frac{2q^2 n_s kT}{\epsilon} \{\exp[-(\psi - \phi_m + \chi)/kT] - \exp[-(\phi - \phi_m)/kT]\}$$

or

$$\frac{d\psi}{dx} = \left(\frac{2q^2 n_s kT}{\epsilon}\right)^{1/2} \{\exp[-(\psi - \phi_m + \chi)/kT] - \exp[-(\phi - \phi_m)/kT]\}^{1/2} \quad (6-19)$$

Integration of Equation 6-19 gives the width of the accumulation region

$$W = \left(\frac{2\epsilon kT}{q^2 N_c}\right)^{1/2} \exp\left[\frac{\phi - \chi}{2kT}\right] \times \left[\frac{\pi}{2} - \sin^{-1}\left\{\exp\left(-\frac{\phi - \phi_m}{2kT}\right)\right\}\right] \quad (6-20)$$

It is clear that when $\phi = \phi_m$, $W = 0$, the contact becomes neutral; that when $\phi - \phi_m < 4kT$, W increases with decreasing barrier height $\phi_m - \chi$;

and that when $\phi - \phi_m > 4kT$, Equation 6-20 reduces to

$$W \approx \frac{\pi}{2} \left(\frac{2\epsilon kT}{q^2 N_c}\right)^{1/2} \exp\left(\frac{\phi - \chi}{2kT}\right) \quad (6-21)$$

Under such a condition, W is independent of the barrier height $\phi_m - \chi$ and the electrode work function ϕ_m . Rather, W depends on the energy separation between the Fermi level and the bottom edge of the conduction band. In other words, it depends on the free carrier density in the bulk of the insulator. This implies that W increases with decreasing free carrier density in the bulk of the insulator.

Although the insulator or semiconductor itself may be intrinsic, the ohmic contact injects free carriers into the insulator in a quantity overwhelming those generated thermally inside the semiconductor. This makes the electrical conduction in the insulator or intrinsic semiconductor become extrinsic and space-charge limited in nature. In fact, even for an insulator, the ohmic contact always tends to inject electrons into the insulator when $\phi_m < \phi$, or to inject holes when $\phi_m > \phi$, raising or lowering the Fermi level in the insulator by an amount of $|\phi - \phi_m|$ in order to align the Fermi levels in the metal and the insulator. Since the ohmic contact acts as a reservoir of free charge carriers, the electrical conduction is controlled by the impedance of the bulk of the insulator (or of the intrinsic semiconductor) and is therefore bulk limited.

Because the injected space charge density decreases with increasing distance from $x = 0$ and reaches the value equal to that generated thermally in the bulk of the insulator at $x = W$, the internal field created by this accumulated space charge will decrease with increasing distance. It is now interesting to see how an applied electric field affects band bending. Figure 6-7 shows that when an average field F_{av} (applied voltage $V = V_1$ divided by specimen thickness) of low magnitude or of the order of the internal field near $x = W$ is applied, the applied field will be equal to and opposite of the internal field at $x = W_o$, where the product of the diffusion field and the charge carrier density is equal to the product of the applied

field and the charge carrier density in the bulk of the semiconductor.¹⁰ In general, we call this point the *virtual cathode* at which $\frac{dV}{dx} = 0$.

This situation is very similar to the space-charge limited situation in vacuum diodes due to thermionic emission. Under equilibrium conditions, the negative potential gradient at $x < W_a$ tends to send back to the contact all the electrons that represent the excess of the SCL current permitted by the insulator. Thus, at $x = W_a$, that is, at the virtual cathode, we can assume that the electrons are released without initial velocity. When the applied voltage is increased to $V_2 (>V_1)$, this field will balance a higher internal field at $x = W'_a (<W_a)$. This also implies that the electron density at $x = W'_a$ is higher than that at $x = W_a$, supplying a higher SCL current at a higher applied field. In Figure 6-7, the distances W , W_a , W'_a , W''_a are greatly exaggerated to illustrate the physical picture.

The higher the applied field, the closer the virtual cathode is to the contact.¹¹ When the applied field is so high that the virtual cathode coincides with the contact at $x = 0$, the effect of space charge ceases (because there is no accumulated charge region) and the conduction becomes ohmic, following Ohm's law. Beyond this, any further increase in applied field will make the conduction change from bulk limited to electrode limited, because the conduction becomes governed by the rate of electron injection from the cathode. However, if $\phi_m - \chi$ is large and the applied field is sufficiently high (but not high enough to cause breakdown), the potential barrier may become so thin that quantum-mechanical tunneling becomes important. The tunneling process through such a barrier will be discussed in Section 6.2.3.

For semiconductors or insulators containing shallow traps confined in a discrete energy level above E_F with an electron ohmic contact, as shown in Figure 6-8, the expression for the width of the accumulated region has been derived by Simmons⁹ and is given by

$$W = \frac{\pi}{2} \left(\frac{2\epsilon kT}{q^2 N_t} \right)^{1/2} \exp \left(\frac{\phi - \chi - E_t}{2kT} \right) \quad (6-22)$$

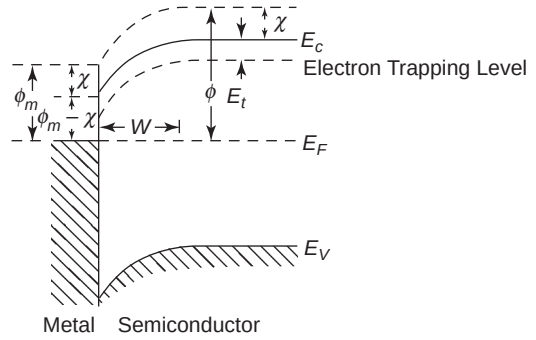


Figure 6-8 Energy band diagram illustrating an electron ohmic contact between a metal and a semiconductor (or an insulator) with shallow electron traps.

for $\phi - \phi_m > 4kT$ where N_t is the density of shallow traps and E_t is the trapping level measured from the bottom edge of the conduction band. For the effects of deep traps and traps of various distributions in energy, see reference.⁹

6.1.3 Surface States

A sudden termination of the lattice of a solid crystal at its surface usually gives rise to the distortion of the band structure near the surface, causing the bending of energy bands in order to bring the solid to an equilibrium state, or in other words, to make the Fermi level at the surface identical to that in the bulk. If a semiconductor surface is brought in intimate contact with a metal whose structure or interparticle distance is different from that of the semiconductor, localized surface states with energies lying within the forbidden energy gap may result because of the dangling bonds or the interruption (or discontinuity) of the periodic lattice structure at the interface. Bardeen¹² was the first to propose the existence of surface states to explain the experimental fact that the contact potential between a metal and a germanium or silicon is independent of the work function of the metal and the conductivity of the semiconductor. Shockley and Pearson¹³ were the first to observe experimentally the existence of surface states on semiconductor thin films.

Surface states originate partly from the discontinuities of the periodic lattice structure at

the surface, leading to the so-called *Tamm states*,¹⁴ associated with an asymmetric termination of the periodic potential and a large separation or weak interaction between atoms (or between molecules), or leading to the so-called *Shockley states*,¹⁵ associated with a symmetrical termination of the periodic potential and a small separation or strong interaction between atoms (or between molecules). Surface states also originate partly from foreign materials adsorbed on the surface. Because surface atoms in general are extremely reactive due to the presence of unsaturated bonds, the crystal surface generally is covered with one or more layers of a compound produced by a reaction between the surface atoms (or molecules) and their surroundings. Surface states also may be created by imperfect structure on the surface, such as abrasion on the surface due to grinding, cutting, etching, polishing, etc. It is important to note that apart from foreign impurities as a cause, there is no direct or unambiguous evidence that surface states lying within the forbidden energy gap must arise because of the sudden termination of an otherwise perfectly periodic lattice structure. However, it would seem reasonable to consider the connection of surface states with chemical binding (dangling bonds) at the surface. A dangling bond means that, in creating the surface, a domain from which an electron that normally participates in chemical binding in the solid has been removed. Created on the surface are surface states, which may overlap the states in the valence and conduction bands. But only those located within the forbidden gap play a role in trapping charge carriers. Surface states that do not involve impurities are referred to as *intrinsic surface states*; those dominated by impurities are *extrinsic surface states*.

Figure 6-9 shows schematically the typical intrinsic surface states formed on the surface of a covalently bonded crystal, such as on the bare surface of silicon or germanium. In this case, the atom at the surface has an unpaired electron in a localized orbital. Since there are no neighbors to bond to on the surface, the dangling bonds tend to capture available electrons there in order to complete their bonds. In fact, a dan-

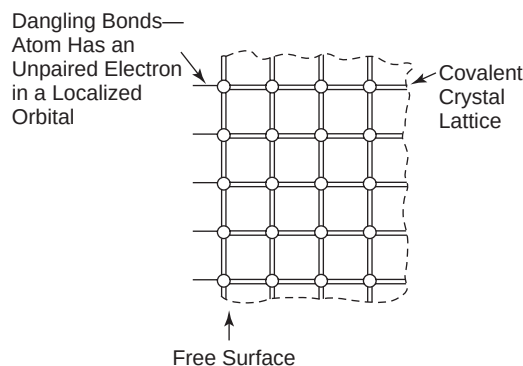


Figure 6-9 Schematic diagram showing the formation of dangling bonds on the surface of a covalently bonded crystal.

gling bond may either accept one electron and act as an acceptor or give up its unpaired electron and act as a donor. The charge neutrality condition, however, can allow only half of the atoms to assume one state; the other half assume the other state. It is interesting to note that for ionically bonded crystals, the wave functions of electrons associated with positive and negative ions overlap insufficiently to create surface states with energy levels located within the forbidden gap, and the outermost shell electrons are very tightly bound in ionic crystals, so the perturbing effect of the surface is small. For molecular crystals in which the molecules are bonded mainly by weak van der Waals forces, the two effects just mentioned for ionic crystals are much more pronounced. So, intrinsic surface states are not important in ionic and molecular crystals. But, there may exist extrinsic surface states created by adsorbed impurities on the surface.

Theoretically, the density of the intrinsic surface states in covalently bonded crystals should be equal to $(N)^{2/3}$, where N is the number of constituent atoms per cubic centimeter, which is about 10^{23} cm^{-3} , so the density should be about 10^{15} cm^{-2} . But experimental data show that the density of surface states is of the order of 10^{11} – 10^{12} cm^{-2} , which is much smaller than 10^{15} cm^{-2} . This difference may be attributed to the following two causes:

- Mutual saturation of the unsaturated bonds of neighboring atoms may take place on the surface. Several investigators have pointed out that the layer of surface atoms may have become distorted in such a way as to help mutual saturation of the unsaturated bonds of neighboring atoms, reducing the chance for the formation of surface states.^{16,17}
- Metal and semiconductor are, in fact, separated by a thin insulating layer. Lax¹⁸ has pointed out that surface states may be associated with flaws or impurities on the surface and that surface state density may not be related to the intrinsic structure of the solid. In practice, this is the case because preparation of the semiconductor surface always involves mechanical polishing, chemical etching, and exposure to an atmosphere containing oxygen. An oxide layer of 5–20 Å in thickness can easily be formed on the semiconductor surface before the deposition of metal. This oxide layer will screen the semiconductor surface from the metal and help

saturation of unsaturated bonds, as shown in Figure 6-10. The barrier height ϕ_B for such an MIS system depends on the thickness of the thin oxide layer, the density of the surface states on the oxide–semiconductor interface, and the applied voltage. A complete analysis of this case is quite mathematically involved and is not discussed in this section. The reader who is interested in the analysis is referred to the excellent book by Rhoderick.¹⁹

In general, gas atoms or molecules having an electron affinity greater than the work function of the semiconductor (or insulator) may capture electrons from the surface. Their behavior is like acceptors (acceptorlike surface states), tending to bend the energy bands upward. For example, if oxygen atoms, which are very electronegative, are adsorbed on the surface of a semiconductor (or insulator) to form surface states within the forbidden gap, these surface states would intercept electrons going toward the conduction band and leave some holes in

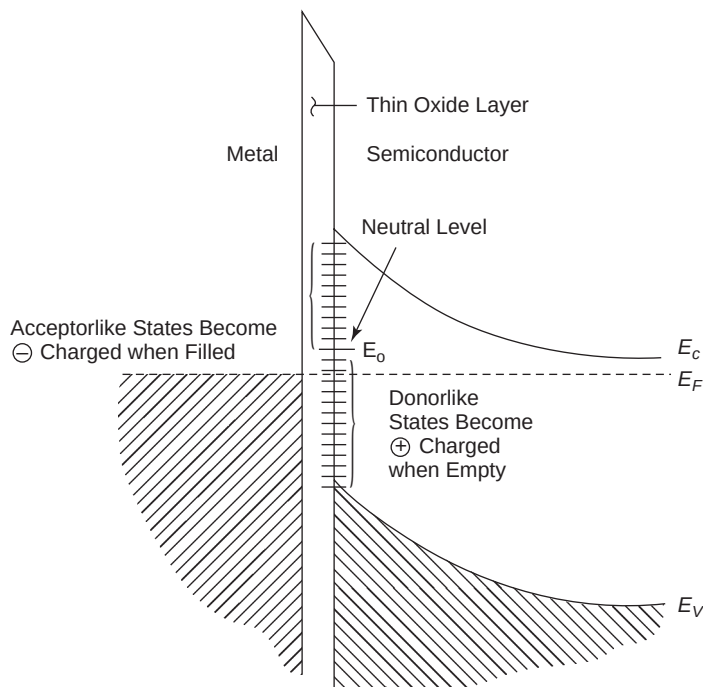


Figure 6-10 Formation of a thin oxide layer to screen the semiconductor surface from the metal and help saturation of unsaturated bonds.

the valence band, thus bending the bands upward. Therefore, electronegative atoms or molecules, such as oxygen, adsorbed on the surface act as acceptors, tending to produce a depletion region in an n-type semiconductor, and an accumulation region in a p-type semiconductor, as shown in Figure 6-11(a) and (b). Similarly, gas atoms or molecules of electropositivity adsorbed on the surface of a semiconductor or an insulator, such as chlorpromazine on anthracene surface,²⁰ will form donorlike surface states, as shown in Figure 6-11 (c) and (d).

In Figure 6-11, qV_s created by surface states dominates the contact potential and is independent of work functions if the surface state density is large. It should be noted that an

acceptorlike or donorlike state can be considered electrically neutral, but the acceptorlike one would become negatively charged after capturing an electron, while the donorlike one would become positively charged after giving up an electron (in other words, capturing a hole). In order to maintain charge neutrality, any excess charge in the surface states must be compensated by the change in free carrier concentration beneath the surface in the solid (equivalent to saying that the holes are attracted to the negatively charged surface or the electrons to the positively charged surface), thus forming a double layer. It can be imagined that the effects of surface states are very important in semiconductors and insulators because the space charge region produced by such a double

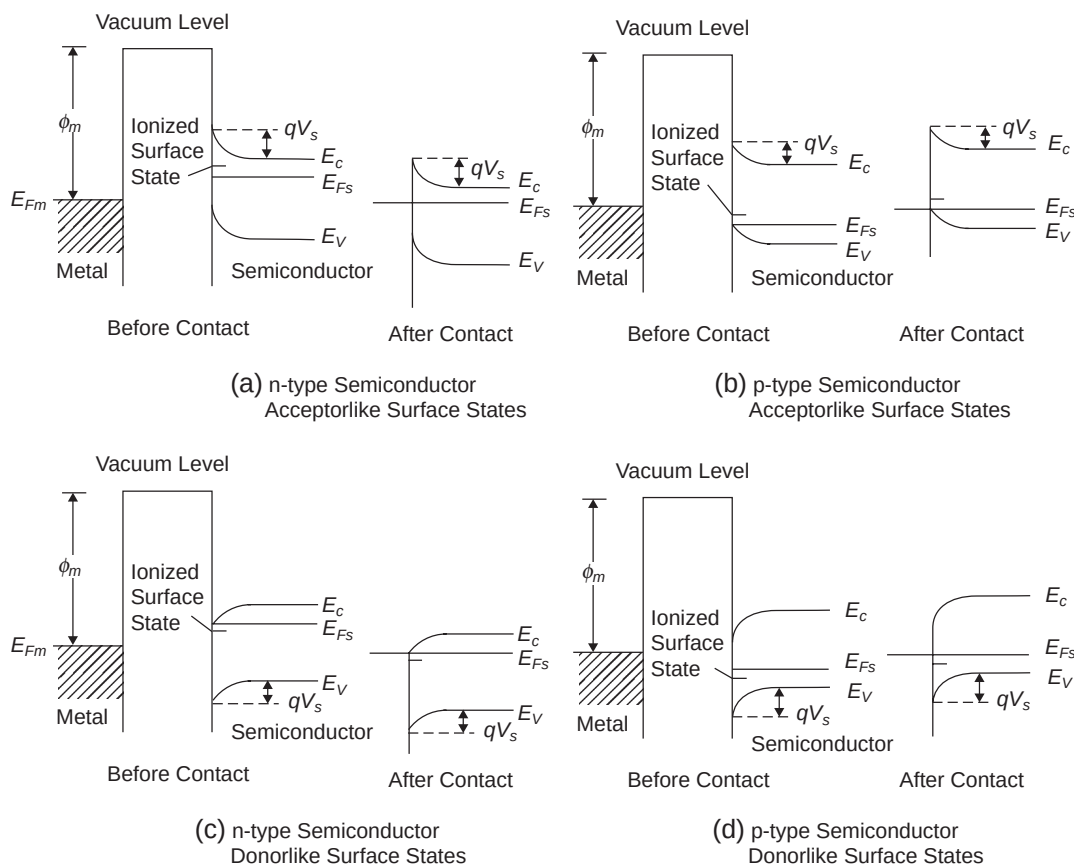


Figure 6-11 Energy level diagrams showing the effects of surface states on band bending: (a) and (b) for acceptorlike surface states, (c) and (d) for donorlike surface states.

layer usually extends to some depth, but these effects are not as important in metals, since a large quantity of free electrons present can easily compensate for any surface charges.

In general, surface states can be divided into fast and slow surface states, depending upon the speed with which they interact with the semiconductor space charge region. Fast surface states exist at the semiconductor–insulator (oxide) interface due mainly to inherent structure of the surface. The time required for their interaction with the carriers near the surface to reach an equilibrium state is of the order of nanoseconds or less and their density is about $10^{12}/\text{cm}^2$, lower than that of slow states. Fast surface states play an important role in carrier recombination processes, which greatly affect the electrical and optical properties of semiconductors. Slow surface states have much longer relaxation times, which are of the order of milliseconds or longer, and their density is of the order of 10^{13} – $10^{15}/\text{cm}^2$.²⁰ Slow surface states carry either positive or negative electric charges, depending on their nature (they may be acceptorlike or donorlike states). Slow surface states generally exist at the outer surface of the oxide layer (i.e., at the oxide–gas interface). Occupation of the slow surface states is affected by the ambient atmosphere, indicating that they originate or participate in the adsorption processes. There are two types of charges in an oxide layer:

Immobile charges: These are associated either with ionic defects in the oxide in which the electronic transfer takes place between the defects and the semiconductor during the formation of the oxide on the semiconductor surface (and usually not afterward), or with adsorbed ambient ions located at the semiconductor–oxide interface, which modify the distribution of semiconductor interface states. The immobile (also called *fixed* or *built-in*) charges do not participate in electron transfer processes across the interfaces.

Mobile charges: These are traps within the oxide. They can react electrically with the semiconductor with a three-dimensional distribution.

Since both fast and slow surface states act as traps, the density of such traps at the surface is much higher than the trap density in the bulk. That surface states act as a stepping stone to assist carrier injection from a metal electrode to polyethylene terephthalate has been reported by Takai et al.²¹ Adsorbed gases such as iodine would also greatly affect carrier injection.^{22,23}

6.2 Charge Carrier Injection through Potential Barriers from Contacts

In the 1920s, when copper oxide and selenium were the best known rectifiers, some people thought that rectification was a bulk effect. After 1930, it became generally accepted that rectification phenomena are associated with the potential barriers near the interface between a metal and a semiconductor. Section 6.1 mentioned that when a depletion layer is formed on the semiconductor side near the interface, the conduction becomes electrode limited because the free carrier concentration in this case is higher in the bulk than what the contact can provide. In this section, we shall consider only those potential barriers leading to an electrode-limited electrical conduction. This implies that current–voltage (J – V) characteristics are controlled by the charge carrier injection from the injecting contacts. There are two possible ways for charge carriers to inject from an electrode to an insulator (or a semiconductor): field-enhanced thermionic emission and quantum-mechanical tunneling.

6.2.1 Potential Barrier Height and the Schottky Effect

The potential barrier at the interface between a metal and an insulator (or a semiconductor) prevents the easy injection of electrons from the metal into the insulator. This section discusses the factors causing the lowering of such a barrier for neutral and blocking contacts.

Neutral Contacts

The phenomenon that the height of the potential barrier is lowered due to the combination of

the applied electric field and the image force is called the *Schottky effect*.²⁴ Figure 6-12 shows that the potential barrier height is $\phi_m - \chi$ if the image force is ignored and the applied electric field is zero. When these two parameters are taken into account, the potential barrier height, measured from the Fermi level of the metal, is given by

$$\psi(x) = \phi_m - \chi - \frac{q^2}{16\pi\epsilon x} - qFx \quad (6-23)$$

It should be noted that the potential energy due to image force $\frac{q^2}{16\pi\epsilon x}$ is not valid at $x = 0$. To avoid this situation in which $x = 0$ occurs, we assume that this expression is valid from $x = x_0$, corresponding to $\frac{q^2}{16\pi\epsilon x_0} = \phi_m - \chi$, to $x = \infty$, and that the image force is constant from $x = 0$ to $x = x_0$. We may also assume that the electron sea in the metal at E_F is extended to x_0 .

The image force tends to attract the emitted electrons back to the metal, while the driving force due to the applied field tends to drive the emitted electrons away from the metal. There is an optimal point at which the net force acting

on the electrons is zero and $\psi(x)$ becomes minimal. By setting $\frac{d\psi(x)}{dx} = 0$, we obtain $x = x_m$, corresponding to a minimum potential barrier height. Thus, we have

$$x_m = \left(\frac{q}{16\pi\epsilon F} \right)^{1/2} \quad (6-24)$$

and the total lowering of the potential barrier height

$$\begin{aligned} \Delta\phi_B &= (\phi_m - \chi) - \phi_B \\ &= \left(\frac{q^3 F}{4\pi\epsilon} \right)^{1/2} = \beta_{sc} F^{1/2} \end{aligned} \quad (6-25)$$

where $\beta_{sc} = \left(\frac{q^3}{4\pi\epsilon} \right)^{1/2}$ is called the Schottky constant.

So the effective potential barrier height can be written as

$$\phi_B = (\phi_m - \chi) - \left(\frac{q^3}{4\pi\epsilon} \right)^{1/2} F^{1/2} = \phi_B - \Delta\phi_B \quad (6-26)$$

which is field dependent.

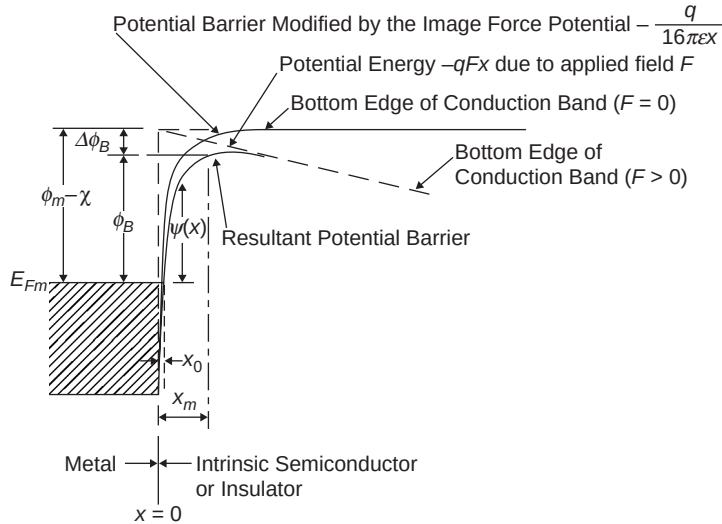


Figure 6-12 Energy level diagram showing the lowering of potential barrier due to the combination of the image force and the applied uniform field for a neutral contact.

Blocking Contacts

Following a simple Schottky barrier model and taking a blocking contact between a metal and an n-type semiconductor as an example (the same principle can be applied straightforwardly to p-type semiconductors or to intrinsic semiconductors or insulators with $\phi_m > \phi_s$ with the bands bending up to block the electron injection, or with $\phi_m < \phi_s$ with the bands bending down to block the hole injections), the width of the Schottky barrier is given by²⁵

$$W = \left[\frac{2\epsilon(\phi_m - \phi_s + qV)}{q^2 N_d} \right]^{1/2} \quad (6-27)$$

W is a function of applied voltage across the junction V . The derivation of Equation 6-27 is based on the following assumptions (Schottky model):

The barrier height ϕ_B is large compared with kT .

$N_{d(\text{ionized})} = \text{constant}$ for $0 < x < W$, and

$$N_{d(\text{ionized})} W \gg \int_0^W n dx, \quad N_{d(\text{ionized})} > P_s = p(x=0)$$

$$N_{d(\text{ionized})} W \gg \int_0^W p dx, \quad N_{d(\text{ionized})} > n_s = n(x=0)$$

$$N_{d(\text{ionized})} = n(x=W)$$

This implies that the barrier is practically depleted of free carriers.

The resistance is much higher in the barrier than outside the barrier, so the applied voltage can be considered completely absorbed across the barrier.

The mean free path of the electrons is small compared to W .

The potential energy of the Schottky barrier measured from the bottom edge of the conduction band at $x = 0$ is

$$-\frac{q^2 N_d}{\epsilon} \left(Wx - \frac{1}{2} x^2 \right) \quad (6-28)$$

This potential energy is equal to zero at $x = 0$ and equal to qV_d (contact potential energy $= \phi_m - \phi_s$ when $x = W$) as shown in Figure 6-13(a). This potential energy will increase with increasing applied voltage V (with negative potential at the metal), as shown in Figure 6-13(b). Thus, the total potential barrier height

measured from the Fermi level of the metal is given by

$$\psi(x) = \phi_m - \chi - \frac{q^2}{16\pi\epsilon x} - \frac{q^2 N_d}{\epsilon} \left(Wx - \frac{1}{2} x^2 \right) \quad (6-29)$$

Obviously, there is an optimal point at $x = x_m$ where $\psi(x) = \phi_B$ is minimal. By setting $\frac{d\psi(x)}{dx} = 0$, we obtain

$$x^3 - Wx^2 + (16\pi N_d)^{-1} = 0 \quad (6-30)$$

since $W \gg x_m$. By neglecting x^3 in Equation 6-30, we have

$$x_m = \frac{1}{4(\pi W N_d)^{1/2}} \quad (6-31)$$

Substituting Equation 6-31 into Equation 6-29 and assuming that $W \gg x_m$, we obtain

$$\phi'_B = \phi_m - \chi - \left[\frac{q^6 (\phi_m - \phi_s + qV) N_d}{2(8\pi)^2 \epsilon^3} \right]^{1/4} \quad (6-32)$$

which again is field dependent. The total lowering of the potential barrier is^{25,26}

$$\Delta\phi_B = \phi_m - \chi - \phi'_B = \phi_B - \phi'_B = \left[\frac{q^6 (\phi_m - \phi_s + qV) N_d}{2(8\pi)^2 \epsilon^3} \right]^{1/4} \quad (6-33)$$

It is important to mention that if $\phi_B < kT$, the electrons easily surmount the barrier and inject from the metal into the semiconductor, so the barrier does not function as a barrier to block the carrier injection. It is equally true that if the wavelength of the electron is larger than W although $\phi_B > kT$, the barrier becomes transparent to the electrons. In this case also, the barrier does not function as a barrier to block the carrier injection.

Here, we shall assume that W is much larger than the wavelength of electrons when considering thermionic emissions, so that the tunneling injection can be neglected. When considering tunneling field emission, we shall also assume that W and ϕ_B have such values that thermionic injection can be neglected. However, it should be noted that under certain

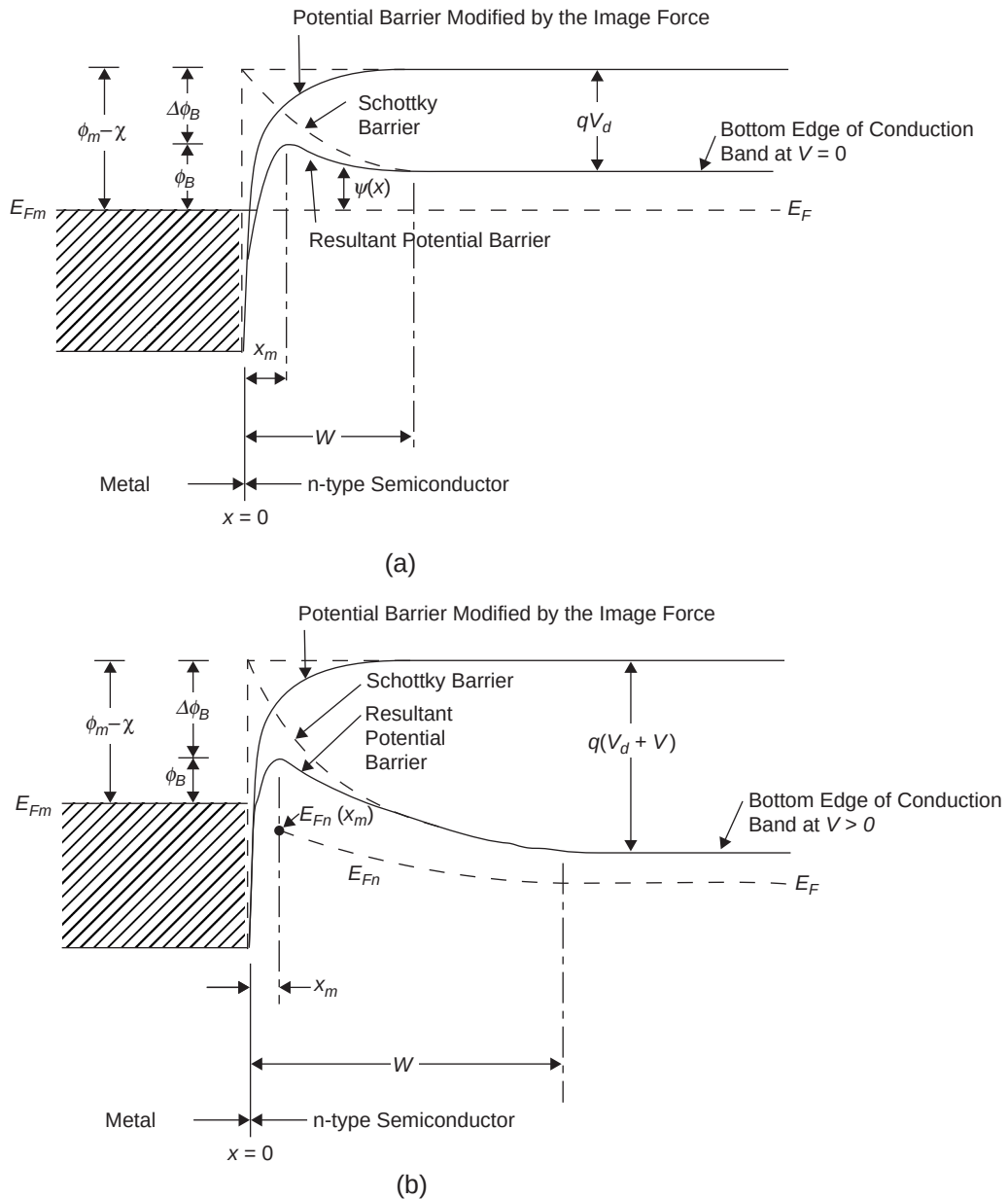


Figure 6-13 Energy level diagram showing the lowering of potential barrier due to the combination of the image force and the depletion layer effects for a blocking contact between a metal and an n-type semiconductor: (a) without applied voltage and (b) with applied reverse-bias voltage V .

conditions, both types of carrier injection may be equally important. Since the barrier width becomes much narrower at energy levels far removed from the Fermi level of the metal, particularly close to the peak of the barrier,

we shall also consider the case of thermally assisted tunneling field emission.

So far, we have dealt only with electrons: The image force attracts electrons back to the metal. The image force also attracts holes back

to the metal. Since the energy of a hole is measured downward from the top of the valence band, the effect of the image force is to bend the valence band upward near the surface of the metal, as shown in Figure 6-14. Unlike for electrons, there is no maximum in the top of the valence band, but the energy band gap is reduced close to the metal surface.¹⁹ It can be

seen from Figure 6-14 that both the barrier height for electron injection ϕ_{Bn} and for hole injection ϕ_{Bp} depend on the applied bias V . For forward bias (metal positively biased), the effective barrier height for electron injection ($\phi_{Bn} - \Delta\phi_{Bn}$) increases and the effective barrier height for hole injection ($\phi_{Bp} - \Delta\phi_{Bp}$) decreases with increasing forward bias. For reverse bias

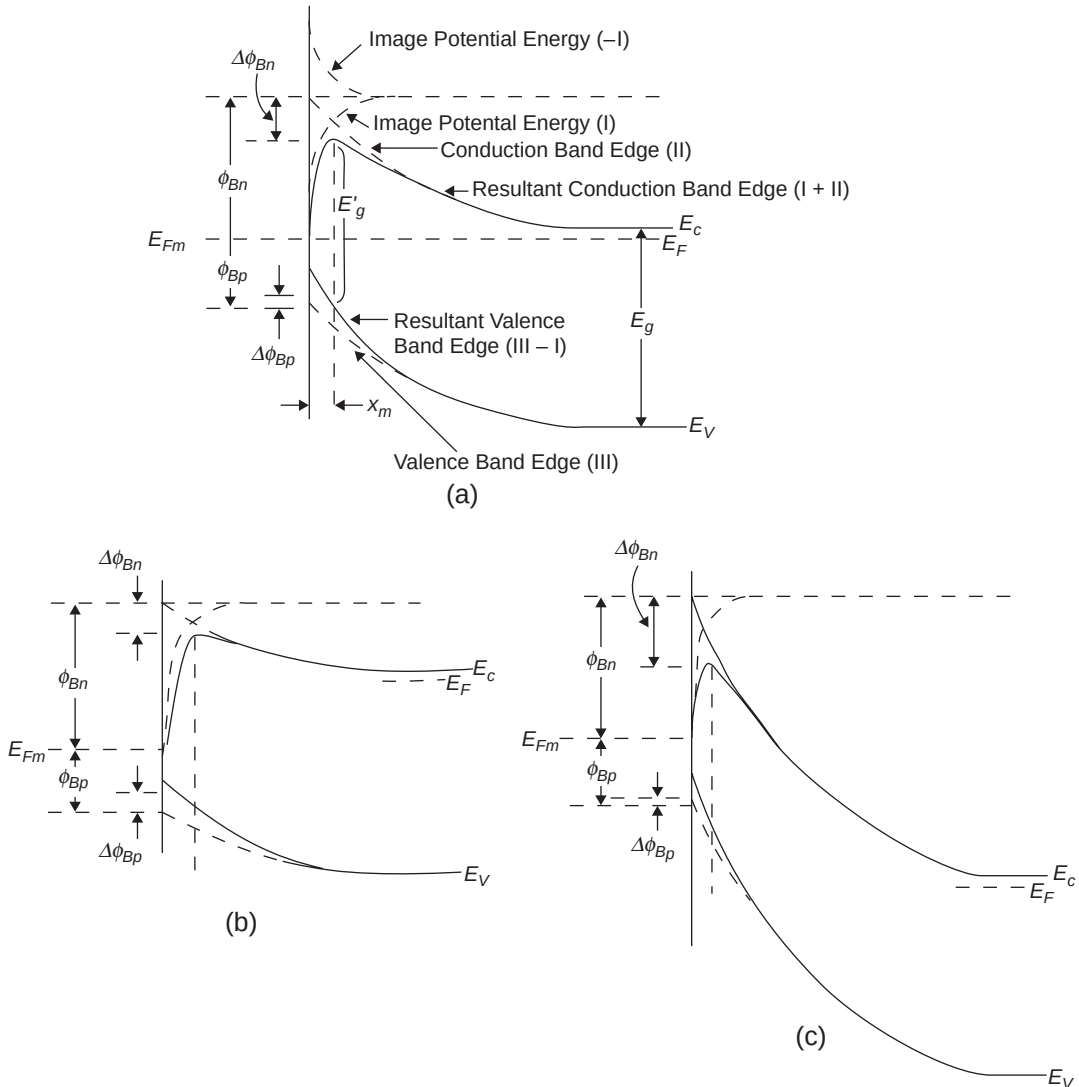


Figure 6-14 The effects of the image force on the effective potential barrier height $\phi_{Bn} - \Delta\phi_{Bn}$ for electron injection, and $\phi_{Bp} - \Delta\phi_{Bp}$ for hole injection under the conditions of (a) zero bias, (b) forward bias with the metal positively biased, and (c) reverse bias with the metal negatively biased. At zero bias, some electrons and holes will be emitted thermally, and they will feel the effects of the image force.

(metal negatively biased), the effective barrier height for electron injection decreases and that for hole injection increases with increasing reverse bias.

It is also important to remember the following concepts:

- The image force that lowers the potential barrier arises from the field produced by the particular electron in the conduction band under consideration. This image force effect is absent if such an electron is not in the conduction band near the top of the barrier.
- The image force reduces the energy band gap E_g close to the metal surface.
- The measurement of the barrier height based on the movement of the conduction electrons (i.e., on the current–voltage characteristics) yields $\phi_{Bn} - \Delta\phi_{Bn} = \phi'_{Bn}$. But if the measurement of the barrier height is based on the space charge (i.e., on the capacitance–voltage characteristics), it yields only ϕ_{Bn} , without the image force effect.

6.2.2 Thermionic Emission

If the electrons emitted from the cathode are not influenced by either space charge or traps—in other words, if all electrons emitted are carried away in the conduction band of the insulator and collected at the anode—the thermionic emission current density is given by¹⁹

$$J = A^* T^2 \exp[-(\phi_B - \Delta\phi_B)/kT] \times [\exp(qV/kT) - 1] = J_o [\exp(qV/kT) - 1] \quad (6-34)$$

where J_o is the saturation current density, which may be written in the form

$$J_o = A^* T^2 \exp[-(\phi_B - \Delta\phi_B)/kT] = qn\bar{v}$$

V is the applied voltage. V is positive for forward bias (metallic cathode positively biased) and negative for reverse bias (metallic cathode negatively biased). $n = N_c \exp[-\phi_B/kT]$,

and $\bar{v} = \left(\frac{2}{\pi} kT/m\right)^{1/2}$. It is obvious that for reverse bias, J should saturate at the value of J_o (i.e., $J = -J_o$ or $|J| = J_o$) because $\exp(qV/kT)$ is

much smaller than 1 if V is negative and $|qV| > 3kT$. The expression is self-explanatory. ϕ_B is given by Equation 6-26 or Equation 6-32, depending on the type of contact, and A^* is the Richardson constant, which is given by

$$A^* = A = \frac{4\pi q k^2 m}{h^3} = 120 A/cm^2 (\text{degree})^2 \quad (6-35)$$

for thermionic emission into a vacuum. For thermionic emission into a semiconductor or an insulator, A^* is quite different from A . The factors governing the value of A^* are discussed next.

Effect of Effective Mass

For isotropic materials, we can write

$$\frac{A^*}{A} = \frac{m^*}{m} \quad (6-36)$$

For anisotropic materials, Equation 6-36 becomes

$$\frac{A^*}{A} = \frac{1}{m} \left(\ell_1^2 m_y^* m_z^* + \ell_2^2 m_z^* m_x^* + \ell_3^2 m_x^* m_y^* \right)^{1/2} \quad (6-37)$$

where ℓ_1 , ℓ_2 , and ℓ_3 are the direction cosines relative to the principal axes of the constant energy ellipsoid, and m_x^* , m_y^* , and m_z^* are the corresponding components of the effective mass tensor.²⁷

Correction Due to Drift and Diffusion of Carriers in the Depletion Region

If the mobility of the carriers in the material is low, it will control the thermionic emission current. There are two approaches to this problem, depending on the type of contact.

Neutral Contacts

We can consider most metal–insulator contacts neutral contacts. Thus, we follow the approach of O'Dwyer et al.^{28,29} The total reverse current density in the insulator is given by

$$J = qn(x)\mu \left[-\frac{d\psi(x)/q}{dx} \right] - qD \frac{dn(x)}{dx} \quad (6-38)$$

Since μ is the electron mobility and assumed to be small and constant, we can use the Einstein relation

$$\frac{D}{\mu} = \frac{kT}{q} \quad (6-39)$$

Thus, Equation 6-38 can be written as

$$J = -n(x)\mu \frac{d\psi(x)}{dx} - kT\mu \frac{dn(x)}{dx} \quad (6-40)$$

The first term on the right side of Equation 6-40 represents the drift current from the metal to the insulator; the second term represents the diffusion current from the insulator to the metal. By solving Equation 6-40 for $n(x)$, O'Dwyer²⁹ has derived the following relation for J as a function of F :

$$J = q\mu N_c (kT/\pi)^{1/2} [4\epsilon(F/q)^3]^{1/4} \times \exp[-(\phi_B - \Delta\phi_B)/kT] \quad (6-41)$$

where N_c is the effective density of states in the conduction band, which is given by

$$N_c = 2 \left(\frac{2\pi m^* kT}{h^2} \right)^{3/2} \quad (6-42)$$

The ratio of Equation 6-41 to Equation 6-34 gives the ratio of the Richardson constant A_2^* , including the diffusion effect, to that of A without the diffusion effect

$$\frac{A_2^*}{A_1^*} = u(2m^*)^{1/2} [4\epsilon(F/q)^3]^{1/4} \quad (6-43)$$

If this ratio is less than unity, the current will be diffusion controlled. This condition will occur if the electron mobilities in the insulator follow

$$\mu < 5F^{3/4} \quad (6-44)$$

in which μ is in $\text{cm}^2 \text{V}^{-1} \text{s}^{-1}$ and F in MV cm^{-1} .²⁹

There is no incontrovertible experimental evidence of diffusion-limited thermionic emission. Experimental data on A_2^* are usually much smaller than what is given by Equation 6-35 or Equation 6-36. There are many factors that may cause this discrepancy. It is likely that thermionic emission is filamentary, because the cathode surface is not microscopically identical in asperity and surface conditions, and the insu-

lator itself is never microscopically uniform. Also, the space charge effect may also play a role in this discrepancy.

Blocking Contacts

Referring to Figure 6-13, the electron density in the region between $x = x_m$ and $x = W$ is given by³⁰

$$n(x) = N_c \exp\{-[\psi(x) - E_{Fn}]/kT\} \quad (6-45)$$

and the current density by

$$\begin{aligned} J &= -n(x)\mu \frac{d\psi(x)}{dx} - kT\mu \frac{dn(x)}{dx} \\ &= -n(x)\mu \frac{dE_{Fn}}{dx} \end{aligned} \quad (6-46)$$

where $\psi(x)$ and E_{Fn} are measured from the Fermi level E_{Fm} , as shown in Figure 6-13. The quasi-Fermi level E_{Fn} is sometimes called *imref*—Fermi spelled backward—to distinguish it from the Fermi level for equilibrium conditions. It is used for describing carrier distribution under nonequilibrium conditions, such as under an applied field. Equations 6-45 and 6-46 are valid only for $x > x_m$. For $0 < x < x_m$, the density of electrons cannot be described by E_{Fn} or be associated with N_c because of the rapid change of the potential energy in the distance comparable to the electron mean free path. Crowell and Sze³⁰ have assumed that the barrier in the region $0 < x < x_m$ acts as a sink for electrons, then the current flow in this region is

$$J = q(N_o - N_m)v_R \quad (6-47)$$

where v_R is the effective recombination velocity, N_o is the quasi-equilibrium electron density at x_m

$$N_o = N_c e^{-\phi_B}/kT \quad (6-48)$$

and N_m is the electron density at x_m when the current is flowing

$$N_m = N_c \exp\{-[\phi_B - E_{Fn}(x_m)]/kT\} \quad (6-49)$$

Using the boundary condition

$$E_{Fn}(x = w) = -qV \quad (6-50)$$

and from Equations 6-45 through 6-49, we obtain

$$\begin{aligned}
J &= \frac{qN_c v_R}{1 + \frac{v_R}{v_D}} \exp\left[-\frac{\phi_B - \Delta\phi_B}{kT}\right] \\
&\quad \times [1 - \exp(-qV/kT)] \\
&\simeq \frac{qN_c v_R}{1 + \frac{v_R}{v_D}} \exp\left[-\frac{\phi_B - \Delta\phi_B}{kT}\right] \\
&= A_2^* T^2 \exp\left(-\frac{\phi_B - \Delta\phi_B}{kT}\right)
\end{aligned} \quad (6-51)$$

for large V , where

$$v_D = \left[\int_{x_m}^W \frac{q}{\mu kT} \exp\left(-\frac{\phi_B - \psi}{kT}\right) dx \right]^{-1} \quad (6-52)$$

is the effective diffusion velocity associated with the diffusion of electrons from $x = W$ to $x = x_m$. Thus,

$$A_2^* = \frac{qN_c v_R v_D}{(v_D + v_R) T^2} \quad (6-53)$$

and the correction factor is

$$\begin{aligned}
\frac{A_2^*}{A_1^*} &= \frac{qN_c v_R v_D}{(v_D + v_R) T^2 A_1^*} \\
&= \left(\frac{2\pi m^*}{kT} \right)^{1/2} \left(\frac{v_D v_R}{v_D + v_R} \right)
\end{aligned} \quad (6-54)$$

In general, if $v_D \gg v_R$, the effect of diffusion is not important, but if $v_R \gg v_D$, the diffusion process is dominant.

Effects of Phonon Scattering and Quantum-Mechanical Reflection

When an electron crosses the peak of the potential barrier, there is a probability that it will be back-scattered by the scattering between the electron and the optical phonon. This effect will reduce the net current over the barrier. Crowell and Sze³¹ have proposed that this effect can be viewed as a small perturbation, and that, neglecting the scattering by acoustic phonons, the probability of electron emission over the peak of the potential barrier is given by³¹

$$f_p \simeq \exp\left(-\frac{x_m}{\lambda}\right) \quad (6-55)$$

where x_m is given by Equation 6-24 and λ is the optical phonon mean free path, which is

$$\lambda = \lambda_o \tanh\left(\frac{E_p}{2kT}\right) \quad (6-56)$$

where E_p is the optical phonon energy and λ_o is the high-energy, low-temperature asymptotic value of the phonon mean free path. They have also suggested that the effect of optical phonon scattering due to f_p can be taken into account by replacing v_R with a smaller recombination velocity $f_p v_R$ in Equations 6-51, 6-53, and 6-54 because $f_p < 1$.

Over the Schottky barrier there is quantum-mechanical reflection of electrons, and below the peak of the barrier there is tunneling of electrons through the thinner part of the barrier. Crowell and Sze³¹ have calculated the ratio f_q of the total current flow, taking into account the effects of electron tunneling and quantum-mechanical reflection, to the current flow neglecting these effects, as a function of electron energy:

$$f_q = \int_{-\infty}^{\infty} \frac{D_q}{kT} \exp(-E/kT) dE \quad (6-57)$$

where D_q is the predicted quantum-mechanical transmission coefficient, and E is the electron energy related to the barrier height and hence the electric field. Therefore, the effective recombination velocity is $f_p f_q v_R$.

Taking the two effects into account, Equation 6-53 must be modified as follows:

$$A_2^{**} = \frac{qN_c f_p f_q v_R v_D}{(v_D + f_p f_q v_R) T^2} \quad (6-58)$$

Other Factors

Apart from the effect of contamination at the interface between the metal and the semiconductor (or insulator), the following factors may also be responsible for the discrepancy between the theoretically expected and the experimental value of A^* :

- The space charge exists in the vicinity of the contact.
- The emitting area is much less than the actual electrode surface area because of filamentary injection.^{32,33}

- The value of the dielectric constant ϵ is less than the DC or static value of ϵ in the vicinity of the contact.^{34–36}
- The actual shape of the potential barrier is different from the ideal ones shown in Figures 6-12 and 6-13 because of imperfections such as surface states.
- The implicit assumption that there is no appreciable electron–electron interaction; thus, no interaction terms are included when the Fermi–Dirac distribution is used to derive the J – V characteristics. This assumption may be valid for low currents but not for high currents.

So far, we have discussed the factors that may affect the value of A^* . In general, if $\phi_B \gg kT$ or if the electron mean free path in the semiconductor or insulator is large compared to the distance over which the potential energy changes by kT near the top of the barrier, the effects of diffusion and electron collision are unimportant and can be neglected.^{37,38} If these effects are unimportant, the thermionic emitted current $\ell n J$ should be linearly proportional to $V^{1/2}$ at a fixed temperature, and $\ell n \frac{J}{T}$ should be

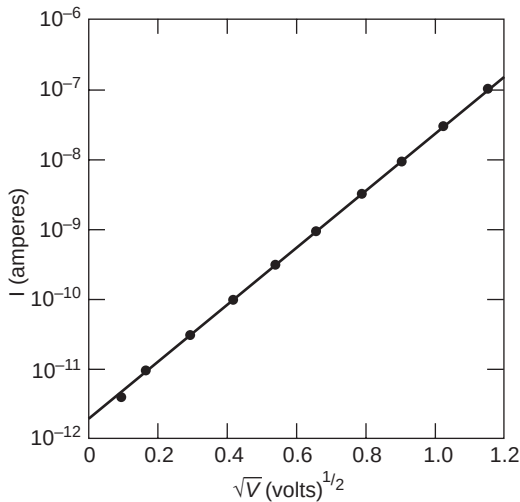


Figure 6-15 The current–voltage (I – V) characteristics of a Zn–ZnO–Au system due to Schottky thermionic emission.

linearly proportional to $\frac{1}{T}$ at a fixed applied voltage. Figure 6-15, with the results from Mead,³⁹ and Figure 6-16, with the results from Pollack,⁴⁰ are typical examples of thermionic emission to an insulator. In Figure 6-16, I is the total current and I_T is the temperature-independent portion possibly due to quantum-mechanical tunneling (field emission).

It should be noted that for semiconductors with fairly high carrier mobilities, such as Si, Ge, and GaAs, Schottky diodes made of these materials conform to thermionic emission theory; the J – V characteristics follow Equation 6-34 closely. For forward bias with $qV > 3kT$, Equation 6-34 may be reduced to

$$J = J_0 \exp(qV/kT) \quad (6-59)$$

However, this ideal characteristic is never observed in practice. The current is usually found to follow

$$J = J_0 \exp(qV/nkT) \quad (6-60)$$

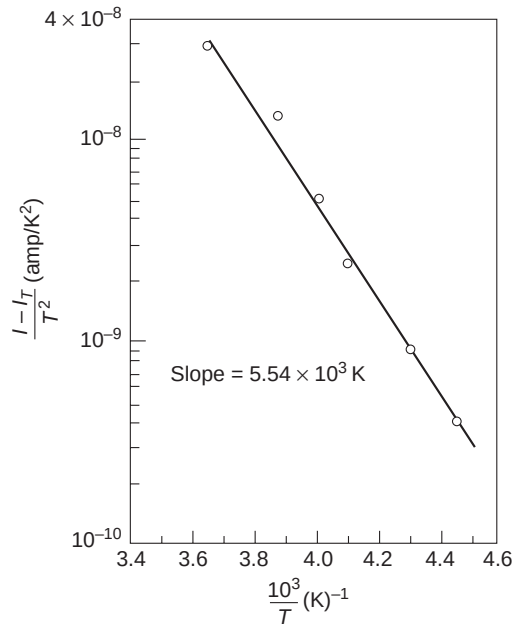


Figure 6-16 $(I - I_T)/T^2$ as a function of $1/T$ for a Pb–Al₂O₃–Pb system.

where n is generally referred to as the *ideality factor*, which is usually greater than unity. There are many factors governing the value of n , including the following:

- The recombination of electrons and holes in the depletion region has not been taken into account in the derivation of Equation 6-34. The recombination is important, particularly for materials with a large potential barrier height and a short lifetime, such as GaAs. Such recombination sometimes leads $n = 2$.
- The effective barrier height $\phi_{Bn} - \Delta\phi_{Bn}$ increases with increasing forward bias.
- With a thin insulating layer between metal and semiconductor, ϕ_B also increases with increasing forward bias.
- At high applied fields, the effects of drift and diffusion current in the barrier region are no longer negligible, implying that pure thermionic emission theory may become inadequate.
- The effect of series resistance at high applied bias cannot be ignored. The actual voltage across the junction should be $V - JR_s$ rather than V , where R_s is the series resistance.

6.2.3 Field Emission

Field emission is defined as the quantum-mechanical tunneling of electrons through a potential barrier from a metal to a semiconductor or an insulator under an intense electric field. At low temperatures, most electrons tunnel at the Fermi level of the metal, constituting field emission (F emission). At intermediate temperatures, most electrons tunnel at any energy level E_m (above the Fermi level of the metal), constituting the so-called thermionic-field or thermally assisted field emission (T - F emission). At very high temperatures, the main contribution is thermionic emission. In this section, we are concerned only with F emission. Since the first treatment of this problem by Fowler and Nordheim,⁴¹ many investigators have modified their treatment for emission into a semiconductor or an insulator rather than into a vacuum.⁴²⁻⁴⁵

Without the Effects of Defects in Solids

The field-emitted current density is given by

$$J = q \int D_T v_x n(E) dE \quad (6-61)$$

where D_T is the quantum mechanical transmission function of the transition probability, defined as the ratio of the transmitted to the incident current; v_x is the electron velocity in the x direction; and $n(E)$ the electron density with energies between E and $E + dE$

$$\begin{aligned} n(E)dE &= g(E)f(E)d(E) \\ &= \frac{8\pi m\sqrt{2mE}}{h^3} f(E)dE \\ &= \frac{4\pi p^2 dp}{h^3} f(E) \end{aligned}$$

Therefore, $n(E)dE$ within $dp_x dp_y dp_z$ is

$$\begin{aligned} n(E)dE &= \frac{4\pi p^2 dp}{h^3} f(E) \frac{dp_x dp_y dp_z}{4\pi p^2 dp} \\ &= \frac{2}{h^3} f(E) dp_x dp_y dp_z \end{aligned} \quad (6-62)$$

Thus, the net current flow from region 1 to region 2 (see Figure 6-17) is

$$J = \frac{2q}{h^3} \int D_T v_{x1} [f_1(E_2) - f_2(E_2)] dp_{x1} dp_{y1} dp_{z1} \quad (6-63)$$

where

$$f_1(E_1) = [\exp\{(E_1 - E_{F1})/kT\} + 1]^{-1} \quad (6-64)$$

$$\begin{aligned} f_2(E_2) &= [\exp\{(E_2 - E_{F2})/kT\} + 1]^{-1} \\ &= [\exp\{(E_1 + qV - E_{F1})/kT\} + 1]^{-1} \end{aligned} \quad (6-65)$$

$$v_{x1} = \frac{\partial \omega}{\partial k} = \frac{\partial E_{x1}}{\partial p_{x1}} \quad (6-66)$$

and

$$D_T = \exp\left\{-\frac{4\pi}{h} \int_{x_1}^{x_2} [2m(\psi_T - E_{x1})]^{1/2} dx\right\} \quad (6-67)$$

D_T is based on the WKB (Wentzel-Kramers-Brillouin) approximation. The condition for its validity is that the electron wavelength is small compared to the region over which appreciable variations in the potential energy of the electrons occurs.⁴⁶ E_{x1} is the portion of the electron

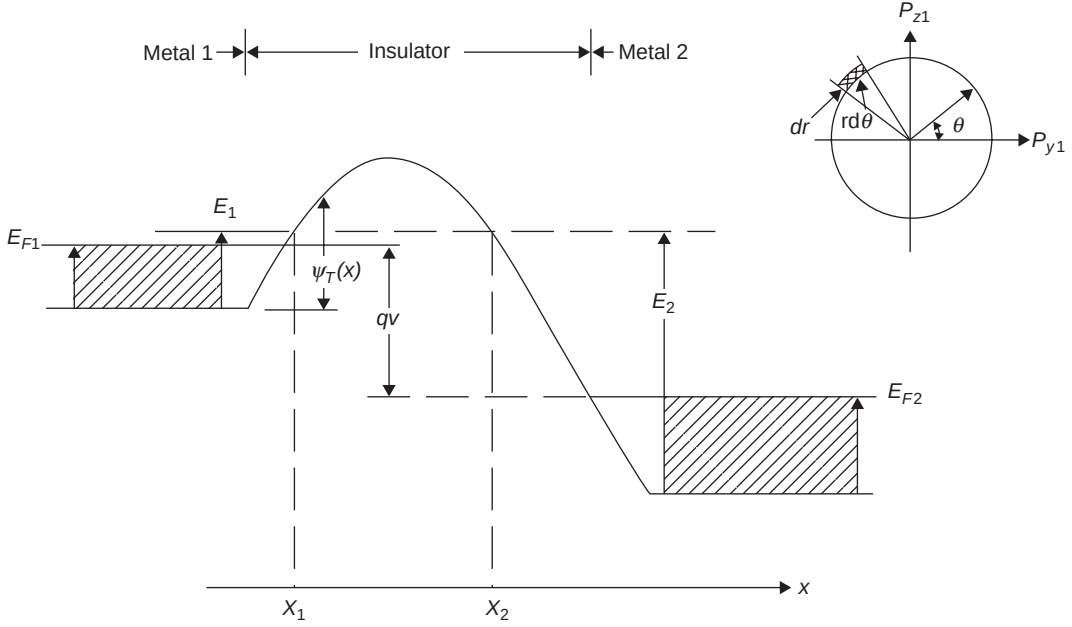


Figure 6-17 Energy level diagram for tunneling analysis. The insulator region may represent an insulating film or a barrier formed near the contact between a metal and an insulator (or a semiconductor).

energy E_1 for electron motion in the x direction. An applied voltage V to favor electron tunneling from region 1 to region 2 will lower E_{F2} , as shown in Figure 6-17.

Assuming that the current flow is dominant from region 1 to region 2 at E_1 , whose energy is close to E_{F1} , and letting

$$\alpha = \frac{4\pi(2m)^{1/2}}{h} \quad (6-68)$$

and

$$\epsilon_x = E_{F1} - E_{x1} \quad (6-69)$$

then Equation 6-67 may be expanded in the Taylor series

$$\begin{aligned} \ell n D_T(E_{x1}) &= -\alpha \int_{x_1}^{x_2} (\psi_T - E_{x1})^{1/2} dx \\ &= -\alpha \int_{x_1}^{x_2} (\psi_T - E_{F1} + \epsilon_x)^{1/2} dx \\ &= -[b_1 + c_1 \epsilon_x + f_1 \epsilon_x^2 + \dots] \end{aligned} \quad (6-70)$$

in which

$$b_1 = \alpha \int_{x_1}^{x_2} (\psi_T - E_{F1})^{1/2} dx \quad (6-71)$$

$$c_1 = \frac{1}{2} \alpha \int_{x_1}^{x_2} (\psi_T - E_{F1})^{-1/2} dx \quad (6-72)$$

and

$$\begin{aligned} f_1 &= \frac{1}{4} \alpha \left[\frac{1}{x_2 - x_1} \left\{ \frac{1}{\psi_T(x_1)} - \frac{1}{\psi_T(x_2)} \right\} \right. \\ &\quad \times \int_{x_1}^{x_2} \frac{dx}{(\psi_T(x) - E_{F1})^{1/2}} - \frac{1}{2} \int_{x_1}^{x_2} \frac{dx}{(\psi_T(x) - E_{F1})^{1/2}} \\ &\quad \times \left. \left(1 - \frac{\psi_T(x)}{x_2 - x_1} \left[\frac{x - x_1}{\psi_T(x_2)} + \frac{x_2 - x}{\psi_T(x_1)} \right] \right) \right] \end{aligned} \quad (6-73)$$

b_1 , c_1 , and f_1 are functions of applied voltage V through the voltage dependence of x_1 and x_2 . Substituting Equations 6-64 through 6-66 into Equation 6-63, we have

$$\begin{aligned} J &= \frac{2q}{h^3} \int_0^\infty [f_1(E_1) - f_2(E_2)] dE_{x1} \\ &\quad \times \int D_T(E_1, p_{y1}, p_{z1}) dp_{y1} dp_{z1} \end{aligned} \quad (6-74)$$

since

$$\begin{aligned} D_T(E_{x1}) &= D_T \left(E_1 - \frac{p_{y1}^2 + p_{z1}^2}{2m} \right) \\ &= D_T(E_1, p_{y1}, p_{z1}) \end{aligned} \quad (6-75)$$

and the conservation law

$$\left. \begin{aligned} p_{y1}^2 + p_{z1}^2 &= p_{y2}^2 + p_{z2}^2 = p_{\perp}^2 \\ E_1 &= E_2 = E \\ E_1 - E_{x1} &= E_{\perp} = \frac{p_{\perp}^2}{2m} \end{aligned} \right\} \quad (6-76)$$

We can write

$$-dE_{x1} = \frac{p_{\perp} dp_{\perp}}{m} = \frac{r dr}{m} \quad (6-77)$$

$$dp_{y1} dp_{z1} = r dr d\theta \quad (6-78)$$

with

$$r^2 = p_{y1}^2 + p_{z1}^2 \quad (6-79)$$

as shown in Figure 6-17. Thus, we obtain

$$J = \frac{4\pi m q}{h^3} \int_0^{\infty} [f_1(E_1) - f_2(E_2)] dE_{x1} \times \int_0^{E_{x1}} D_T(E_{x1}) dE_{x1} \quad (6-80)$$

Integration of Equation 6-80 by part yields

$$J = \frac{4\pi q m k T}{h^3} \int_0^{\infty} D_T(E_{x1}) \times \left(\frac{1 + \exp[(E_{F1} - E_{x1})/kT]}{1 + \exp[(E_{F1} - E_{x1} - qV)/kT]} \right) dE_{x1} \quad (6-81)$$

Substituting Equation 6-70 into Equation 6-81 and integrating it, we obtain

$$J = \frac{A^* T^2 \exp(-b_1)}{(c_1 k T)^2} \frac{\pi c_1 k T}{\sin(\pi c_1 k T)} \times [1 - \exp(-c_1 V)] \quad (6-82)$$

where $A^* = \frac{4\pi q k^2 m^*}{h^3}$, which is the Richardson constant (see Equation 6-35).

We consider only two terms in Equation 6-70; this is justified because

$$\frac{1}{kT} - c_1 > (2f_1)^{1/2} \quad (6-83)$$

in order to satisfy the condition for WKB approximation. This also implies that $c_1 k T < 1$ and that no singularity would be involved in Equation 6-82.

The temperature dependence of J can be easily found from Equation 6-82

$$\frac{J(T)}{J(T=0)} = \frac{\pi c_1 k T}{\sin(\pi c_1 k T)} \approx 1 + \frac{1}{6} (\pi c_1 k T)^2 + \dots \quad (6-84)$$

and $c_1 k T < 1$

Considering a Schottky barrier, shown in Figure 6-18(b), and neglecting the effect of the image force for simplicity, we have for low temperatures

$$\left. \begin{aligned} x_1 &= 0 \\ x_2 &= \frac{\epsilon \phi_B}{q^2 N_d W} \\ \psi_T(x) &= E_{Fm} + \phi_B - \frac{q^2 N_d}{\epsilon} \left(Wx - \frac{1}{2} x^2 \right) \\ E_{F1} &= E_{Fm} \end{aligned} \right\} \quad (6-85)$$

Using these parameters, integration of Equations 6-71 and 6-72 gives

$$b_1 = \frac{1}{E_{\infty}} \left\{ \phi_B^{1/2} (\phi_B + qV)^{1/2} + qV \ell n \left[\frac{(\phi_B + qV)^{1/2} + \phi_B^{1/2}}{(qV)^{1/2}} \right] \right\} \quad (6-86)$$

and

$$c_1 = \frac{1}{E_{\infty}} \ell n \left[\frac{(\phi_B + qV)^{1/2} + \phi_B^{1/2}}{(qV)^{1/2}} \right] \quad (6-87)$$

where

$$E_{\infty} = \frac{2q}{\alpha} \left(\frac{N_d}{2\epsilon} \right)^{1/2} \quad (6-88)$$

If $|qV| > \phi_B$, then Equations 6-86 and 6-87 become

$$b_1 = 2\phi_B^{3/2} / 3E_{\infty} (\phi_B + qV)^{1/2} \quad (6-89)$$

$$c_1 = \phi_B^{3/2} / E_{\infty} (\phi_B + qV)^{1/2} \quad (6-90)$$

Equation 6-82 becomes

$$J = \frac{A^* T^2 \pi E_{\infty} \exp[-2\phi_B^{3/2} / 3E_{\infty} (\phi_B + qV)^{1/2}]}{\left(\frac{kT[\phi_B / (\phi_B + qV)^{1/2} \times \sin\{\pi k T [\phi_B / (\phi_B + qV)]^{1/2} / E_{\infty}\}]}{1} \right)} \quad (6-91)$$

For low temperatures, Equation 6-91 reduces to

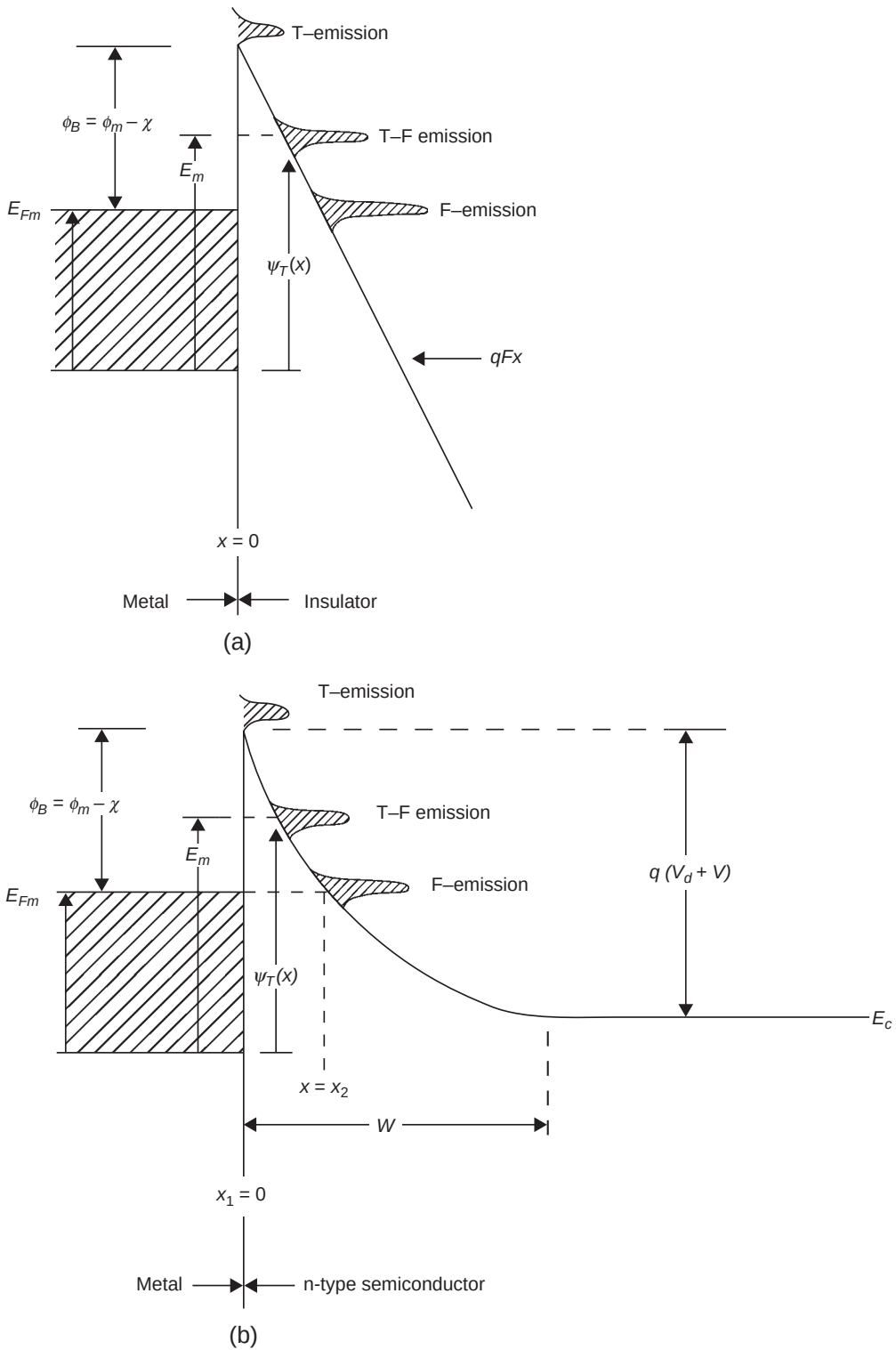


Figure 6-18 Schematic energy level diagrams showing thermionic (T) emission, thermionic–field (T – F) emission, and field (F) emission for (a) neutral contacts and (b) blocking contacts.

$$J = A^* T^2 \left(\frac{E_\infty}{kT} \right)^2 \frac{(\phi_B + qV)}{\phi_B} \times \exp \left[-\frac{2\phi_B^{3/2}}{3E_\infty(\phi_B + qV)^{1/2}} \right] \quad (6-92)$$

Considering a triangular barrier, as shown in Figure 6-18(a), we have

$$\left. \begin{aligned} x_1 &= 0 \\ x_2 &= \frac{\phi_B}{qF} \\ \psi_T(x) &= E_{Fm} + \phi_B - qFx \\ E_{F1} &= E_{Fm} \end{aligned} \right\} \quad (6-93)$$

Using these parameters, integration of Equation 6-71 and 6-72 gives

$$b_1 = \alpha \frac{2\phi_B^{3/2}}{3qF} \quad (6-94)$$

$$c_1 = \alpha \frac{\phi_B^{1/2}}{qF} \quad (6-95)$$

Thus, the field emission current is

$$J = \frac{A^* T^2 \pi \exp(-2\alpha\phi_B^{3/2}/3qF)}{(\alpha\phi_B^{1/2}kT/qF) \sin(\pi\alpha\phi_B^{1/2}kT/qF)} \quad (6-96)$$

For low temperatures

$$J = \frac{A^* T^2}{\phi_B} \left(\frac{qF}{\alpha kT} \right)^2 \exp \left[-\frac{2\alpha\phi_B^{3/2}}{3qF} \right] \quad (6-97)$$

Equations 6-92 and 6-97 are essentially equivalent to the Fowler–Nordheim equation, since for the Schottky barrier the field at the interface is proportional to the square root of the effective barrier height. It is most likely that the field emission is filamentary^{32,33} and the current is space-charge limited under high fields. If so, all expressions for field emission must be modified. However, the experimental results showing a linear relation between $\ln J/F^2$ and $1/F$, such as those shown in Figure 6-19, may be taken as an indication of field emission. The experimental results are from Lenzlinger and Snow.⁴⁷

With the Effects of Defects in Solids

Obviously, the effects of image force and carrier traps greatly affect the efficiency of field emission. If these effects are taken into account,

the integrals for b_1 and c_1 may be written in terms of complete elliptic integrals of the first and second kinds. Several investigators have attempted to solve this problem.^{29,48,49} The analysis is quite mathematically involved, and such a detailed treatment is beyond the scope of this chapter. However, we would like to describe briefly the effects of the image force and carrier traps.

Electrons injected from the electron-injecting contact to the insulator specimen are partly captured by traps, forming a negatively trapped space charge (i.e., a homo-space charge) near the injecting contact. The rate of electron trapping may be expressed as

$$\frac{dn_t(t)}{dt} = \frac{\sigma J(t)}{q} [N_t - n_t(t)] \quad (6-98)$$

where $n_t(t)$ and N_t are, respectively, the concentrations of the filled and the total traps, σ is the effective capture cross-section of the traps, and $J(t)$ is the injected current density.⁵⁰ If $J(t)$ is due mainly to the Fowler–Nordheim (FN) type of tunneling injection, then $J(t)$ can be written as

$$J(t) = \frac{q^2 m_o [F_c(t)]^2}{16\pi^2 h m^* [q\phi_B(t)]} \times \exp \left\{ \frac{4(2m^*)^{1/2} [q\phi_B(t)]^{3/2}}{3hqF_c(t)} \right\} \quad (6-99)$$

where h is the Planck constant, m^* and m_o are, respectively, the effective mass and the rest mass of the electrons, $\phi_B(t)$ is the effective potential barrier height in eV at the electron-injecting contact, and $F_c(t)$ is the effective field at the electron-injecting contact.⁴⁹ The apparent barrier height at the metal–insulator contact is given by

$$\phi_{BP} = \phi_m - \chi \quad (6-100)$$

where ϕ_m is the work function of the metal electrode and χ is the electron affinity of the insulator. In the tunneling region, the potential barrier is not triangular if the image force is taken into account. The image force lowering is given by

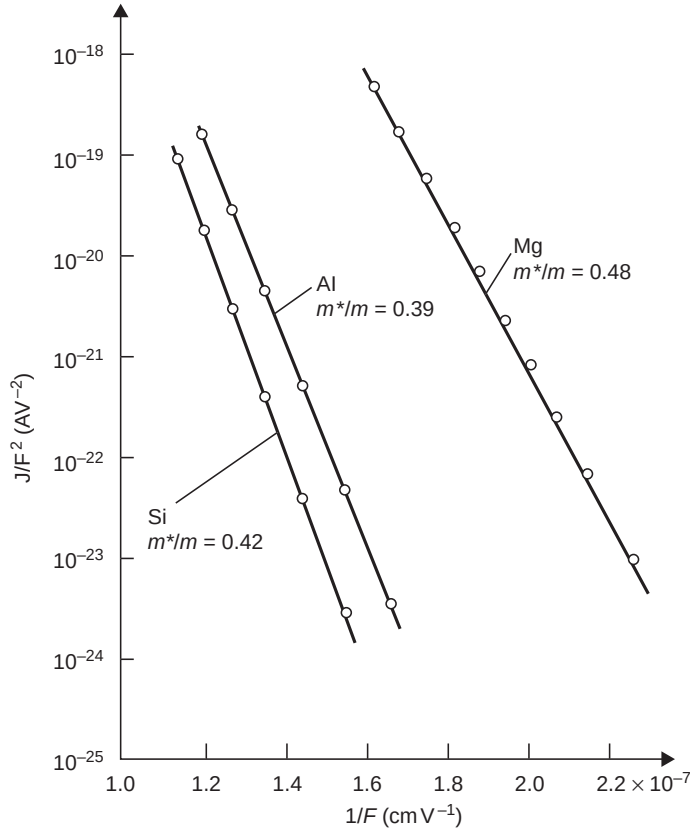


Figure 6-19 Fowler-Nordheim plot of current density as a function of electric field for field emission from various metal electrodes into SiO_2 at room temperature. m^*/m is the ratio of electron effective mass to rest mass.

$$\Delta\phi(t) = \left(\frac{q^3 F_c(t)}{4\pi\epsilon} \right)^{1/2} \quad (6-101)$$

which is similar to Equation 6-25.⁴⁹ Thus, the effective barrier height for the metal-insulator contact, as shown in Figure 6-20(a) or Figure 6-21(a), can be expressed as

$$\phi_B(t) = \phi_{BP} - \Delta\phi(t) = \phi_m - \chi - \Delta\phi(t) \quad (6-102)$$

The electric field at the electron-injecting contact is lowered by the internal field $F_i(t)$ in the opposite direction to the applied field. Thus, $F_c(t)$ for an MIM system, as shown in Figure 6-20(a), can be written as

$$F_c(t) = \frac{V}{d} - F_i(t) \quad (6-103)$$

If the carrier injection is from the semiconductor-insulator contact (i.e., a p-type

silicon-insulator contact for the present case) in an MIS system, as shown in Figure 6-21(b), the potential barrier height is given by⁵¹

$$\begin{aligned} \phi_B(t) &= \phi_{BP} - \Delta\phi(t) \\ &= \phi_s - \psi_s(t) - \chi - \Delta\phi(t) \end{aligned} \quad (6-104)$$

and $F_c(t)$ by

$$F_c(t) = \frac{V(t) + \phi_{ms} + \psi_s(t)}{d} - F_i \quad (6-105)$$

where ϕ_s is the work function of silicon, $\psi_s(t)$ is the surface potential at the silicon-insulator interface, and ϕ_{ms} is the work function difference equal to $\phi_m - \phi_s$. The internal field F_i at the injecting contact created by the trapped electron space charge is given by

$$F_i = \frac{Q_t}{\epsilon} \left(1 - \frac{x_o}{d} \right) \quad (6-106)$$

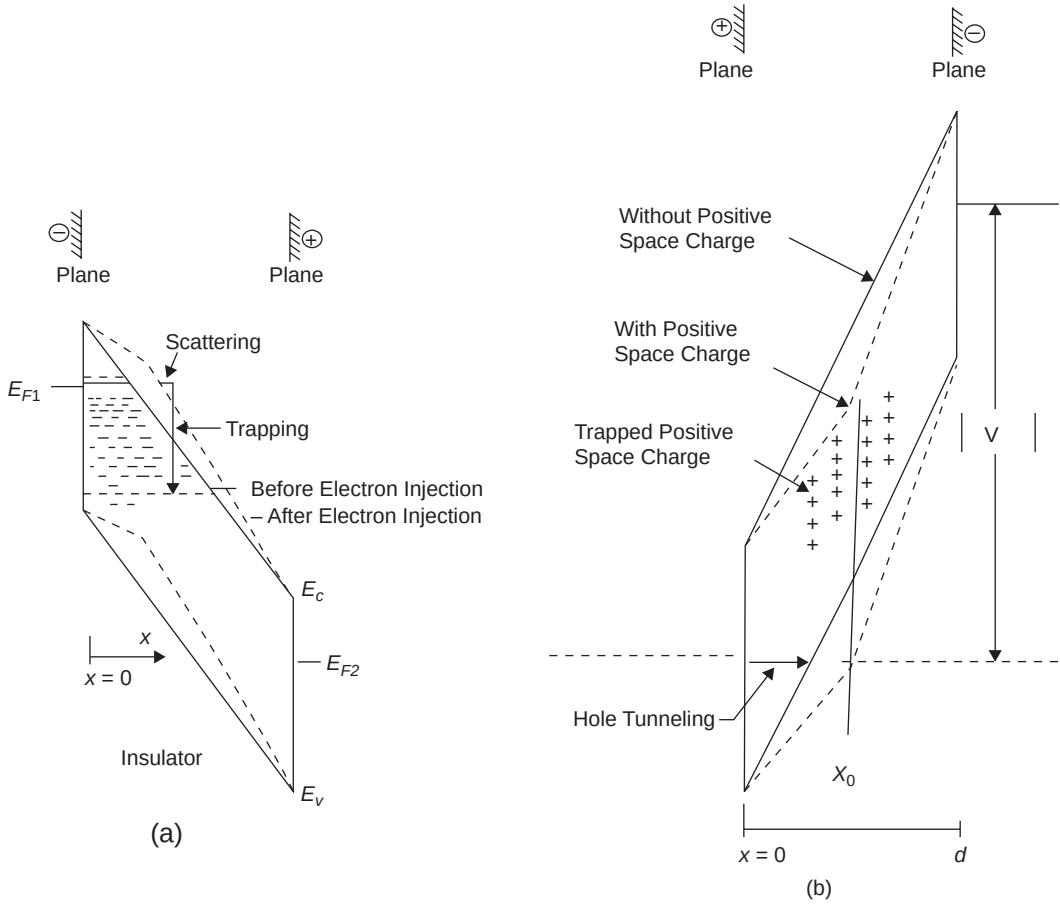


Figure 6-20 Schematic diagrams for an MIM system illustrating (a) an electron tunneling injection, electron trapping, and build-up of trapped electron space charge modifying the potential distribution, and (b) hole tunneling injection, hole trapping, and build-up of trapped hole space charge modifying the potential distribution. (The image force effect is not shown for clarity.)

where x_0 is the location of the centroid of the trapped electron space charge measured from the silicon–insulator interface (i.e., the electron-injecting contact), as shown in Figure 6-21(c), and Q_t is the total trapped charge in the specimen.^{52,53} Using the boundary condition $n_t(t) = 0$ when $t = 0$, the solution of Equation 6-98 gives

$$n_t = N_t \left[1 - \exp \left(-\frac{\sigma}{q} \int_0^t J dt \right) \right] \quad (6-107)$$

and the total trapped charge can be expressed as

$$Q_t = q \int_0^t n_t dx = q N_t d \left[1 - \exp \left(-\frac{\sigma}{q} \int_0^t J dt \right) \right] \quad (6-108)$$

The first term of Equation 6-103 or Equation 6-105 is the average applied field F . Thus, the rate of the change of the effective field at the injecting contact is

$$\frac{dF_c}{dt} = \frac{dF}{dt} - \frac{dF_i}{dt} \quad (6-109)$$

If we use linear ramp voltages (i.e., if the applied voltage increases linearly with time) with the ramp rate r_g for the measurement of the high-field J – F characteristics, then Equation 6-109 can be simplified to

$$\frac{dF_c}{dt} = r_g - r_i \quad (6-110)$$

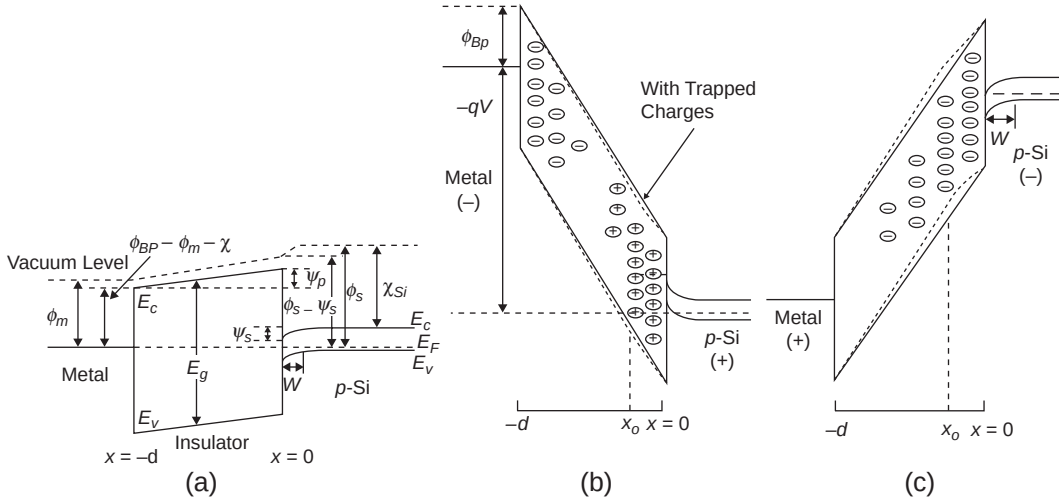


Figure 6-21 Schematic diagrams for an MIS system illustrating the energy band diagrams (a) in the absence of applied bias, (b) with the presence of trapped electrons and trapped holes for the metal gate electrode negatively biased, and (c) with the presence of mainly trapped electrons for the metal gate electrode positively biased. (The image force effect is not shown for clarity.)

where r_i is the rate of the change of the internal field at the injecting contact, which can be written as

$$\begin{aligned} r_i &= \frac{dF_i}{dt} = \frac{d}{dt} \left[\frac{Q_t}{\epsilon} \left(1 - \frac{x_o}{d} \right) \right] \\ &= \frac{1}{\epsilon} \left(1 - \frac{x_o}{d} \right) \frac{dQ_t}{dt} - \frac{Q_t}{\epsilon d} \frac{dx_o}{dt} \end{aligned} \quad (6-111)$$

Since the homo-space charge is located very near the injecting contact, $x_o < d$. Therefore, in most cases the first term of Equation 6-111 is larger than the second term. Thus, r_i is always positive. By differentiating Equation 6-108 and substituting it into Equation 6-111, we obtain

$$\begin{aligned} r_i &= \frac{dF_i}{dt} = \frac{qN_c}{\epsilon} \left\{ \left[\frac{\sigma J}{q} (d - x_o) + \frac{dx_o}{dt} \right] \right. \\ &\quad \times \exp \left(-\frac{\sigma}{q} \int_0^t J dt \right) - \frac{dx_o}{dt} \left. \right\} \end{aligned} \quad (6-112)$$

From Equations 6-110 and 6-112, we can have three possible J - F characteristics, as shown in Figure 6-22:

1. $r_g > r_i$: The injection current J always increases with increasing F .
2. $r_g = r_i$: The injection current J tends to become saturated. Since r_i is a function of J , if r_i does not change, J would not

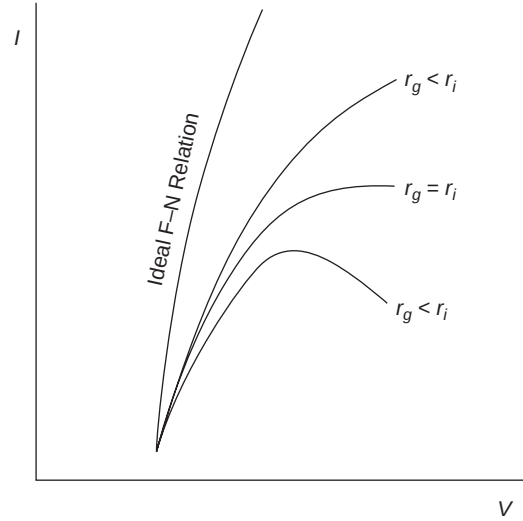


Figure 6-22 The ramp I-V characteristics for three possible cases— $r_g > r_i$, $r_g = r_i$, and $r_g < r_i$ —and showing the effects of carrier traps on the Fowler-Nordheim tunneling injection.

change, thus forming a ledge in the J - F characteristics within a certain range of F .

3. $r_g < r_i$: The injection current J , after reaching its peak value, may decrease with increasing F , forming a negative differential resistance region. In this case, N_c , σ , or

both are large, and x_o may move toward the injecting contact as t is increased.

It can be seen from Equation 6-112 that r_i is controlled by many parameters. Depending on the carrier mobility, the capture cross-section of the traps, and the spatial distribution of the trap concentration, the first group of injected carriers moving away from the injecting contact will be trapped somewhere, forming a first sheet of space charge, which then tends to retard the oncoming subsequent groups of injected carriers, gradually enhancing their interaction with bulk traps. In other words, the bulk traps are gradually filled, starting from the first sheet of trapped carrier space charge toward the injecting contact. It is likely that the centroid of the trapped charge x_o is not constant but moves toward the injecting contact as the applied voltage is increased. For linear ramp voltages, dx_o/dt should always be negative. This makes r_i increase with increasing injection current level.

Tu and Kao⁵⁴ have computed the I - V characteristics for electron injection from p -Si to polyimide (PI) films based on Equations 6-99 through 6-108, using Maple's computer program in conjunction with the fourth-order Runge-Kutta numerical method⁵⁵ and the following physical parameters for p -Si/polyimide contact:

Electron trap capture cross section: $\sigma = 10^{-16} \text{ cm}^2$

Relative permittivity of PI: $\varepsilon / \varepsilon_o = 3.4$ with $\varepsilon_o = 8.85 \times 10^{-14} \text{ F cm}^{-1}$

Apparent barrier height at the injecting contact: $\phi_{BP} = 3.8 \text{ eV}$

Thickness of the PI film: $d = 6500 \text{ \AA}$

Area of the injecting contact: $A = 1.77 \times 10^{-2} \text{ cm}^2$

Ramp rate of the applied voltage: $g = 1.7 \text{ Vs}^{-1}$ (or $0.026 \text{ MV cm}^{-1} \text{ s}^{-1}$)

For this computation, we chose the p -Si/PI contact as the electron-injecting contact and assumed, for simplicity, $x_o = 0$, implying that the centroid of the trapped electron concentration is located near or at the electron-injecting p -Si/PI contact. We computed the I - V characteristics

for various total trap concentrations. Figure 6-23 shows these characteristics for $N_t = 7.7 \times 10^{17} \text{ cm}^{-3}$, $6.2 \times 10^{17} \text{ cm}^{-3}$, $4.6 \times 10^{17} \text{ cm}^{-3}$, and $3.0 \times 10^{17} \text{ cm}^{-3}$. It can be seen that for the occurrence of a ledge in the I - V characteristics, the trapped electron concentration must be large enough to make the rate of increase of the internal field approach the ramp rate of the applied field.

Typical experimental I - V characteristics for this MIS system are shown in Figure 6-24. Region I represents the displacement current region. This displacement current, which is due to $C(dV/dt)$, where C is the total capacitance of the system, is constant up to the threshold field for electron injection. In Region II, the current in the first ramp cycle is much larger than that in the second and the third cycles at the same electric field, indicating clearly the buildup of a negative space charge. Furthermore, for the first ramp cycle, the threshold field is about 1.4 MV cm^{-1} . During the first cycle, most traps may have been filled; after the first cycle, the same trapped electrons may remain trapped, creating an internal field opposite to the applied field. Thus, a higher threshold field is required for the Fowler-Nordheim (FN) tunneling injection. This is why the threshold field becomes

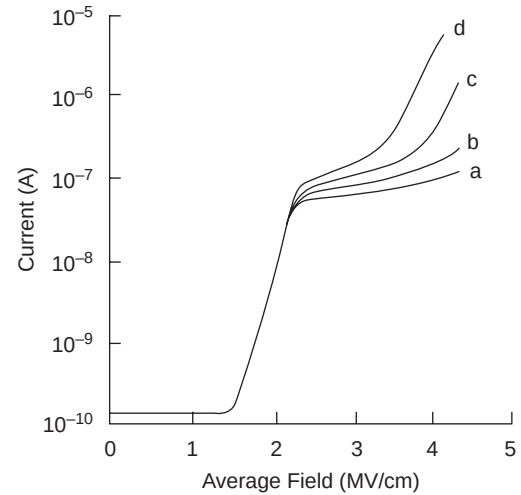


Figure 6-23 Computed I - V characteristics ($F = Vd^{-1}$) with the p -Si/PI contact as the electron-injecting contact for $x_o = 0$ and (a) $N_t = 7.7 \times 10^{17}$, (b) $N_t = 6.2 \times 10^{17}$, (c) $N_t = 4.6 \times 10^{17}$, and (d) $N_t = 3.0 \times 10^{17} \text{ cm}^{-3}$.

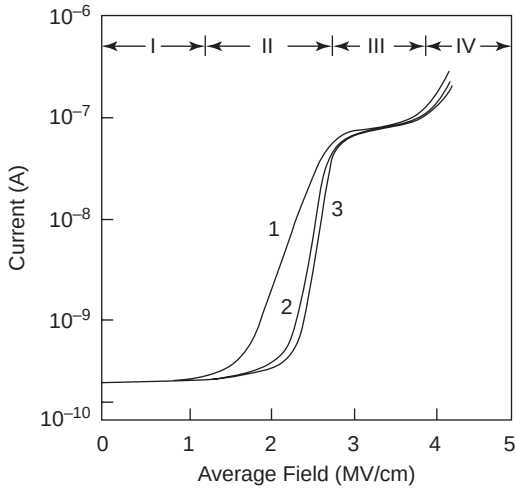


Figure 6-24 Typical experimental I - V characteristics of the metal-polymide-p-Si (MIS) system with the gate positively biased and the ramp rate of $0.026 \text{ MV cm}^{-1} \text{ s}^{-1}$. 1, 2, and 3 denote the first, second, and third ramp cycle, respectively.

2.0 MV cm^{-1} for the second cycle and even higher for the third, as shown in Figure 6-24. This also implies that the net negatively trapped charge increases and its centroid tends to move closer to the electron-injecting contact after each stressing cycle.

In Region III, the ledge current appears almost at the same location for the first, second, and third ramp cycles, indicating that the detrapping may become important at high fields, possibly due to the Poole-Frenkel type of detrapping. In Region IV, we can consider that all traps have been filled and that hole injection may also start from the metal electrode, resulting in double injection, which leads to a rapid increase in carrier multiplication and final destructive breakdown. The general shape of the experimental I - V characteristics is very similar to that of the theoretical characteristics shown in Figure 6-23. By comparing Figures 6-23 and 6-24, the total electron trap concentration in the PI films may be of the order of 10^{17} cm^{-3} .

6.2.4 Thermionic-Field Emission

In the intermediate temperature range (when $c_1 kT > 1$), most electrons tunnel at an energy

level E_m , which is lower than $\phi_B + E_{Fm}$ but higher than E_{Fm} , as shown in Figure 6-18. For a Schottky barrier, Stratton⁵⁶ has derived an expression for J as a function of V using a procedure similar to that given in Section 6.2.3. The thermionic-field emission current density is given by⁴⁵

$$J = J_s \exp(qV/E') \quad (6-113)$$

where

$$J_s = A^* T^2 \left(\frac{\pi E_\infty}{k^2 T^2} \right)^{1/2} \left[qV + \frac{\phi_B}{\cosh^2(E_\infty/kT)} \right]^{1/2} \times \exp\left(-\frac{\phi_B}{E_o}\right) \quad (6-114)$$

$$E' = E_\infty [E_\infty/kT - \tanh(E_\infty/kT)]^{-1} \quad (6-115)$$

and

$$E_o = E_\infty \coth(E_\infty/kT) \quad (6-116)$$

It can easily be shown that for T - F emission, the energy level E_m , which represents the peak of the energy distribution of emitted electrons, will be such that

$$c_1(E_m)kT = 1 \quad (6-117)$$

and that the energy distribution of emitted electrons is a Gaussian distribution with a half-width⁵⁷

$$\Delta = 2(\ell n 2)^{1/2} E_\infty^{1/2} [qV + \phi_B / \cosh^2(E_\infty/kT)]^{1/2} \quad (6-118)$$

where

$$c_1(E_m) = \frac{1}{E_\infty} \ell n \left[\frac{(\phi_B + qV)^{1/2} + (\phi_B + E_{Fm} - E_m)^{1/2}}{(E_m + qV - E_{Fm})^{1/2}} \right] \quad (6-119)$$

and

$$E_m = E_{Fm} + \frac{\phi_B - qV \sinh^2(E_\infty/kT)}{\cosh^2(E_\infty/kT)} \quad (6-120)$$

The above results are valid only in a certain temperature range, given by the two conditions

$$c_1 kT > 1 \quad (6-121)$$

$$D_T(E_m) < 1/e$$

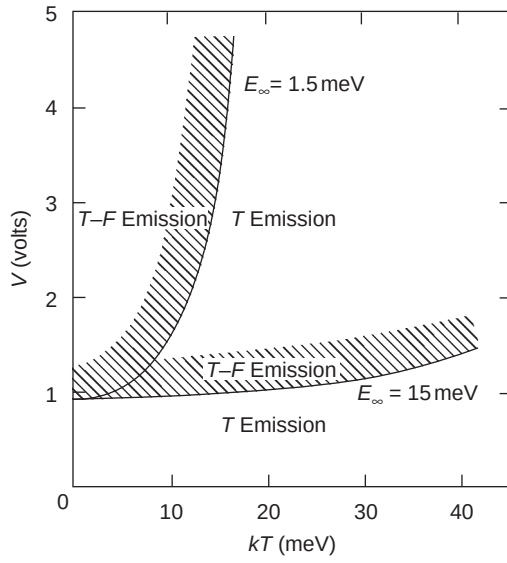


Figure 6-25 Minimum applied reverse bias (V) for thermionic field (T - F) emission as a function of temperature for Au-GaAs Schottky barrier for two values of the parameter E_{∞} . T - F and T denote, respectively, thermionic field emission and thermionic emission.

This implies that $E_m < E_{Fm} + \phi_B$. When $E_m \geq E_{Fm} + \phi_B$, the thermionic-field emission will change to thermionic emission. The minimum voltage to be applied for thermionic-field emission is

$$qV > \phi_B + \frac{3E_{\infty}}{2} \frac{\cosh^2(E_{\infty}/kT)}{\sinh^2(E_{\infty}/kT)} \quad (6-122)$$

The plot of Equation 6-122 for a typical Schottky barrier—a gold-gallium-arsenide barrier—of barrier height of 0.95 eV is shown in Figure 6-25. The typical energy distribution of the density of electrons emitted from the metal (Au) into the semiconductor (GaAs) as a function of applied voltage for the temperature 140°C is shown in Figure 6-26. The results are from Padovani and Stratton.⁴⁵

6.3 Tunneling through Thin Dielectric Films between Electrical Contacts

If an insulator is sufficiently thin or contains a large number of imperfections, or both, electrons may tunnel directly from one electrode to

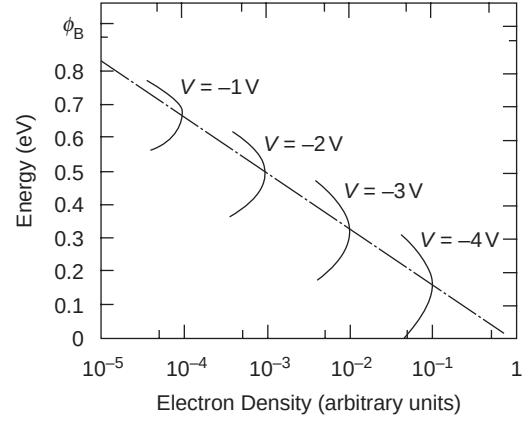


Figure 6-26 Energy distribution of the density of the thermionic field-emitted electrons from the Au electrode to GaAs at 140°C as a function of applied reverse bias (V).

the other and constitute a measurable current without involving the movement of carriers in the conduction band or in the valence band of the insulator. The tunneling current-voltage characteristics depend on the thin film fabrication procedures and, particularly, on the properties of the interfaces between the insulating film and the electrodes. In addition, these characteristics depend on the intrinsic properties of the insulator and the work functions of the metallic electrodes. It is most likely that the tunneling is filamentary and the current density is not uniform over the electrode surface, even with a planar symmetrical electrode geometry. Since Frenkel⁵⁸ reported his approximate analysis of electron tunneling through a thin insulating film, several investigators have extended and elaborated on his work.^{48,56,59-62} For some excellent and comprehensive review articles, see references 9, 26, 42, 44, 61, and 62.

6.3.1 Analysis Based on a Generalized Potential Barrier

A great many features of tunneling phenomena in insulating films are essentially of a one-dimensional nature. If the potential barrier for tunneling extends in the x direction, the momentum components of the electrons in the y and z directions, which are normal to the

direction of the current flow, can be considered merely fixed parameters. Thus, the probability that an electron at the energy level E_x can penetrate a potential barrier of height $\psi_T(x)$ and of width $S_2 - S_1$ as shown in Figure 6-27 can be calculated by the well known WKB or WKBJ (Wentzel-Kramers-Brillouin-Jeffreys) approximation method. Using this method, it can be shown that the number of electrons tunneling per second from electrode 1 to electrode 2, as shown in Figure 6-27, is given by

$$N_1 = \frac{4\pi m}{h^3} \int_0^{E_m} D_T(E_x) dE_x \int_0^\infty f(E) dE_r \quad (6-123)$$

The number of electrons tunneling per second from electrode 2 to electrode 1 is given by

$$N_2 = \frac{4\pi m}{h^3} \int_0^{E_m} D_T(E_x) dE_x \int_0^\infty f(E + qV) dE_r \quad (6-124)$$

The current density due to the net flow of electrons from electrode 1 to electrode 2 through

the forbidden energy gap of the insulating film is given by

$$J = q(N_1 - N_2) = \int_0^{E_m} D_T(E_x) \xi dE_x \quad (6-125)$$

where E_m is the maximum energy of the electrons in the electrode, $D_T(E_x)$ is the probability that an electron at the energy level E_x can penetrate a potential barrier of height $\psi_T(x)$ and of width $S_2 - S_1$.⁶¹ This probability is derived on the basis of the WKBJ approximation and given by

$$D_T(E_x) = \exp\left(-\frac{4\pi}{h} \int_{S_1}^{S_2} \{2m[\psi_T(x) - E_x]\}^{1/2} dx\right) \quad (6-126)$$

ξ is generally called the supply function, and ξdE_x represents the difference between the number of electrons having energy in the range of E_x to $E_x + dE_x$ incident on one side per second

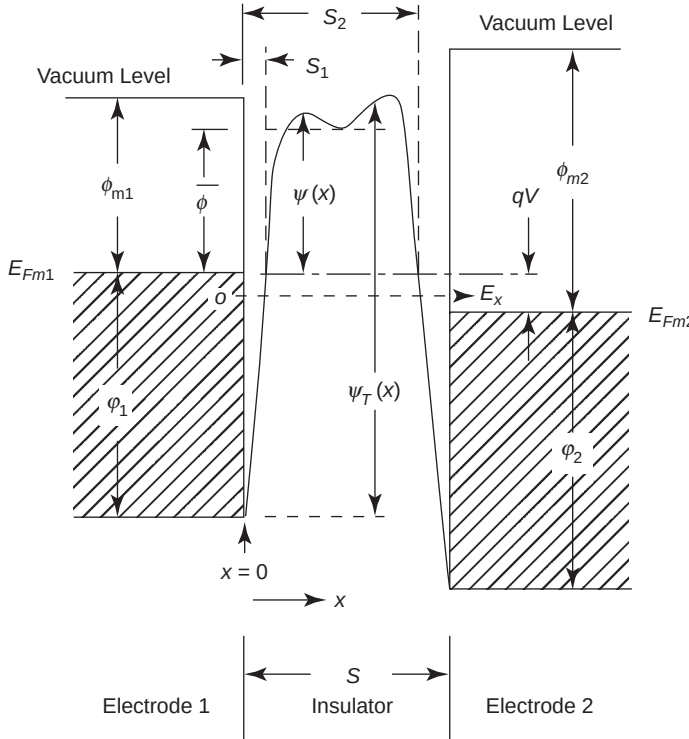


Figure 6-27 General potential barrier for a metal-insulating film-metal system. The width of the barrier for tunneling at the energy level E_x is $S_2 - S_1$.

per unit area, and those incident on the opposite side of the barrier. Thus, ξ is given by

$$\xi = \xi_1 - \xi_2 \quad (6-127)$$

$$\xi_1 = \frac{4\pi m q}{h^3} \int_0^\infty f(E) dE_r \quad (6-128)$$

$$\xi_2 = \frac{4\pi m q}{h^3} \int_0^\infty f(E + qV) dE_r \quad (6-129)$$

and

$$\left. \begin{aligned} E &= \frac{1}{2} m (v_x^2 + v_y^2 + v_z^2) \\ E_r &= \frac{1}{2} m (v_y^2 + v_z^2) \end{aligned} \right\} \quad (6-130)$$

For a generalized barrier, the barrier height can be written as

$$\psi_T(x) = \phi_1 + \psi(x) \quad (6-131)$$

Substituting Equation 6-131 into Equation 6-126 and simplifying it, we obtain

$$D_r(E_x) \approx \exp[-C(\phi_1 + \bar{\phi} - E_x)^{1/2}] \quad (6-132)$$

where

$$C = \frac{4\pi}{h} (2m)^{1/2} \beta [S_2 - S_1] \quad (6-133)$$

$$\bar{\phi} = \frac{1}{|S_2 - S_1|} \int_{S_1}^{S_2} \psi(x) dx \quad (6-134)$$

$$\beta = 1 - \frac{1}{8\bar{g}^2 |S_2 - S_1|} \int_{S_1}^{S_2} [g - \bar{g}]^2 dx \quad (6-135)$$

and

$$\left. \begin{aligned} \bar{g} &= \frac{1}{|S_2 - S_1|} \int_{S_1}^{S_2} g(x) dx \\ g &= [\psi_T(x) - E_x]^{1/2} \end{aligned} \right\} \quad (6-136)$$

For $T = 0$

In this case, Equations 6-127 through 6-129 become

$$\xi_1 = \frac{4\pi m q}{h^3} (\phi_1 - E_x) \quad (6-137)$$

$$\xi_2 = \frac{4\pi m q}{h^3} (\phi_1 - E_x - qV) \quad (6-138)$$

$$\left. \begin{aligned} \xi &= \xi_1 - \xi_2 = \frac{4\pi m q}{h} (qV), & 0 < E_x < \phi_1 - qV \\ &= \frac{4\pi m q}{h^3} (\phi_1 - E_x), & \phi_1 - qV < E_x < \phi_1 \\ &= 0 & E_x > \phi_1 \end{aligned} \right\} \quad (6-139)$$

Substitution of Equations 6-132 and 6-133 into Equation 6-125 yields

$$J \approx J_0 (\bar{\phi} \exp(-C\bar{\phi}^{1/2}) - (\bar{\phi} + qV) \exp[-C(\bar{\phi} + qV)^{1/2}]) \quad (6-140)$$

where

$$J_0 = \frac{q}{2\pi h} (\beta |S_2 - S_1|)^{-2} \quad (6-141)$$

Equation 6-140 can be applied to any shape of potential barrier, provided that the mean barrier height $\bar{\phi}$ is known. Of course, if the J - V characteristic is known, then $\bar{\phi}$ can be determined. The first term on the right side of Equation 6-140 can be interpreted as the current flow from electrode 1 to electrode 2, the second term as the current flow from electrode 2 to electrode 1. For low applied voltages, $qV \rightarrow 0$, $\beta \approx 1$ and $\bar{\phi} \gg qV$, Equation 6-140 becomes

$$J \approx J_0 qV [C\bar{\phi}^{1/2}/2 - 1] \exp(-C\bar{\phi}^{1/2}) \quad (6-142)$$

J is a linear function of V for very low voltages at $T = 0$ or very low temperatures.

For $T > 0$

In this case, Equations 6-128 and 6-129 become

$$\xi_1 = \frac{4\pi m q k T}{h^3} \ell n(1 + \exp[(E_{Fm1} - E_x)/kT]) \quad (6-143)$$

and

$$\xi_2 = \frac{4\pi m q k T}{h^3} \ell n(1 + \exp[(E_{Fm1} - E_x - qV/kT)]) \quad (6-144)$$

Assuming that the current flow is due predominantly to electrons with energies close to E_{Fm1} , the integral in Equation 6-126 can be expanded with respect to θ_x , which is

$$\theta_x = E_{Fm1} - E_x = \phi_1 - E_x \quad (6-145)$$

By carrying out a Taylor expansion, Equation 6-126 becomes

$$\ell n D_T(E_x) = -[b_1 + c_1 \theta_x + f_1 \theta_x^2 + \dots] \quad (6-146)$$

where

$$b_1 = \frac{4\pi(2m)^{1/2}}{h} \int_{s_1}^{s_2} [\phi_1 + \psi(x) - E_x]^{1/2} dx \quad (6-147)$$

$$c_1 = \frac{2\pi(2m)^{1/2}}{h} \int_{s_1}^{s_2} [\phi_1 + \psi(x) - E_x]^{-1/2} dx \quad (6-148)$$

which are functions of V through $\psi(x)$, which is a function of V . For most practical cases

$$1 - c_1 kT > kT(2f_1)^{1/2} \quad (6-149)$$

the term with the quadratic and high orders can be neglected.⁴⁸ Thus, applying Equation 6-125, we obtain

$$J = \frac{4\pi m q k T}{h^3} \int_0^{E_m} \exp(-[b_1 + c_1(\phi_1 - E_x)]) \times \ell n \left(\frac{1 + \exp[(\phi_1 - E_x)/kT]}{1 + \exp[(\phi_1 - E_x - qV)/kT]} \right) dE_x \quad (6-150)$$

After integration,⁵⁶ we have

$$J = \frac{4\pi m q}{h^3 c_1} \exp(-b_1) [1 - \exp(-c_1 V)] \frac{\pi c_1 k T}{\sin \pi c_1 k T} \quad (6-151)$$

Thus, at a given applied voltage, the ratio of the tunneling current through a thin insulating film for $T > 0$ to that for $T = 0$ is

$$\begin{aligned} \frac{J(T > 0)}{J(T = 0)} &= \frac{\pi c_1 k T}{\sin \pi c_1 k T} \\ &\simeq \frac{\pi c_1 k T}{\pi c_1 k T - (\pi c_1 k T)^3/3! + \dots} \\ &\simeq 1 + \frac{1}{6} (\pi c_1 k T)^2 \end{aligned} \quad (6-152)$$

Equations 6-151 and 6-152 are identical to Equations 6-82 and 6-84, indicating that the basic tunneling processes in field emission and in tunneling through a thin film are identical. The only differences between these two cases are the values of b_1 and c_1 , which depend on the profile of the potential barrier. However, the slight quadratic dependence of the current on

temperature is characteristic of the tunneling process. Figure 6-28 shows the good agreement between the experimental results and Equation 6-152 for temperature dependence of tunneling current through Al_2O_3 films. The results are from Hartman and Chivian.⁶³ It should be noted that at $T = 0$, Equation 6-151 is more general than Equation 6-140 and can be used for any shape of the potential barrier without the need to approximate the shape of the barrier to a rectangular one before calculations.

6.3.2 Elastic and Inelastic Tunneling

The tunneling current through a thin insulating film is determined by the supply function and the effective height and the effective width of the potential barrier, which are strongly dependent on the barrier profile. According to energy conservation, the electrons in electrode 1 can undergo the transition at constant energy into empty states in electrode 2 when a voltage V is applied to the system, as shown in Figure 6-29. The tunneling in which the energies of the electrons remain unchanged during transition is generally referred to as *elastic* tunneling. However, the energies of the tunneling electrons may change at certain applied voltages under the following conditions:

- If the impurities or traps at a certain energy level in the insulator are ionized by inelastic collisions with the tunneling electrons, this process will involve changes in both the supply function and the barrier profile.
- If the tunneling electrons lose part of their energies to excite the vibrational modes of atomic or molecular species in the insulator through inelastic collisions, this process may be thought of as a perturbation to the barrier profile.^{64,65}
- If the electrons tunnel through an MIS system, electrons may tunnel to surface states in the semiconductor energy gap with the subsequent recombination of the trapped electrons with holes in the valence band.^{66,67}

The tunneling in which the tunneling electrons give part of their energy to one of the inelastic

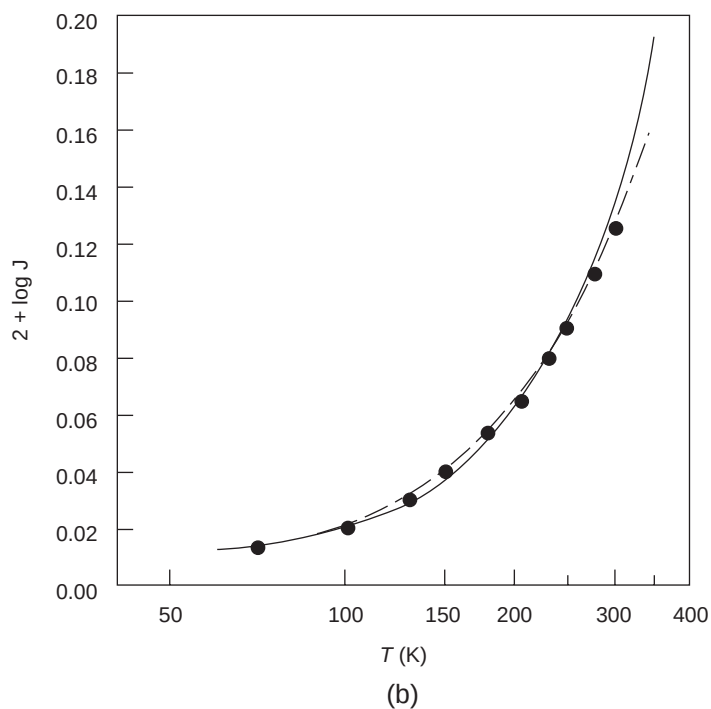
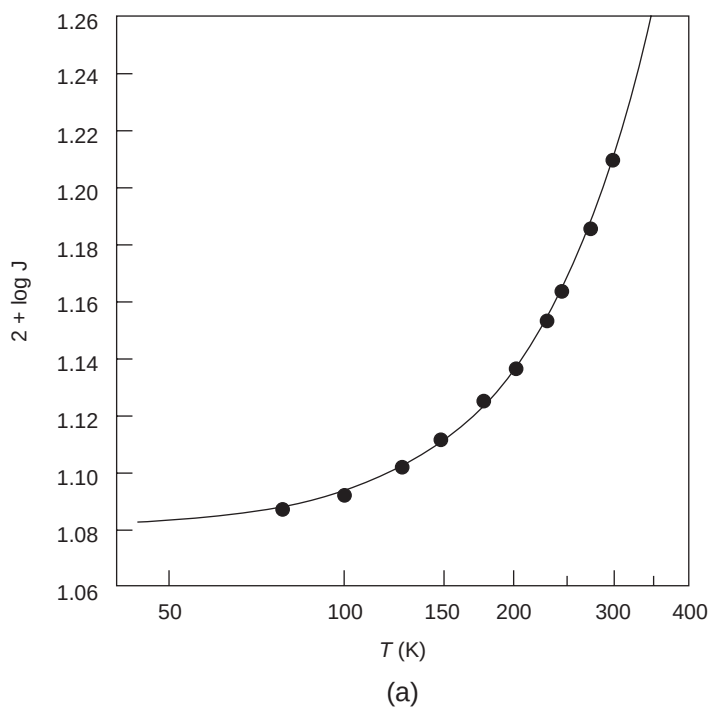


Figure 6-28 Temperature dependence of the tunneling current density in an Al-Al₂O₃-Al system with the oxide thickness of 30 Å; (a) 0.5 V and (b) 0.05 V. The points are experimental and the solid curves theoretical results.

transfer processes previously mentioned during transition, is referred to as *inelastic* tunneling.

When the energy difference ΔE between the energy level of the tunneling electron in electrode 1 and that of the same electron in electrode 2 is equal to $\hbar\omega_0$, where $\hbar\omega_0$ is the excitation energy of one type of internal mode in the insulator, a new channel for tunneling opens up and the total tunneling current increases, as shown in Figure 6-29, provided that $qV \geq \Delta E = \hbar\omega_0$ and that empty states in electrode 2 are available for the tunneling. If the insulator has more than one type of internal mode that can be excited, more kinks like the one shown in Figure 6-29 will appear in the J - V curve. The tunneling J - V characteristics may be employed as a tunnel spectroscopy to measure the energy levels of adsorbed or adsorbed impurities in MIM or MIS systems. Since the magnitude of the tunneling current resulting from inelastic processes is smaller than that resulting from elastic processes, it is generally necessary to plot the first or second derivatives of the current with respect to voltage, versus voltage in order to reveal the inelastic processes.

Inelastic tunneling has been theoretically analyzed by several investigators.^{38,65,68} In inelastic tunneling, the electrons are expected to tunnel from the states on one side to the empty states with lower energy on the other side. At the same time, the impurities in the insulator are excited from their ground states to excited states. The excited states of the impurities inside the barrier can be interpreted as a perturbation to the height of the barrier, because the barrier height depends on the interaction of tunneling electrons with impurities. Based on the analysis of Pollack and Seitchik,³⁸ the inelastic portion of the tunneling current can be expressed as

$$J_{\text{inelastic}} = \frac{2\pi^2 m}{h^2 \phi_0} \left| \left\langle \phi_{\text{ex}} \left| \int_0^S U_{\text{in}} dx \right| \phi_{\text{gr}} \right\rangle \right|^2 J_{\text{elastic}} \quad (6-153)$$

where J_{elastic} is the same tunneling current given in section 6.3.1 and Equations 6-151 and 6-152, S is the width of the insulator, ϕ_{gr} and ϕ_{ex} are, respectively, the wave functions of the impurities in the ground and the excited states, and U_{in} is the interaction energy between tunneling electrons and impurities, which depends on the type of internal mode.

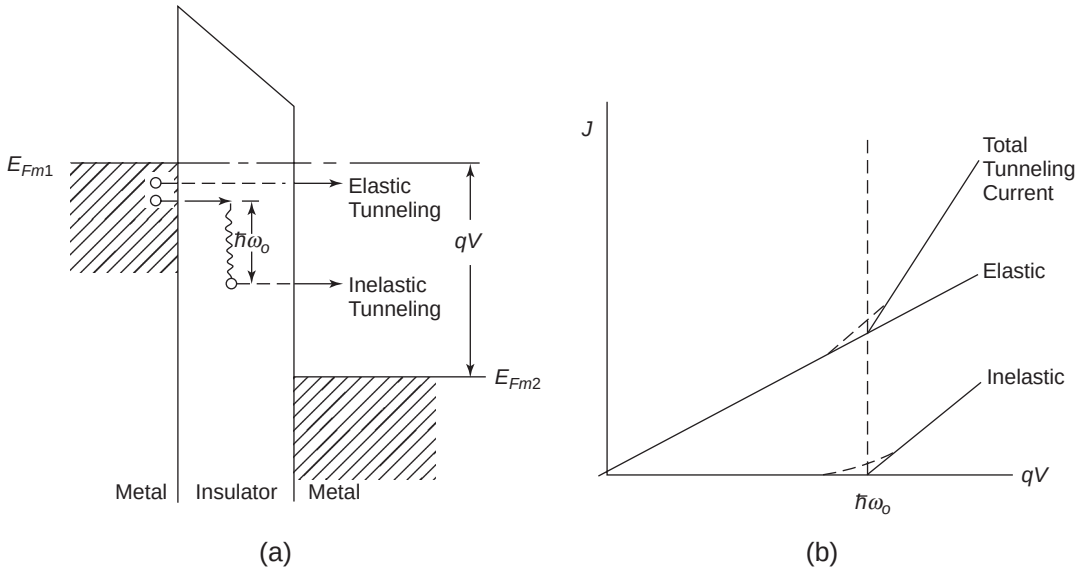


Figure 6-29 (a) Elastic tunneling and inelastic tunneling via an energy-loss process (such as one involving excitation energy $\hbar\omega_0$); (b) the J - V characteristics for elastic and inelastic tunneling. Solid lines are for $T = 0\text{ K}$, and dashed lines include the effect of thermal smearing for $T > 0$.

For example, electron-dipole interaction energy due to a dipole moment p_x associated with the molecular vibration located at one of the electrodes and taking into account the nearest image of the dipole is

$$U_{in} = 2qxP_x / (x^2 + r_{\perp}^2)^{3/2} \quad (6-154)$$

The electron-induced dipole interaction energy due to a dipole induced by the electric field of the electron and taking into account the nearest image of the induced dipole is

$$U_{in} = -4q^2x^2\alpha / (x^2 + r_{\perp}^2)^3 \quad (6-155)$$

in which the polarizability α usually varies with the vibration of the molecule.

Inelastic tunneling due to the excitation of impurities in the barrier has been observed by Lambe and Jaklevic.⁶⁵ They observed the changes in tunneling conductance of Al – Al₂O₃ – Pb systems at $T = 4.2$ K at definite voltages. The voltages have been identified as

$$V = \frac{\hbar\omega_o}{q}, \text{ directly relating to vibrational}$$

frequencies of bending and stretching vibrational mode (C—H or O—H) of hydrocarbons present in the oxide film as impurities. The coupling of the electron and the impurity in this case may be associated with the interaction of the electron with the dipole moment of the C—H and O—H bonds in the adsorbed hydrocarbons, as shown in Figure 6-30. These investigators have plotted $\frac{d^2J}{dV^2}$ versus V for the

Al – Al₂D₃ – Pb system and compared this result to the infrared spectrum of the bulk specimen of the hydrocarbon, which was deliberately adsorbed on the Al – Al₂O₃ – Pb system used for obtaining the tunneling J - V characteristics. These results are also shown in Figure 6-30 and are in good agreement with each other. The results are from Lambe and Jaklevic.⁶⁵

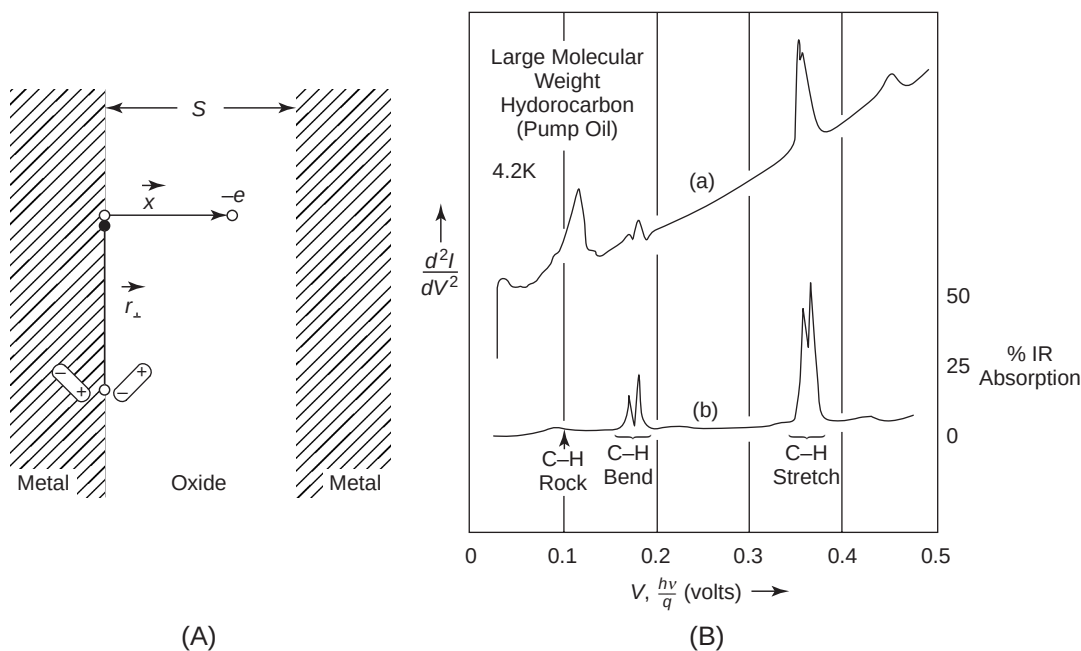


Figure 6-30 (A) The interaction of a tunneling electron with a dipole in the oxide barrier (insulator) and its image in the metal. (B) (a) The tunneling d^2I/dV^2 spectrum of an Al-oxide-Pb system with a monolayer of a large molecular-weight hydrocarbon adsorbed on the oxide; (b) the infrared spectrum of the bulk hydrocarbon, of which a sample was adsorbed on the oxide used to obtain (a). The origin of the various modes is indicated in the figure.

6.3.3 Tunneling through a Metal–Thin Insulating Film–Semiconductor (MIS) System

Numerous studies of tunneling phenomena in MIS systems have been reported, with considerable attention to the use of these phenomena as a spectroscopic tool for the investigation of the electronic band structure of the semiconductor surface, as well as for electronic applications.^{66,67,69–81} The difference between a metal–semiconductor junction and an MIS system can easily be seen from the energy band diagrams shown in Figure 6-31 and Figure 6-2. To use Figure 6-31 to describe the features of the tunneling through the MIS system, it is convenient to define some important symbols as follows:

$\bar{\phi}_e$, $\bar{\phi}_h$: The average barrier heights of the insulator for electron tunneling between the metal and the conduction band, and for hole tunneling between the metal and the valence band of the semiconductor, respectively, which are, in general, functions of applied voltage

S: The barrier width of the insulator, which is the thickness of the insulator (e.g., the oxide layer)

V_i , V_{io} : The voltage drop across the insulator with and without applied voltage, respectively

ψ_s , ψ_{so} : The surface potential at the semiconductor surface due to band bending with and without applied voltage, respectively

E_{cs} , E_{cso} : The energy level of the conduction band edge at the semiconductor surface with and without applied voltage, respectively

E_{vs} , E_{vso} : The energy level of the valence band edge at the semiconductor surface with and without applied voltage, respectively

From Figure 6-31, the applied voltage V will appear partially across the insulator and partially across the semiconductor. Thus,

$$V = V_i + \psi_s + (\phi_m - \phi_s)/q \quad (6-156)$$

For simplicity, we assume that there are no charges present inside the insulator. There-

fore, the electric field in it is uniform and is given by

$$E_i = \frac{V_i}{S} \quad (6-157)$$

The presence of the thin insulating film will make the distance between E_{Fm} and the semiconductor conduction band edge E_{cs} at the interface dependent on the applied voltage, because the voltage drop V_i across the insulating film is dependent on applied voltage and so is E_{cs} or E_{vs} . This causes the voltage dependence of the tunneling current through the thin insulating film by the following factors:

The voltage dependence of the tunneling transmission coefficient due to the voltage dependence of the average barrier heights $\bar{\phi}_e$ and $\bar{\phi}_h$ for tunneling

The voltage dependence of the supply function due to the voltage dependence of $|E_{Fm} - E_{cs}|$ or $|E_{Fm} - E_{vs}|$

For the MIS (n-type) system shown in Figure 6-31, when a forward bias brings E_{Fm} in alignment with E_{vs} , the tunneling hole current J_{mv} will begin to increase. (This means that electrons in the valence band see a large number of empty states in the metal; electrons tunneling from the valence band to the metal are equivalent to the holes tunneling from the metal to the valence band.) This condition is

$$\begin{aligned} qV_{w(\text{Forward})} &= E_{Fs} - E_{vs} \\ &= E_{Fs} - [E_v + q\psi_{sw(\text{Forward})}] \end{aligned} \quad (6-158)$$

With an increase in the forward bias, J_{cm} also increases because the electron concentration n in the conduction band increases; but both J_{mc} (electron flow from the metal to the conduction band) and J_{vm} (hole flow equivalent to electron flow from the metal to the valence band) decrease. When a reverse bias brings E_{Fm} in alignment with E_{cs} , the tunneling electron current J_{cm} (electron flow from the metal to the conduction band) will begin to increase. This condition is

$$\begin{aligned} qV_{w(\text{Reverse})} &= E_{cs} - E_{Fs} \\ &= [E_c + q\psi_{sw(\text{Reverse})}] - E_{Fs} \end{aligned} \quad (6-159)$$

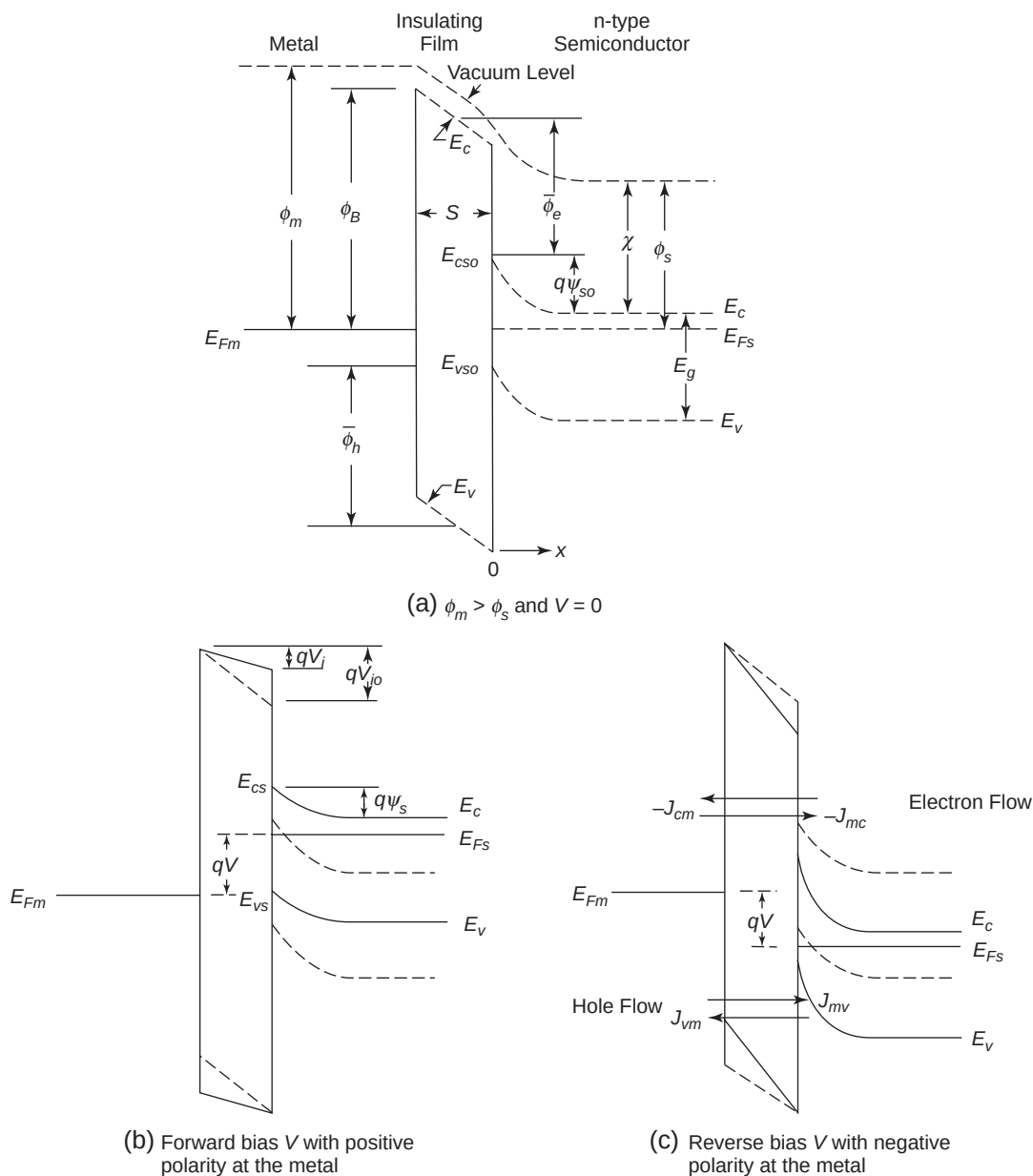


Figure 6-31 Energy band diagrams for a metal-insulator-semiconductor system. ----- at zero bias, — under an applied bias.

With an increase in the reverse bias, J_{vm} also increases because the band bending results in an increase in the number of empty states in the valence band. Both J_{cm} and J_{mv} , however, become negligible under the reverse-bias condition. Therefore, in plotting J - V or G - V

characteristics, where G is the dynamic conductance dJ/dV , there is a wide region of low conductance, and beyond the ends of this region the conductance rises rapidly. This region of low conductance is generally referred to as the *conductance well*.^{66,82} From Equations

6-158 and 6-159, the width of the conductance well is

$$V_w = \frac{E_g}{q} + [\psi_{sw(\text{Reverse})} - \psi_{sw(\text{Forward})}] \quad (6-160)$$

$$= V_{w(\text{Reverse})} + V_{w(\text{Forward})}$$

The presence of interface states located in the energy band gap at the insulator–semiconductor interface affect V_w through $\psi_{sw(\text{Reverse})}$ and $\psi_{sw(\text{Forward})}$. The charge in the interface states tends to shield the semiconductor from the metal and makes communication between the metal and the semiconductor more difficult. This leads to a smaller voltage dependence of semiconductor band bending and hence a narrower conductance well.

For the MIS (n-type) system shown in Figure 6-31, the expression for the majority carrier (electron) tunneling currents for one-dimensional tunneling is⁷⁸

$$J_n = J_{cm} - J_{mc} = A_n T^2 \exp(-\bar{\phi}_e^{1/2} S) \times \exp[-(q\psi_{so} + E_c - E_{Fs})/kT] \quad (6-161)$$

$$\times [\exp(qV/n_e kT) - 1]$$

and the minority carrier (hole) tunneling current is

$$J_p = J_{mv} - J_{vm} = A_p T^2 \exp(-\bar{\phi}_h^{1/2} S) \times \left\{ \exp\left[\frac{E_{vo} - E_{Fpo}}{kT}\right] - \exp\left[\frac{E_{vo} - E_{Fm}}{kT}\right] \right\} \quad (6-162)$$

where

$$A_n = \frac{4\pi q m_{te} k^2}{h^3}$$

$$A_p = \frac{4\pi q m_{th} k^2}{h^3}$$

m_{te} = the effective mass for electrons with momentum transverse to the barrier

m_{th} = the effective mass for holes with momentum transverse to the barrier

n_e = the ideality factor for electrons, which is equal to $V/(\psi_{so} - \psi_s)$

E_{Fpo} = the hole quasi-Fermi level at the semiconductor surface ($x = 0$)

In Equation 6-161, V is positive for the forward bias and negative for the reverse bias; in Equations 6-162, the bias dependence of J_p is obtained through the bias dependence of E_{vo} and E_{Fpo} . It should be noted that $(-\bar{\phi}_e^{1/2} S)$ in $\exp(-\bar{\phi}_e^{1/2} S)$ and $(-\bar{\phi}_h^{1/2} S)$ in $\exp(-\bar{\phi}_h^{1/2} S)$ are dimensionless and come from the following expression

$$\exp\left[-\frac{4\pi s}{h}(2m\bar{\phi})^{1/2}\right] \approx \exp[-1.01\bar{\phi}^{1/2} S] \quad (6-163)$$

$$\approx \exp[-\bar{\phi}^{1/2} S]$$

if $\bar{\phi}$ is expressed in electron volts and S in angstroms.^{61,78}

In MIS systems, we can adjust the applied voltage to increase the minority carrier injection ratio, defined as the ratio of the minority carrier current to the total current. This can be easily realized from Figure 6-31 and Equations 6-161 and 6-162. On the basis of this principle, Card and Smith⁸¹ have observed green luminescence in an Au-SiO₂-GAP (n-type) system under forward bias. Figure 6-32 shows the green light output power of the Au-SiO₂-GAP system as a function of the thickness of SiO₂

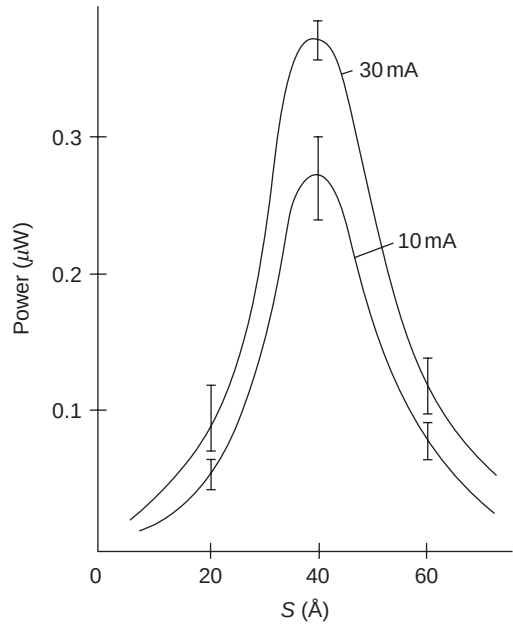


Figure 6-32 Green light output power as a function of SiO₂ film thickness for Au-SiO₂-GaP system for two values of tunneling current under forward bias.

films for two values of tunneling current under forward bias. For a fixed value of current, the light output power (or light intensity) increases with S , reaches a peak, and then decreases with increasing S . This indicates that the minority carrier injection ratio $[J_p/(J_n + J_p) \sim J_p/J_n]$ initially increases with S (when $E_{Fm} \sim E_{vs}$), but after reaching the peak, it decreases with S because the barrier width increases, causing a rapid decrease in the transmission coefficient for tunneling.

6.3.4 Effects of Space Charges and Traps on Tunneling Efficiency and Impurity Conduction

In general, insulating films contain traps. The space charge due to trapped electrons (the trapped space charge) may seriously affect the profile of the potential barrier and hence the tunneling J - V characteristics. If there is no space charge, the applied voltage reduces the effective height and width of the potential barrier for electron tunneling, as shown in Figure 6-33(b). Suppose that the net space charge density is zero in the region $0 < x < S_2$ and is constant in the region $S_2 < x < S$, where $|S_2|$ is the barrier width for tunneling. The space charge will alter the profile of the potential barrier by increasing the barrier width to limit electron tunneling, as shown in Figure 6-33(c). Thus, the space charge effects can be accounted for by extending the straight-line portion of the actual barrier from $x = S'_2$ to $x = S$ and using an effective bias V' in place of the actual bias V to calculate the tunneling distance $|S_2 - S_1|$ and the average barrier height $\bar{\phi}$.

Several investigators have studied these effects of space charge and reported that free carrier space charge without traps is ineffective in lowering the tunneling current, even for extremely low free carrier mobilities, that a high trap density can severely limit the tunneling current, and that low free carrier mobilities enhance space charge effects.^{83,84}

Insulating or semiconducting films always contain impurities, which may be unavoidably present or may be deliberately doped in the film specimens. These impurities form impurity

states in the forbidden energy gap. When the localized electronic wave functions of the impurity states overlap, an electron bound to one impurity state can tunnel to an unoccupied impurity state without involving activation into the conduction band. This tunneling process between impurity sites is referred to as *impurity conduction*.⁸⁵ The mobility of an electron moving in the impurity states is very small since it depends on interaction between widely spaced impurities, so this conduction mechanism usually becomes predominant at low temperatures due to a low concentration of carriers in the conduction and valence bands. This depends, however, on impurity concentration and the energy levels of the impurity states, which control the probability of tunneling from impurity site to impurity site and the number of electrons taking part in this tunneling process.

In semiconductors, the impurity conduction process is possible only if the material is compensated (i.e., if the material contains both donor and acceptor impurities). This condition for impurity conduction was put forward by Mott⁸⁶ and Conwell⁸⁷ and confirmed experimentally by Fritzsche.⁸⁸⁻⁹⁰ For example, if the donor concentration N_D is larger than the acceptor concentration N_A in a compensated n-type semiconductor, all the acceptors will be occupied and become negatively charged; only $N_D - N_A$ donors remain occupied and neutral at low temperatures, as shown in Figure 6-34. If impurity state A and impurity state B are at the same energy level, the overlap of the wave functions between these two sites will enable the movement of an electron from an occupied to an empty donor site without involving activation into the conduction band. If the impurity state A is located at a lower energy level than the impurity state B , then thermal energy (phonon), supplied by lattice vibrations of the material, is required to assist electron tunneling from A to B . The field created by the charged acceptors and donors will split the energy levels of donor states; therefore, an electron can tunnel from one impurity state to another only by exchanging energy with phonons. The applied voltage will alter the energy level difference between sites, thereby making the tunneling probability

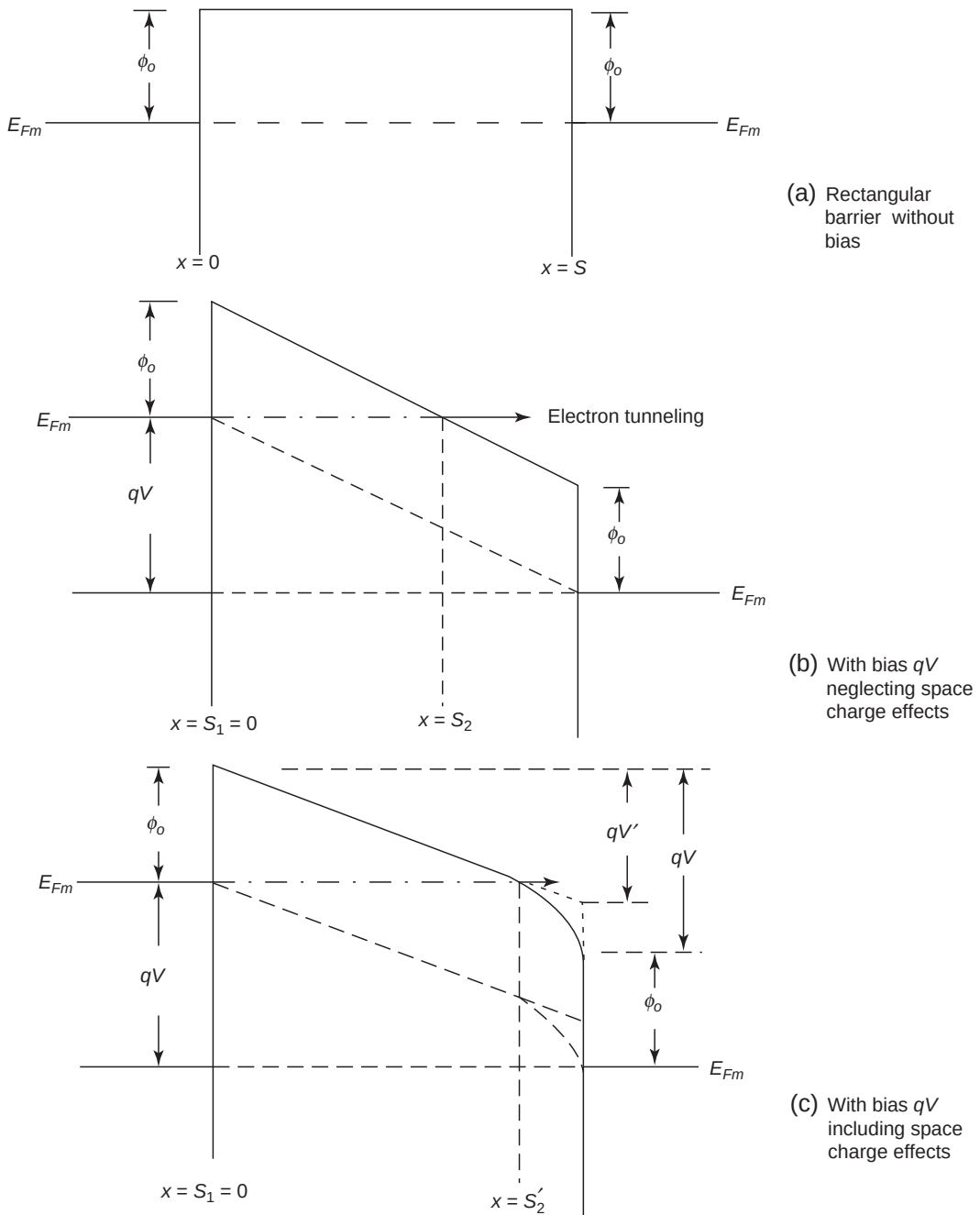


Figure 6-33 Energy band diagrams for a thin insulating film between two similar metal electrodes with and without space charge effects.

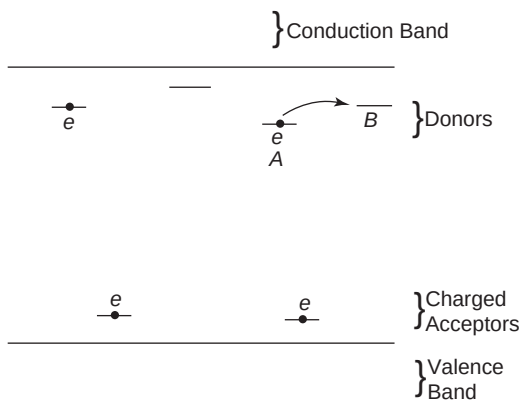


Figure 6-34 An electron from an occupied impurity state A hopping to an empty impurity state B due to thermally assisted tunneling in a compensated n-type semiconductor with $N_D > N_A$. (After Mott and Davis 1971.)

higher in one direction than in the other. A similar process can readily be realized in compensated p-type semiconductors ($N_A > N_D$), but in this case, electrons tunnel through acceptor impurity sites.

The following are the most significant features of impurity conduction observed in resistivity–temperature characteristics:

- Resistivity is strongly dependent on impurity concentration.
- The plot of $\ln \rho$ versus $1/T$ exhibits a finite slope, indicating that a thermal activation energy is required for electron tunneling between sites when the impurity concentration is small.
- Activation energy decreases with increasing impurity concentration and becomes zero when the impurity concentration reaches a certain critical value or higher. This indicates that with impurity concentrations higher than such a critical value, carriers move freely without involving thermal activation.

Some typical experimental results obtained by Fritzsche and Cuevas⁹⁰ in compensated p-type germanium semiconductors are shown in Figure 6-35. A similar impurity conduction phenomenon has also been observed in tantalum oxide thin films,⁹¹ in silicon monoxide films,⁹² in nickel oxide,^{93,94} in vanadium phosphate glasses,⁹⁵ and in many other materials.^{85,96}

6.4 Charge Transfer at the Metal–Polymer Interface

Electron (or hole) injection from a metallic contact to a polymer has been considered due to electron transfer via the surface states in the polymer.^{97,98} Several investigators have proposed a similar electron transfer mechanism.^{99–101} The density of surface states in the polymer is likely to be much higher than that of bulk trapping states, due partly to direct exposure of the surface to the external environment and partly to the metallization process in forming the metallic contact. The surface states are distributed in the energy band gap and extended to a depth of x_s from the interface, which may be about 300 Å.¹⁰² Beyond x_s , the bulk trapping states are dominant, but their density is smaller. Within the region dominated by surface states, electrons thermally activated can tunnel from the metallic electrode to the surface states, and then move up to the conduction band by thermal activation, as shown in Figure 6-36. In the region containing bulk trapping states, electrons cannot tunnel from state to state because of the large separation between states.

For polyethylene, it is likely that the bulk electron trapping states are distributed following Gaussian distribution between E_c and E_F , while the bulk hole trapping states are also distributed in the same manner between E_F and E_v . It is also likely that the electron traps are acceptorlike and the hole traps are donorlike localized states. In thermal equilibrium, all surface and bulk trapping states can be assumed to be empty above the Fermi level E_F and occupied below E_F . However, when electrons are injected under an applied field from the metallic electrode to the conduction band of the polymer via the surface states, the major trapping will take place between the quasi-Fermi level and the thermal equilibrium Fermi level. Initially, the injected current is large, but as the traps gradually become filled, the current gradually reduces to the trap-controlled SCL current. Depending on the density, the distribution and the depth of traps in the energy band gap and the trap filling and thermal detrapping processes can be very slow for large

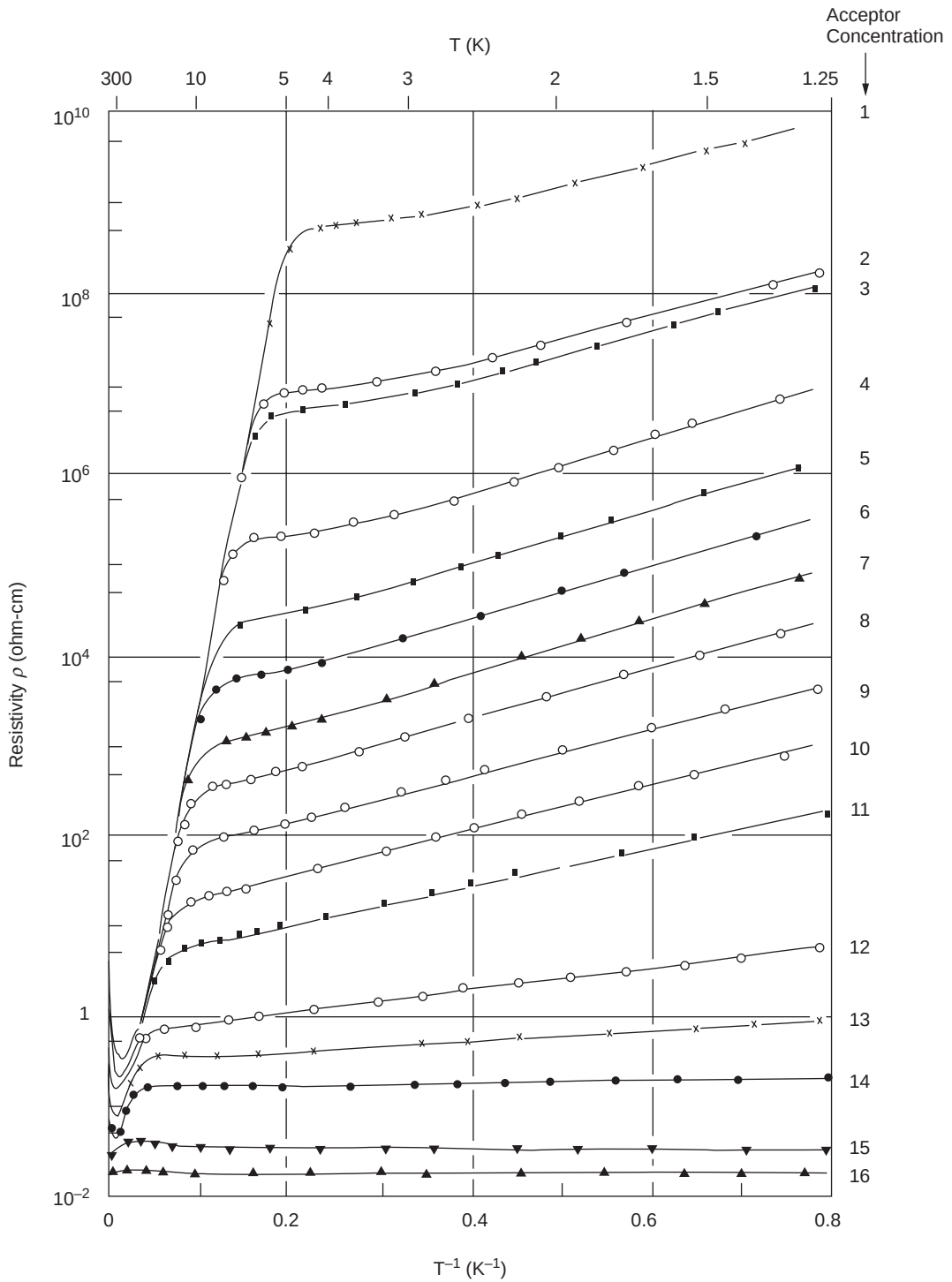


Figure 6-35 Resistivity ρ in ohm-cm as a function of temperature for compensated p-type germanium semiconductors with compensation $= N_D/N_A = 0.4$ for the acceptor concentrations in cm^{-3} : (1) 7.5×10^{14} ; (2) 1.4×10^{15} ; (3) 1.5×10^{15} ; (4) 2.7×10^{15} ; (5) 3.6×10^{15} ; (6) 4.9×10^{15} ; (7) 7.2×10^{15} ; (8) 9.0×10^{15} ; (9) 1.4×10^{16} ; (10) 2.4×10^{16} ; (11) 3.5×10^{16} ; (12) 7.3×10^{16} ; (13) 1.0×10^{17} ; (14) 1.5×10^{17} ; (15) 5.3×10^{17} ; (16) 1.35×10^{18} .

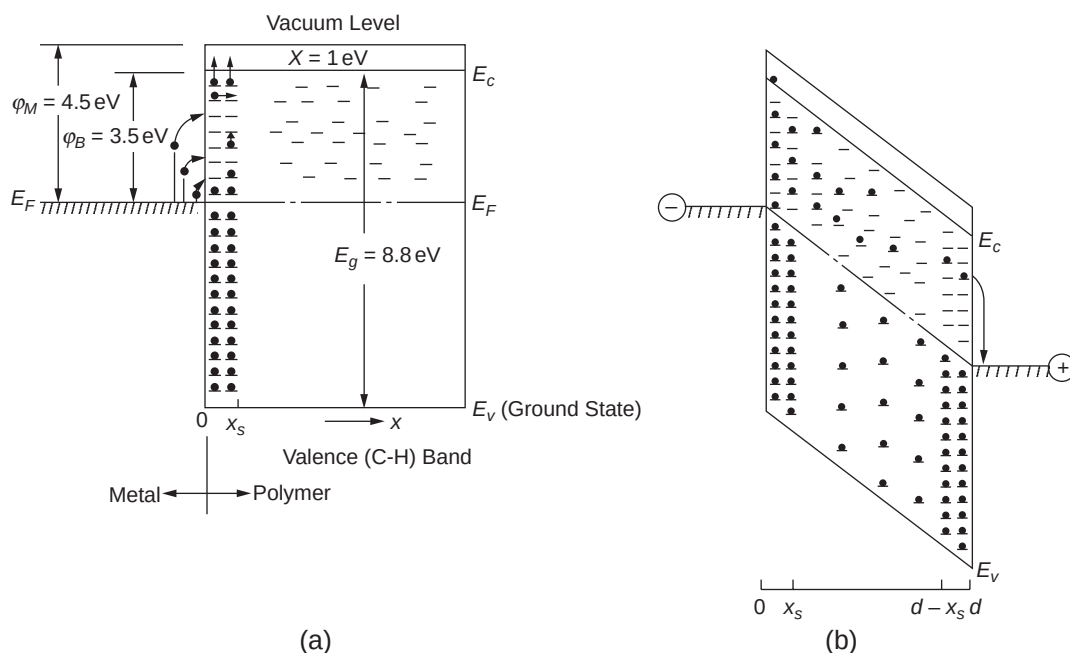


Figure 6-36 Simple energy band diagrams illustrating the distributions of surface and bulk trapping states in energy and in space: (a) without an applied field and (b) with an applied field, showing the electrons tunneling from the Au cathode to the polyethylene via surface states and moving up to the conduction band. (The hole traps in the bulk are not shown for clarity.)

bandgap materials such as polyethylene. This is why it usually takes a very long time for the charging current under a step-function voltage to decay to a steady-state value.

References

1. W.R. Harper, *Contact and Frictional Electrification*, (Oxford University Press, Oxford, 1967).
2. E. Wigner and J. Bardeen, *Phys. Rev.*, **48**, 84 (1935).
3. J. Bardeen, *Phys. Rev.*, **49**, 653 (1936).
4. W. Shockley, *Electrons and Holes in Semiconductors*, (Van Nostrand, New York, 1950).
5. C. Herring and M. H. Nichols, *Rev. Mod. Phys.*, **21**, 185 (1949).
6. W. H. Brattain and C. G. B. Garrett, "Surface States," in "Methods of Experimental Physics," *Solid State Physics*, vol. 6, edited by K. Lark-Horovitz and V. A. Johnson, (Academic Press, New York, 1959), pp. 136–144.
7. V. K. Zworykin and E. G. Ramberg, *Photoelectricity and Its Applications*, (Wiley, New York, 1949).
8. F. Seitz, *Modern Theory of Solids*, (McGraw-Hill New York, 1940).
9. J. G. Simmons, *J. Phys. Chem. Solids*, **32**, 1987 and 2581 (1971).
10. A. Rose, *Concepts in Photoconductivity and Allied Problems*, (Wiley Interscience, New York, 1963).
11. G. T. Wright, *Solid State Electron.*, **2**, 165 (1961).
12. J. Bardeen, *Phys. Rev.*, **71**, 717 (1947).
13. W. Shockley and G. L. Pearson, *Phys. Rev.*, **74**, 232 (1948).
14. I. Tamm, *Z. Phys.*, **76**, 849 (1932), and *Phys. Z. Sowjetunion*, **1**, 733 (1932).
15. W. Shockley, *Phys. Rev.*, **56**, 317 (1939).
16. D. Hanemann, *Proc. International Conf. on Semiconductor Physics in 1960* (Czechoslovak Academy of Science, Prague, 1961), p. 540.
17. H. E. Farnsworth, R. E. Schlier, T. H. George, and R. M. Burger, *J. Appl. Phys.*, **26**, 252 (1955).

18. M. Lax, Proc. International Conf. on Semiconductor Physics in 1960, (Czechoslovak Academy of Science, Prague, 1961), p. 484.
19. E. H. Rhoderick, *Metal-Semiconductor Contacts*, (Clarendon, Oxford, 1978).
20. W. H. Brattain and J. Bardeen, Bell System Tech. J., 32, 1-41 (1953).
21. Y. Takai, T. Osawa, K. C. Kao, T. Mizutani, and M. Teda, Japan J. Appl. Phys., 14, 473 (1975).
22. M. M. Labes, O. N. Rudyj, and P. L. Kronick, J. Am. Chem. Soc., 84, 499 (1962).
23. H. Banser, Z. Phys., 188, 423 (1965).
24. W. Schottky, Naturwiss., 26, 843 (1938); also Z. Phys., 113, 367 (1939) and Z. Phys., 118, 539 (1942).
25. B. R. Gossick, *Potential Barriers in Semiconductors*, (Academic Press, New York, 1964).
26. J. G. Simmons, "Electronic Conduction through Thin Insulating Films," in *Handbook of Thin Film Technology*, (McGraw-Hill, New York, 1970), pp. 14.1-14.50.
27. C. R. Crowell, Solid State Electron., 8, 395 (1965), and also 12, 55 (1969).
28. P. R. Emtage and J. J. O'Dwyer, Phys. Rev. Lett., 16, 356 (1966).
29. J. J. O'Dwyer, *The Theory of Electrical Conduction and Breakdown in Solids*, (Clarendon, Oxford, 1973).
30. C. R. Crowell and S. M. Sze, Solid State Electron., 9, 1035 (1966).
31. C. R. Crowell and S. M. Sze, J. Appl. Phys., 37, 2683 (1966).
32. K. C. Kao, IEEE Trans. Electr. Insul., EI-11, 121 (1976).
33. W. Hwang and K. C. Kao, J. Chem. Phys., 60, 3845 (1974).
34. H. K. Henisch, *Semiconductor Contacts*, (Clarendon, Oxford, 1984).
35. S. M. Sze, C. R. Crowell, and D. Kahng, J. Appl. Phys., 35, 2534 (1964).
36. K. K. Thornber, T. C. McGill, and C. A. Mead, J. Appl. Phys., 38, 2384 (1967).
37. H. A. Bethe, "Theory of the Boundary Layer of Crystal Rectifiers," MIT Radiation Laboratory Report 43-12 (1942).
38. S. R. Pollack and J. A. Seitchik, "Electron Transport through Insulating Films," in *Applied Solid State Science*, vol. 1, edited by R. Wolfe, (Academic Press, New York, 1969), pp. 343-383.
39. C. A. Mead, "Electron Transport in Thin Insulating Films," in *Basic Problems in Thin Film Physics*, edited by R. Niedermayer and H. Mayer, Proc. of 1965 International Symposium held at Clausthal-Gottingen (1966), p. 674.
40. S. R. Pollack, J. Appl. Phys., 34, 877 (1963).
41. R. H. Fowler and L. Nordheim, Proc. Roy. Soc. (London), 119A, 173 (1928).
42. C. B. Duke, "Tunneling in Solids," *Solid State Physics Supplement 10*, (Academic Press, New York, 1969).
43. R. H. Good and W. Muller, "Field Emission," in *Handbuch der Physik*, vol. 21, (Springer-Verlag, Berlin, 1956), pp. 176-231.
44. E. Burstein and S. Lundqvist (Eds.), *Tunneling Phenomena in Solids*, (Plenum Press, New York, 1969).
45. F. A. Padovani and R. Stratton, Solid State Electron., 9, 695 (1966).
46. R. Stratton, "Tunneling in Schottky Barrier Rectifiers," in *Tunneling Phenomena in Solids*, edited by E. Burstein and S. Lundqvist, (Plenum Press, New York, 1969), pp. 105-126.
47. M. Lenzlinger and E. H. Snow, J. Appl. Phys., 40, 278 (1969).
48. E. L. Murphy and R. H. Good, Phys. Rev., 102, 1464 (1956).
49. S. Tiwari, *Compound Semiconductor Device Physics*, (Academic Press, New York, 1992).
50. T. H. Ning and H. N. Yu, J. Appl. Phys., 45, 5373 (1974).
51. E. H. Nicollian and J. R. Brews, *MOS (Metal-Oxide-Semiconductor) Physics and Technology*, (Wiley, New York, 1982).
52. D. J. McMaria, J. Appl. Phys., 47, 4073 (1976).
53. D. J. McMaria, R. Ghez, and D. W. Dong, J. Appl. Phys., 51, 4830 (1980).
54. N. R. Tu and K. C. Kao, J. Appl. Phys., 85, 7267 (1999).
55. B. Caranahan, H. A. Luther, and J. O. Wilkes, *Applied Numerical Methods*, (Wiley, New York, 1969).
56. R. Stratton Phys. Rev., 125, 67 (1962).
57. R. Stratton, Phys. Rev., 135, 794 (1964).
58. J. Frenkel, Phys. Rev., 36, 1604 (1930).
59. A. Sommerfeld and H. Bethe, in *Handbuch der Physik*, vol. XXIV/2, edited by H. Geiger and K. Scheel, (Springer-Verlag, Berlin, 1933), p. 450.
60. R. Holm, J. Appl. Phys., 22, 569 (1951).
61. J. G. Simmons, J. Appl. Phys., 34, 1793 (1963), and also 34, 2581 (1963).
62. J. G. Simmons, J. Appl. Phys., 35, 2472 (1964); also 35, 2655 (1964).

63. T. E. Hartman and J. S. Chivian, *Phys. Rev.*, **134A**, 1094 (1964).
64. R. C. Jaklevic and J. Lambe, *Phys. Rev. Lett.*, **17**, 1139 (1966).
65. J. Lambe and R. C. Jaklevic, *Phys. Rev.*, **165**, 821 (1968).
66. J. Shewchum, A. Waxman, and G. Warfield, *Solid State Electron.*, **10**, 1165 (1967).
67. P. V. Gray, *Phys. Rev.*, **140A**, 179 (1965).
68. D. J. Scalapino and S. M. Marcus, *Phys. Rev. Lett.*, **18**, 459 (1967).
69. P. V. Gray, *Phys. Rev. Lett.*, **9**, 302 (1962).
70. J. Shewchum, M. A. Green, and F. D. King, *Solid State Electron.*, **17**, 563 (1974).
71. J. Shewchum, R. Singh, and M. A. Green, *J. Appl. Phys.*, **48**, 765 (1977).
72. W. E. Dahlke and S. M. Sze, *Solid State Electron.*, **10**, 865 (1967).
73. L. B. Freeman and W. E. Dahlke, *Solid State Electron.*, **13**, 1483 (1970).
74. M. A. Green, F. D. King, and J. Shewchum, *Solid State Electron.*, **17**, 551 (1974).
75. L. Esaki and P. J. Stilles, *Phys. Rev. Lett.*, **14**, 902 (1965).
76. L. L. Chang, P. J. Stilles, and L. Esaki, *IBM Journal Res. Dev.*, **10**, 484 (1966).
77. L. L. Chang, P. J. Stilles, and L. Esaki, *J. Appl. Phys.*, **38**, 4440 (1967).
78. H. C. Card and E. H. Rhoderick, *J. Phys. D*, **4**, 1589 and 1602 (1971).
79. H. C. Card and E. H. Rhoderick, *Solid State Electron.*, **15**, 993 (1972).
80. H. C. Card and E. H. Rhoderick, *Solid State Electron.*, **16**, 365 (1973).
81. H. C. Card and B. L. Smith, *J. Appl. Phys.*, **42**, 5863 (1971).
82. A. Waxman, J. Shewchum, and G. Warfield, *Solid State Electron.*, **10**, 1187 (1967).
83. D. V. Geppert, *J. Appl. Phys.*, **33**, 2993 (1962) and also **34**, 490 (1963).
84. E. Pittelli, *Solid State Electron.*, **6**, 667 (1963).
85. N. F. Mott and W. D. Twose, *Adv. Phys.*, **10**, 107 (1961).
86. N. F. Mott, *Can. J. Phys.*, **34**, 1356 (1956).
87. E. M. Conwell, *Phys. Rev.*, **103**, 51 (1956).
88. H. Fritzsche, *J. Phys. Chem. Solids*, **6**, 69 (1958).
89. H. Fritzsche, *Phys. Rev.*, **115**, 336 (1959).
90. H. Fritzsche and M. Cuevas, *Phys. Rev.*, **119**, 1238 (1960).
91. C. A. Mead, *Phys. Rev.*, **128**, 2088 (1962).
92. J. G. Simmons and R. R. Verderber, *Radio Electron. Eng.*, **34**, 81 (1967).
93. A. J. Bosman and C. Crevecoeur, *Phys. Rev.*, **144**, 763 (1966).
94. A. J. Springthorpe, J. G. Austin, and B. A. Smith, *Solid State Commun.*, **3**, 143 (1965).
95. A. P. Schmid, *J. Appl. Phys.*, **39**, 3140 (1966).
96. N. F. Mott and E. A. Davis, *Electronic Processes in Non-Crystalline Materials*, (Clarendon, Oxford, 1979).
97. Y. Murata, *Japan J. Appl. Phys.*, **18**, 1 (1979).
98. Y. Murata, T. Hodoshima, and S. Kittaka, *Japan J. Appl. Phys.*, **18**, 2215 (1979).
99. C. B. Duke and T. J. Fabish, *Phys. Rev. Lett.*, **37**, 1075 (1976).
100. T. J. Fabish and C. B. Duke, *J. Appl. Phys.*, **48**, 4256 (1977).
101. T. J. Lewis, in *Polymer Surfaces*, edited by D. T. Clark and W. F. Feast (Wiley, New York, 1979), p. 65.
102. C. G. Garton, *J. Phys. D: Appl. Phys.*, **7**, 1814 (1974).

7 Electrical Conduction and Photoconduction

The hole is really an abstraction which gives a convenient way of describing the behavior of the electrons. The behavior of the holes is essentially a shorthand way of describing the behavior of all the electrons.

William Shockley

Electrical conduction and photoconduction are essentially governed by the manner of generating charge carriers and their transport in a material. If the carriers are generated by any means other than optical excitation (i.e., if they can be generated in the dark), the conduction is referred to as *dark electrical conduction* or simply *electrical conduction*. If the carriers are generated primarily by optical excitation, the conduction is referred to as *photoconduction*. In this chapter, the discussion is divided into two parts: Part I deals with electrical conduction and Part II with photoconduction.

PART I: ELECTRICAL CONDUCTION

7.1 Introductory Remarks

The electrical conductivities of materials range from those of superconductors through those of metals and semiconductors to those of highly resistive insulators. Electrical conductivity can be classified into three categories:

Intrinsic conductivity: Charge carriers are generated in the material based on its chemical structure only.

Extrinsic conductivity: Charge carriers are generated by impurities in the material, which may be introduced into it by fabrication processes or deliberately doped into it for a specific purpose.

Injection-controlled conductivity: Charge carriers are injected into the material mainly from metallic electrodes through a metal-material interface.

As far as insulators are concerned, the origin of the charge carriers for intrinsic or extrinsic conductivity is by no means clear. In this chapter, we shall cover only electrical conduction in dielectric solids, just a fraction of the large volume of work on electrical conduction,¹⁻⁶ and discuss generally applicable principles in hopes of providing a framework for an understanding of this diverse behavior.

At low fields—this means at average applied fields F (i.e., the applied voltage V divided by the specimen thickness d) lower than the threshold (or critical) field F_{th} for switching on significant injection of carriers from the carrier-injecting contacts—electrical conductivity follows the empirical equation given by

$$\sigma = \sigma_o \exp(-E_\sigma/kT) \quad (7-1)$$

where σ_o is the preexponential factor and E_σ is the activation energy. Neither the interpretation of σ_o nor E_σ is straightforward; it is rather ambiguous because electrical conduction involves various transport processes. The basic equation for total conductivity can be written as

$$\begin{aligned} \sigma_T &= q(\mu_n n + \mu_p p) + q(\mu_- n_- + \mu_+ n_+) \\ &= \sigma + \sigma_{ion} \end{aligned} \quad (7-2)$$

where n and p are, respectively, the concentrations of electrons and holes; μ_n and μ_p are, respectively, the average mobilities of electrons and holes; n_- and n_+ are, respectively, the concentrations of negative and positive ions; μ_- and μ_+ are, respectively, the average mobilities of negative and positive ions; and σ and σ_{ion} are, respectively, the electronic and ionic con-

ductivities. Under certain conditions, electrical conduction may involve both electronic and ionic conduction. For simplicity, we shall discuss these two types of electrical conduction separately.

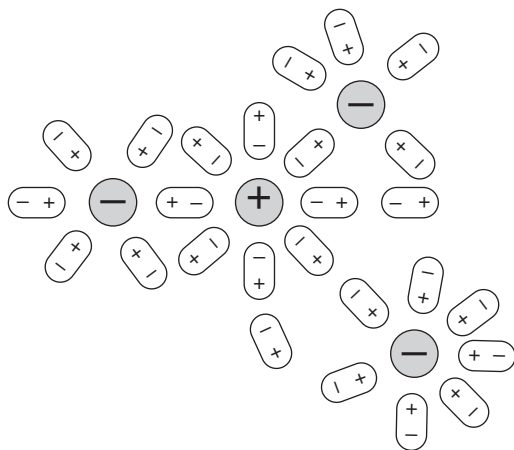


Figure 7-1 Schematic illustration of the screening of ions by surrounding polarized molecules or atoms.

7.2 Ionic Conduction

Ionic conduction in dielectric solids involves the transport of ions which have a mass like their surroundings. An ion with a charge q in a solid will polarize its surroundings. As a result, the polarized atoms or molecules will rearrange themselves to form a dipole hole, providing a screening effect on the ion, as shown in Figure 7-1.

Since the polarization of the surroundings reduces the electrostatic energy of the ion, the effect can be considered an ionic trap hindering the motion of the ion. This is equivalent to the creation of a potential barrier; an ion must move by an activation process from one potential well to another by surmounting the barrier height, as shown in Figure 7-2(a). An ion not only polarizes the surroundings but is also polarized by the opposite charges of the surrounding dipoles as it moves past them. This causes an increase in the height of the potential barrier. In short, interaction between ions and

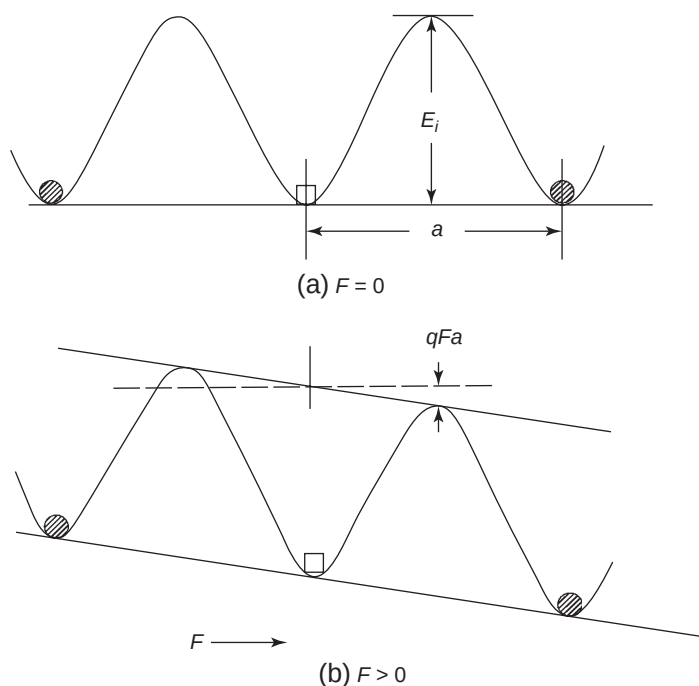


Figure 7-2 Schematic illustration of the potential energy barrier hindering the transport of a cation vacancy in the lattice: (a) in the absence of an electric field and (b) with an applied field F .

the surroundings hinders the movement of ions. However, ions require energy to surmount the barrier height, as well as vacant sites into which they can move. Vibration of ions involving an energy kT can cause ions and vacancies to exchange sites. Ionic conduction can be classified into two categories: intrinsic and extrinsic, based mainly on the manner of generating ions.

7.2.1 Intrinsic Ionic Conduction

For ionic crystals, ionic conduction is mainly intrinsic because the crystals have thermally created vacant sites in the lattice through which host ions can move. The simplest types of thermally created vacant sites (lattice defects) are the well known Frenkel and Schottky defects. A *Frenkel defect* is formed when an ion originally at a lattice site moves to an interstitial position, which implies that this process generates two imperfections: a vacancy in the lattice and an interstitial ion. A *Schottky defect* is formed when an ion originally at a lattice site diffuses to a surface position, creating one imperfection: a vacancy in the lattice. Since the volume and the surface of a crystal must be electrically neutral, Schottky defects must be created in pairs: one vacancy created by displacing an anion and the other by displacing a cation. It has been found that in most alkali-halide ionic crystals, probability for the formation of Schottky defects is much higher than for Frenkel defects.⁷ It can be imagined that for large ions, vacant lattice sites must be present for the ionic movement, whereas small ions can move through the interstitial space. For nonionic solids, some structures may provide channels, allowing ions space to move.

Ionic conduction is due simply to the transport of negative and positive ions. If the material contains several different species of ions, the ionic conductivity may be written as

$$\sigma_{\text{ion}} = q \sum_j \mu_j n_j \quad (7-3)$$

where μ_j and n_j are, respectively, the mobility and the concentration of ions belonging to species j . To describe the transport process clearly, we will assume for simplicity that the conductivity is due to the movement of one

type of ion. We will take the NaCl crystal as example, since it is an important and much investigated material. An NaCl lattice contains mainly Schottky defects, and the Na^+ ion is much smaller than the Cl^- ion. So the ionic conduction in NaCl can be considered almost entirely due to the movement of Na^+ via Schottky defects.

As $\sigma_{\text{ion}} = qu_+n_+$, the density of ions n_+ can be considered the density of vacancies in the lattice in which Na^+ ions are missing. This is analogous to the case of hole conduction in semiconductors, for which the density of holes in the valence band is the density of quantum states that are empty (unoccupied by electrons). So in ionic crystals, the movement of ions is in fact the movement of vacancies.

In the absence of an electric field, the probability per unit time for a vacancy to move to a neighboring position is given by

$$w = v_o \exp\left(-\frac{E_i}{kT}\right) \quad (7-4)$$

where v_o is the number of attempted escapes per second, which is the vibration frequency of the ions surrounding the vacancy, and E_i is the activation energy, which is the height of the potential barrier. Based on a simple one-dimensional model, the cation vacancy has equal probability of escape to the right or to the left in the absence of an applied field. Under an applied field, the probability changes because the barrier height is changed by an amount of qFa , where a is the lattice constant as shown in Figure 7-2.

For the vacancy moving in the direction of the field, the probability becomes

$$w' = v_o \exp\left[-\left(E_i + \frac{1}{2}qFa\right)/kT\right] \quad (7-5)$$

and that in the direction opposite to the field becomes

$$w'' = v_o \exp\left[-\left(E_i - \frac{1}{2}qFa\right)/kT\right] \quad (7-6)$$

Thus, the velocity of the vacancy can be written as

$$\begin{aligned} v &= a(w'' - w') \\ &= av_o \exp\left(-\frac{E_i}{kT}\right) \left[2 \sinh\left(\frac{qFa}{2kT}\right)\right] \end{aligned} \quad (7-7)$$

In the case of low fields, $qFa \ll kT$, the last term $\sinh(qFa/2kT)$ can be approximately equal to $qFa/2kT$ (i.e., the first term of its expansion series). Then Equation 7-7 can be simplified to

$$v = \frac{a^2 q v_o F}{kT} \exp\left(-\frac{E_i}{kT}\right) \quad (7-8)$$

Hence, the mobility can be expressed as

$$\mu_+ = \frac{a^2 q v_o}{kT} \exp\left(-\frac{E_i}{kT}\right) \quad (7-9)$$

For a three-dimensional NaCl lattice, a central cation vacancy can jump to any of the 12 neighboring cation sites; each has a distance $\sqrt{2}a$ from the vacancy. If the applied field is in 100 or any crystal direction, four possible jumping directions are perpendicular to the field, which will not contribute to the conductivity; four jumping directions are in the direction of the field; and the remaining four are in the direction opposite to the field. Taking this jumping probability into account, the mobility becomes

$$\mu_+ = \frac{4a^2 q v_o}{kT} \exp\left(-\frac{E_i}{kT}\right) \quad (7-10)$$

Since the vacancy is part of the Schottky defect, we can assume $n_+ = n_s$ where n_s is the equilibrium density of Schottky defects, which is given by

$$n_s \approx N \exp\left(-\frac{E_s}{2kT}\right) \quad (7-11)$$

where N is the density of the cations in the crystal and E_s is the Gibbs free energy for the formation of a pair of Schottky defects. Thus, the ionic conductivity of NaCl-type ionic crystals can be written as

$$\sigma_{\text{ion}} = \frac{BN(4a^2 q v_o)}{kT} \exp\left[-\left(\frac{E_i + E_s/2}{kT}\right)\right] \quad (7-12)$$

where B is a temperature-dependent constant taking into account the effect of the lattice vibration. By letting

$$(\sigma_{\text{ion}})_o = \frac{BN(4a^2 q v_o)}{kT}$$

and

$$E_\sigma = E_i + \frac{E_s}{2}$$

Equation 7-12 can be simplified to

$$\sigma_{\text{ion}} = (\sigma_{\text{ion}})_o \exp\left(-\frac{E_\sigma}{kT}\right) \quad (7-13)$$

which is similar to Equation 7-1. It can be seen that $(\sigma_{\text{ion}})_o$ is temperature dependent. B also varies with temperature, which makes $(\sigma_{\text{ion}})_o$ relatively less sensitive to temperature. It should be noted that this simple analysis for intrinsic ionic conduction is based on an ideal model for binary ionic crystals such as NaCl and KCl.

Vacancies might also be introduced into the crystal by dopants. For example, a small amount of SrCl_2 doped into NaCl would introduce Na vacancies for the Sr ions. In this case, the ionic conductivity depends on the dopant concentration and the temperature. Typical $\sigma_{\text{ion}} - T$ characteristics are shown schematically in Figure 7-3. The slope in the high-temperature region (intrinsic) reflects the activation energy E_σ , which consists of the energy required for the creation of vacancies and that required for the movement of the vacancies.

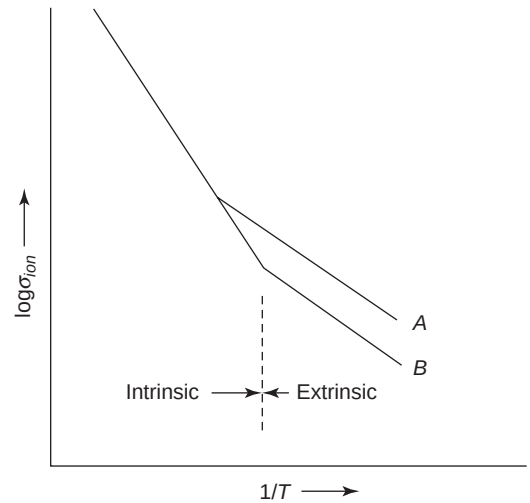


Figure 7-3 Schematic illustration of the ionic conductivity - temperature ($\sigma_{\text{ion}} - T$) characteristics of ionic crystals. Dopant concentration for A is larger than that for B.

The slope in the low-temperature region (extrinsic) reflects the energy E_i required mainly for the movement of the vacancies.

7.2.2 Extrinsic Ionic Conduction

The charge carriers for intrinsic ionic conduction are the vacancies created in the lattice sites unoccupied by the host ions such as Na^+ in the NaCl lattice. This occurs only in pure ionic crystals. However, in nonionic solids, ionic conduction is extrinsic, depending almost entirely on the nature and concentration of ionic impurities. Insulating materials such as polymers may contain ions (e.g., ionomers, polyelectrolytes), groups capable of ionizing, or groups into which ionic materials have been introduced. In most insulating materials, water is always a source of ions. It is generally not easy to identify the ions experimentally. For insulating polymers, it is reasonable to assume that they are derived mainly from fragments of polymerization catalyst, degradation and dissociation products of the polymer itself, absorbed water, and other impurities introduced into the polymer during fabrication processes. For example, insulating polymers such as PVC may contain H_3O^+ , Na^+ , K^+ , OH^- , Br^- , etc.

A molecule AB can be dissociated into A^+ and B^- ions:



In thermal equilibrium, the concentration of ions n_k can be expressed as

$$n_k = [A^+] = [B^-] \quad (7-15)$$

An equilibrium dissociation constant K_d , according to the mass action law, is defined as

$$K_d = \frac{[A^+][B^-]}{[AB]} = \frac{n_k^2}{(1-b)[AB]_0} \quad (7-16)$$

$$= \frac{b^2[AB]_0}{1-b}$$

where $[AB]_0$ and $[AB]$ are, respectively, the concentrations of the total original molecules and the nondissociated molecules and b is the fraction of the molecules that are dissociated. The dissociation is a thermally activated process, and therefore K_d can be written as

$$K_d = K_{d0} \exp(-E_d/\varepsilon_r kT) \quad (7-17)$$

where K_{d0} is a constant, E_d is the activation energy for dissociation, and ε_r is the dielectric constant. It can be seen from Equations 7-16 and 7-17 that K_d increases very fast with increasing dielectric constant.

If AB molecules are the only dissociable species present in the material, then their concentration can be expressed as

$$N_o = [AB]_o \quad (7-18)$$

and the extrinsic ionic conductivity can be written as

$$\sigma_{\text{ion}} = qbN_o(u_- + u_+) \quad (7-19)$$

Since b increases with increasing K_d , the dielectric constant plays a very important role in ionic conductivity. This is also why water absorbed by the material would cause a great increase in ionic conductivity.

7.2.3 Effects of Ionic Conduction

The positively charged cations moving toward the cathode and the negatively charged anions moving toward the anode under an applied electric field will create hetero-space charges near the electrodes. If the charges of the ions are not neutralized at the electrodes, they will accumulate there. These hetero-space charges may trigger the injection of electrons from the cathode or holes from the anode when the accumulated amount of these charges reaches a certain critical value.

In general, ions, when arriving at the electrodes, are not neutralized. They may react chemically with the electrode material and produce deposits that may alter the interface behavior at the metal-material contacts. Furthermore, because of the formation of hetero-space charges, the field distribution between electrodes becomes very nonuniform. The transport of ionic mass in the material will also change the spatial distribution of the dielectric properties, resulting in the formation of a multilayer capacitor. Based on the simplest Maxwell-Wagner two-layer capacitor model, the behavior of such a capacitor under AC fields

has been analyzed. It was found that both the overall dielectric constant and the conductivity depend on the frequency, the relative difference in layer thickness, the dielectric constant, and the conductivity between the two layers.⁸

It is well known that the charging current decays with time immediately after the application of a step-function voltage across an insulating material between two metallic electrodes, as shown in Figure 7-4. We refer to this metal-insulating material-metal structure as the MIM system. This current decay phenomenon occurs in all insulating materials at any applied electric fields, low or high. In this section, we will consider only cases at low fields. There are three possible mechanisms that may be responsible for this phenomenon. These mechanisms are ionic conduction, dipolar polarization, and electronic conduction.

It is generally accepted that the current decay phenomenon is associated with ionic conduction. Obviously, for materials having a high dielectric constant or containing ionic impurities, this phenomenon is due to ionic conduction. At low applied fields, the electrical conduction current in insulating materials is usually so small

that determining the type of charge carrier responsible for the conduction is very difficult. Some investigators have used the conductivity's pressure dependence to determine the type of charge carrier in insulating polymers.⁹ If the conductivity decreases with increasing pressure, the conduction is ionic because of the decrease in free volume, which is required for the movement of ions. If the conduction is electronic, the conductivity should increase with increasing pressure because of increasing overlap of wave functions. For nonpolymeric hard insulating materials, the interpretation of the electrical conductivity's pressure dependence may not be straightforward.

However, for good nondipolar insulating materials, it is unlikely that ionic conduction would make a noticeable contribution to the total conduction current, in either the transient or the steady-state region, for the following reasons:

- The dissociation at room temperature for materials with a low dielectric constant cannot produce sufficient ions for significant ionic conduction, particularly when the amount of ionic impurities can be reduced to a negligible level through careful preparation and handling processes.
- The hetero-space charges formed near the electrodes produced by ionic conduction may trigger the injection of electrons from the cathode and holes from the anode when the amount of accumulated space charges reaches a certain critical level. In this case, we would expect the conduction current to increase after the transient current decay period has elapsed. But such an expected phenomenon does not occur, implying that the concentration of hetero-space charges is either zero or very small and that the ionic conduction component, if any, is negligibly small compared to the dominant electronic conduction component.
- For ionic conduction, we would expect that electrolytic products would be accumulated near the electrodes as the ions arrive there. So far, such products have not been detected in good insulating polymers, indicating that

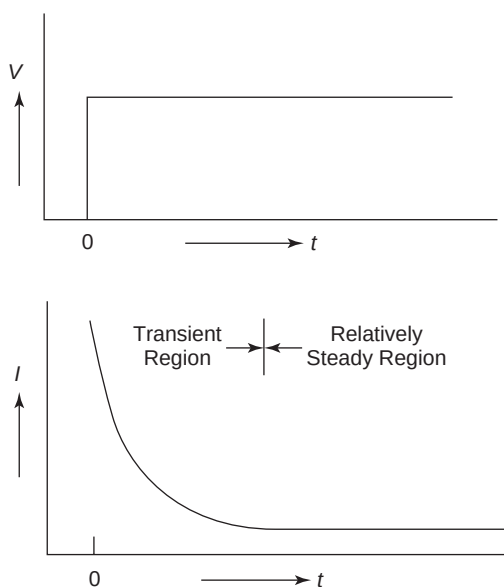


Figure 7-4 Decay of charging current with time after the application of a step-function DC voltage across the MIM system.

such products may not exist or may be too few to be detected. This also means that the ionic conduction component, if any, would be negligibly small.

- It is difficult to imagine how ions can be replenished continuously to the solid specimen to maintain the steady current flow after the transient current period has elapsed.

Regarding the second possible mechanism, dipolar polarization, we have the following comments. The time involved in the charging current decay region is usually longer than a few tens of seconds. If the current decay transient is due to the polarization (i.e., due to the displacement current and not the conduction current), then the polarization must be due to the orientation of dipoles available in the material. Good insulating polymers generally have a low dielectric constant (of the order of 3), implying that they are mainly nondipolar. Of course, dipolar impurities such as moisture (H_2O) and other foreign impurities may be introduced into the polymers during fabrication processes. However, it is not expected that the concentration of such impurities would be large enough to contribute significantly to this transient phenomenon. Otherwise, their presence would significantly enhance both the dielectric constant and the conductivity, but this is not the case. Furthermore, dipolar polarization may induce a current transient, but would not produce space charges, implying that the specimen should remain neutral. In insulating polymers, space charges are always formed near the electrodes after the sample is subjected to electrical stressing for a period of time. On the basis of these facts, it is very unlikely that dipolar polarization, if there were any, would play a noticeable role in the current transient phenomenon.

The following section discusses the third possible mechanism: electronic conduction.

7.3 Electronic Conduction

Insulating materials generally have a narrow energy bandwidth, a large energy band gap

(>5 eV), a large carrier effective mass, a large concentration of various localized states in the band gap, and hence a very low conductivity at low fields. Most insulating polymers are non-crystalline in structure and nondipolar. Considering an ideal perfect crystal, the intrinsic concentrations of electrons n and of holes p are equal, and n is about $1.5 \times 10^{10} \text{ cm}^{-3}$ for a band gap of 1.1 eV, as in silicon. The intrinsic carrier concentration is given by

$$n = p = n_i = (N_c N_v)^{1/2} \exp(-E_g/2kT) \quad (7-20)$$

where N_c and N_v are the effective densities of states in the conduction and the valence bands, respectively, E_g is the energy band gap, and k and T are the Boltzmann constant and the absolute temperature, respectively. For the same ideal perfect crystal, with E_g changed from 1.1 eV to 5.0 eV (other parameters remaining unaltered), n_i will reduce to $3 \times 10^{-23} \text{ cm}^{-3}$, according to Equation 7-20. This means that there is less than one electron per cm^3 . The electrical conductivity is defined as

$$\sigma = qu_n n \quad (7-21)$$

where u_n is the electron mobility and q is the electronic charge. Based on this equation, the intrinsic electrical conductivity would be negligibly small compared to the finite values of σ for all insulating materials reported in the literature,^{3,10,11} the conductivity of polyethylene, for example, is about $10^{-17} (\Omega\text{-cm})^{-1}$. This simple calculation tells us that for $E_g \geq 5 \text{ eV}$, the electrical conduction is extrinsic and the intrinsic contribution can be entirely ruled out. It is interesting to see what carrier concentration is required to give an electrical conductivity of $10^{-17} (\Omega\text{-cm})^{-1}$. Taking $u_n = 10^{-10} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ for polyethylene, the carrier concentration would be about $6 \times 10^{11} \text{ cm}^{-3}$, based on Equation 7-21. Now, the question arises as to what these carrier species are and where they come from. There are two major types of carrier species: ions and electrons (or holes). The previous section mentioned that ions are not the species responsible for electrical conduction.

Tahira and Kao¹⁰ were the first to identify experimentally the charge carrier species and the mechanism responsible for the charging

current decay transient in polyethylene. They used a simple experimental arrangement, shown in Figure 7-5, and an MIM sandwich configuration with one electrode made of a semitransparent thin gold film of about 200 Å in thickness. The applied step-function DC field

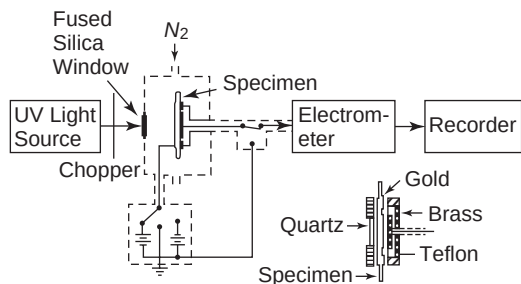


Figure 7-5 Experimental arrangement for the measurement of photocurrent superimposed on the charging current.

was 60 kV cm^{-1} . The ultraviolet (UV) light was produced by a 180 W deuterium lamp, and the illuminating beam was chopped to produce rectangular light pulses with a rise-and-fall time of about 10^{-4} s . Both the charging current J and the superimposed photocurrent J_{ph} were measured simultaneously after the application of a step-function DC field and a series of light pulses to illuminate the specimen through the gold illuminated electrode. Typical results, given in Figure 7-6, show that the photocurrent depends strongly on the magnitude of the charging current. The total current is

$$J_T = J + J_{ph} \quad (7-22)$$

The photocurrent decays much faster when the illuminated electrode is negatively biased than when it is positively biased, indicating clearly that the lifetime of the photogenerated carriers is shorter for the former than for the latter case.

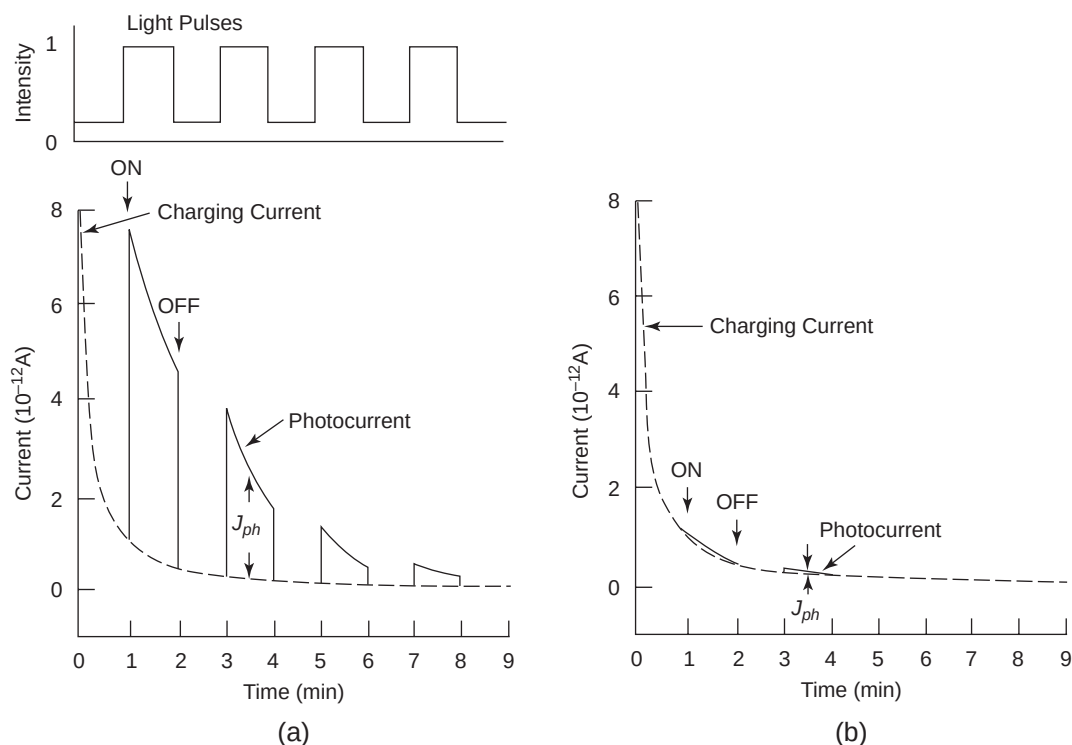


Figure 7-6 Photocurrent superimposed on the charging current under exciting radiation by a series of rectangular UV pulses for an illuminated electrode (a) positively biased and (b) negatively biased at $F = 60 \text{ kV cm}^{-1}$ for low-density polyethylene.

It is most likely that the probability for photo-generation of free holes is much higher than for free electrons by UV light of wavelengths ranging from 200 nm to 300 nm in polyethylene with a band gap of about 8.8 eV. The electrons injected from the injecting contact are quickly captured by the acceptorlike traps, forming a trapped electron space charge there. These trapped electrons behave as hole traps, tending to capture the photogenerated free holes. In the region near the electron-injecting contact, there are plenty of trapped electrons, which act as deep hole traps. Thus, in this region, the rate of photogeneration of free holes is smaller than the rate of annihilation of holes; in other words, the hole lifetime is shorter. This is why the photocurrent decays very rapidly with time. When the anode is illuminated, the photocurrent decays slowly with time, indicating that the holes have a longer lifetime. In that region near the anode, the major traps are the originally existing donorlike hole traps.

This experiment provides the following important information:

- The concentration of the originally existing acceptorlike electron traps is higher than that of the originally existing donorlike hole traps.
- After being filled by electrons, the acceptorlike traps act as deep hole traps, and their concentration near the electron-injecting contact is much higher than that of the originally existing donorlike hole traps.
- The charging current decay phenomenon is due to the electron trap-filling process.
- At low fields, electron injection from the contact is predominant because the potential barrier height for electron injection is lower than that for hole injection. Thus, hole injection can be ignored.

The trapped electrons near the injecting contact create a homo-space charge and hence an internal field opposite to the applied field, reducing the effective field for electron injection from the contact. This is why the charging current decays with time. After the transient decay period, the conduction current will finally reach

a quasi-steady state value. This implies that the electrical conduction has reached a dynamic equilibrium condition under which the rate of electrons being trapped is approximately equal to the rate of trapped electrons being thermally detrapped (i.e., thermal emission of electrons from traps), and that the contact injects just enough electrons to replenish those lost at the anode. In insulating polymers, the trapped electrons are highly concentrated near the injecting contact because of very low electron mobility. Thus, beyond the space charge region, the remaining portion of the specimen may be considered free of space charge. Therefore, the field in the remaining portion may be assumed to increase linearly with increasing applied field. This may be the reason why, at low fields, the quasi-steady state conduction current follows Ohm's law closely. Similar phenomena have also been observed in polypropylene and polyimide. Based on this argument, it can be concluded that at low fields, electrical conduction in both the transient and the quasi-steady state regions is predominantly electronic in nature.

7.3.1 Electrical Transport

There are several mechanisms that may account for electrical transport in dielectric solids. We shall discuss each of them briefly.

Band Conduction

The most important result of applying quantum mechanics to electronic properties of solids is that electrons are allowed only in certain energy levels grouped in bands, separated by energy gaps, which are generally called *forbidden gaps* or *band gaps*. Based on the crystal structure with a perfectly periodic lattice, an electron can propagate freely through the lattice as a wave characterized by a wave vector k . So, an electron behaves as a particle as well as a wave, and its behavior is governed by the wave vector

$$k = \frac{2\pi}{\lambda} \quad (7-23)$$

the energy

$$E = h\nu = \frac{h\nu}{\lambda} \quad (7-24)$$

and the momentum

$$p = m_n^* v \quad (7-25)$$

where λ is the de Broglie wavelength, ν is the frequency, v is the velocity, m_n^* is the effective mass of the electron, and h is the Planck constant. Based on wave mechanics, the momentum is expressed as

$$p = \frac{h}{\lambda} = \hbar k \quad (7-26)$$

where $\hbar$ is equal to $h/2\pi$. Thus, from Equations 7-23 through 7-26, the relation between the energy and the momentum of the electron can be written as

$$E = \frac{\hbar^2 k^2}{2m_n^*} \quad (7-27)$$

Each atom has its own discrete energy levels in the s , p , d , f , . . . orbitals, but when many atoms are brought together to form a solid, the electron charges in each orbital will overlap between adjacent atoms, and the energy levels of isolated atoms will split to form energy bands.

Since electron waves move in the structure with a periodic potential, the interaction between the electrons and the periodic potential gives rise to the formation of a band structure.¹² The traveling electron waves undergo constructive and destructive interference as they interact with the ion cores of the lattice. This interaction leads to the creation of energy band gap, separating bands of allowed propagation. This is analogous to an electrical filter with stop bands and pass bands. When the electron waves moving perpendicular to the lattice planes have the wave vectors $k = 2\pi/\lambda = n\pi/a$ with $n = 1, 2, 3, \dots$, Bragg reflection occurs at the lattice planes, so the wave cannot propagate. The condition for Bragg reflection is

$$2a = n\pi$$

or

$$k = \frac{n\pi}{a} \quad (7-28)$$

For example, a silicon atom has an electronic structure $1s^2 2s^2 2p^6 3s^2 3p^2$ and a carbon atom has an electronic structure $1s^2 2s^2 2p^2$. Both atoms are similar; there are eight available quantum states in the outermost s and p orbital levels, but only four of these are filled for bonding. A diamond (crystalline C) has N atoms and $6N$ electrons per unit volume. In the $2p$ orbital, there are six quantum states, but only two quantum states are filled. As the interatomic spacing decreases, these energy levels split into bands, as shown in Figure 7-7. As the $2s$ and $2p$ bands grow, they merge into a single band composed of a mixture of energy levels. This band of mixing $2s - 2p$ levels contains $8N$ available states but only $4N$ are filled. As the interatomic spacing approaches the equilibrium value (i.e., the lattice constant of diamond), this band splits into two bands separated by an energy gap E_g . The upper band, containing $4N$ states of antibonding, is called the *conduction band*. The lower band, also containing $4N$ states of sp^3 bonding, is called the *valence band*. At $T = 0$, all states in the valence band are occupied by electrons and all states in the conduction band are empty. But at $T > 0$, some electrons at the top of the valence band make a transition to the bottom of the conduction band and occupy some states there, as shown in Figure 7-8.

It should be noted that Equation 7-27 was originally derived for Sommerfeld's free electron model for metals. To take into account the interaction between the electron and the periodic potential, the effective mass m^* (instead of the rest mass m) is used for electrons. Also, Equation 7-27 is for a simple, spherical, constant energy surface. For an ellipsoidal constant energy surface, the effective masses along the longitudinal and the transverse directions are different and Equation 7-27 is no longer valid. In fact, for many semiconductors, such as silicon and germanium, the constant energy surfaces are not spherical and the E - k relation is not parabolic.¹³ However, we use Equation 7-27 for the sake of simplicity because this book emphasizes the physical concepts rather than detailed mathematical analysis. In fact, the calculation of the energy band structure is quite mathematically involved, and for most dielec-

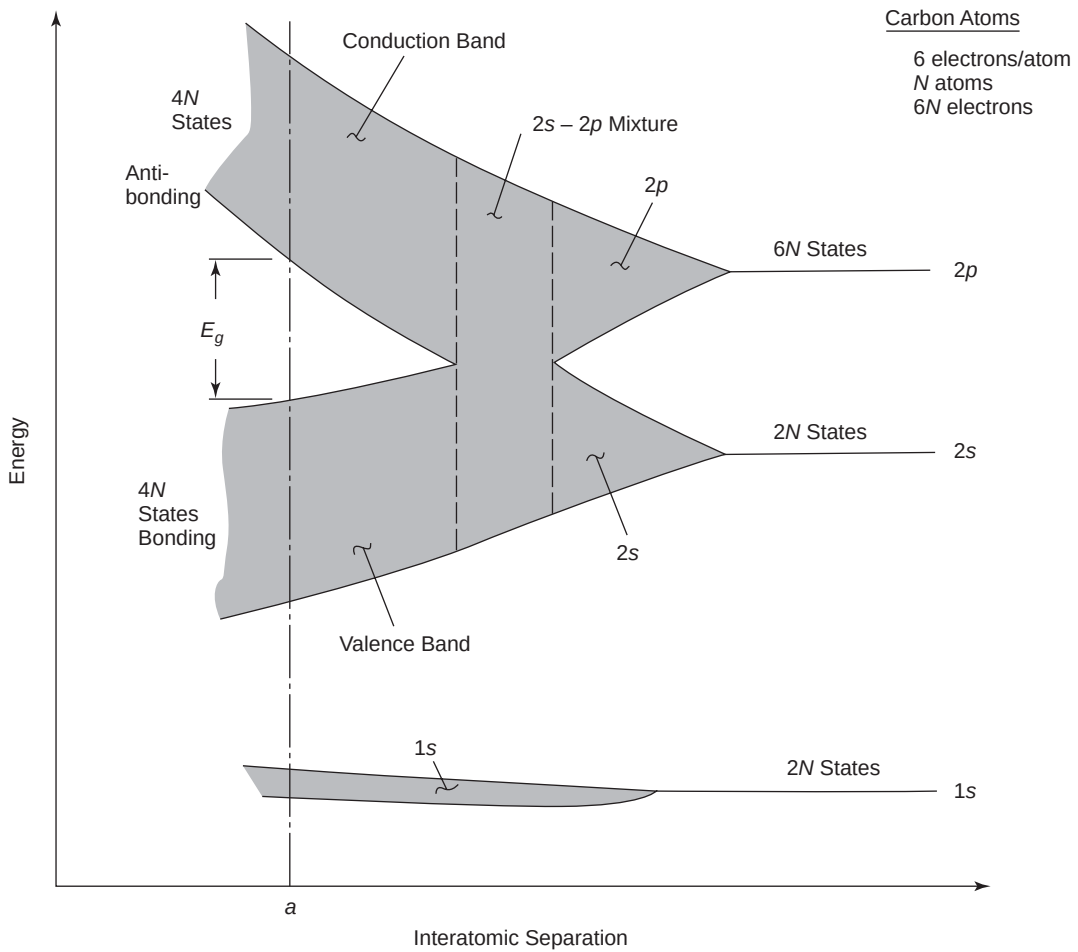


Figure 7-7 Formation of energy bands in a diamond crystal formed by putting together isolated carbon atoms. The thermal-equilibrium interatomic separation is a (i.e., the lattice constant).

tric solids information about the band structure is still lacking. So a detailed discussion of band structure is beyond the scope of this chapter.

Two typical types of band structure are shown, however, in Figure 7-8, in which the various symmetry points mean that at the Brillouin zone Γ : $2\pi/a$ (0, 0, 0) is at the zone center; L: $2\pi/a$ (1/2, 1/2, 1/2) is at the zone edge along (111) axes, and X: $2\pi/a$ (0, 0, 1) is at the zone edge along (100) axes in the first Brillouin zone.¹⁴ In general, the structure is more complicated than in Figure 7-8, but this simplified structure gives a clear picture of the basic features. The energy difference between the

bottom of the conduction band (termed the conduction band edge E_c) and the top of the valence band (termed the valence band edge E_v) is the forbidden gap E_g . For silicon, E_c occurs at $k = 0$, but E_v occurs at $k = k_c$ along the [100] direction, implying that the crystal momentum $\hbar k$ is different from the particle momentum mv . See Figure 7-8(a). In this case, the crystal momentum is $\hbar k$, but the particle momentum is zero when the kinetic energy of the electron is zero at E_c . Thus, for silicon, when an electron makes a transition from the valence band to the conduction band, it requires not only an energy E_g , but also a momentum to change its momentum.

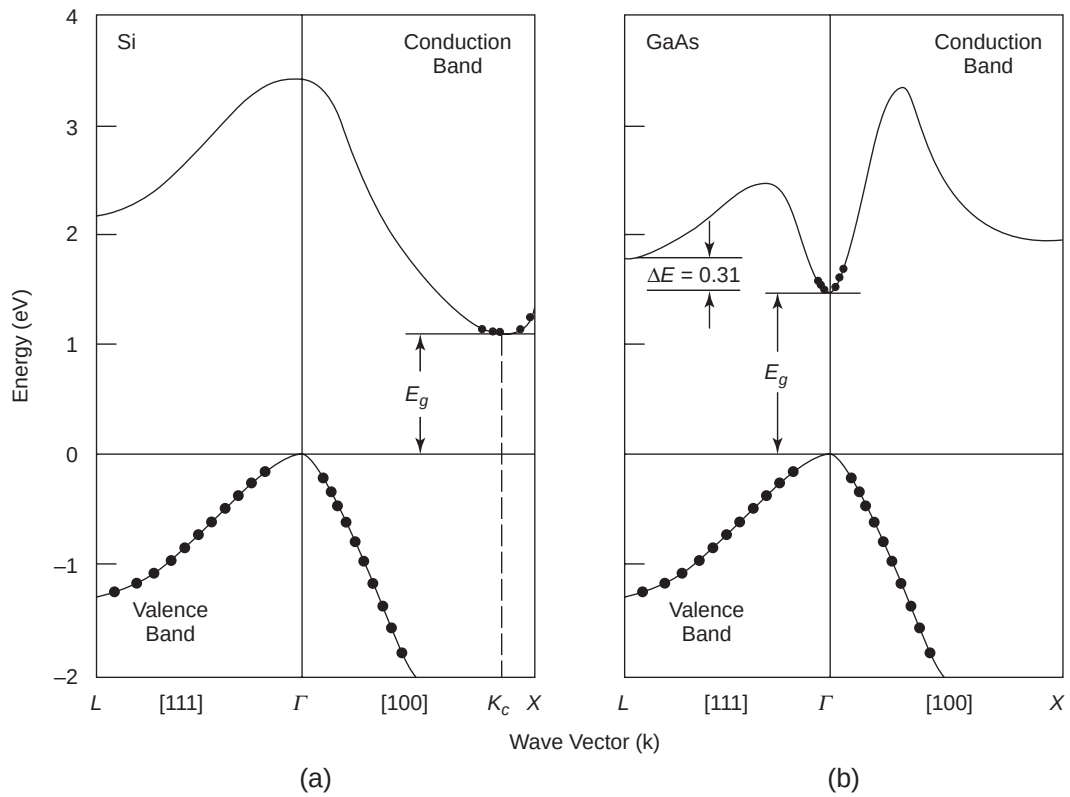


Figure 7-8 Energy band structures of (a) Si (indirect band gap) and (b) GaAs (direct band gap). Dots (●) indicate electrons in the conduction band. The empty states (without dots) on the top of the valence band are holes at $T > 0$.

Momentum is conserved via a phonon interaction. A phonon is a quantum of lattice vibration and has a characteristic energy E_{phonon} . So a transition from E_v to E_c involves either phonon emission or phonon absorption. If the transition is caused by optical excitation, the photon provides an energy $h\nu$ but cannot provide a momentum. Therefore, if the photon energy $h\nu \leq E_g$, the transition will involve phonon absorption. If $h\nu > E_g$, the transition will involve phonon emission. Silicon is an indirect band-gap semiconductor. Elementary semiconductors, including silicon, germanium, etc., belong to this category.

If E_c and E_v occur at the same value of k , as in GaAs in Figure 7-8(b), the transition does not require a change in crystal momentum, so GaAs is a direct gap semiconductor. Most III–V compound semiconductors, such as GaAs, InP,

InAs, and InSb, belong to this category. Direct gap semiconductors have a direct optical transition, small electron effective masses, and high electron mobilities. That is why these materials are widely used in light-emitting diodes, semiconductor lasers, and high-speed or high-frequency devices.

In general, an electron can move freely in a field-free space or in a space with a constant potential field because there is no force acting on the moving electron. If a voltage is applied between two electrodes, producing an electric field F in the space, then the electron will experience a force qF , driving it to move along the field direction. In this case, the velocity of the electron is not constant at a constant F but increases with time until it reaches the positive electrode. Thus, the velocity of an electron moving in a lattice with a periodic potential

field would fall and rise and fall again. In this case, the instantaneous velocity has no meaning; we must use the mean velocity over a distance many times the atomic spacings.

An electron can move without being scattered if its energy E and momentum k lie in the allowed ranges and if the potential field is strictly periodic.¹² As it moves, an electron is acted upon by the periodically varying field. If we average the parameters describing the motion over several periods, we can describe the motion of the electron due to an externally applied electric field or magnetic field by equations of the same form as those for an electron moving in free space, provided that the effective mass m^* for the electron (instead of its rest mass) is used. For steady-state motion, the mean velocity of the electron can be written as

$$\bar{v} = \frac{1}{\hbar} \frac{dE}{dk} \quad (7-29)$$

and the acceleration as

$$\begin{aligned} \frac{d\bar{v}}{dt} &= \frac{1}{\hbar} \frac{d}{dt} \left(\frac{dE}{dk} \right) = \frac{1}{\hbar} \frac{d}{dk} \left(\frac{dE}{dk} \right) \frac{dk}{dt} \\ &= \mathcal{F} \frac{1}{\hbar^2} \frac{d^2 E}{dk^2} \end{aligned} \quad (7-30)$$

since $dk/dt = \mathcal{F}$ where $\mathcal{F}$ is the force acting on the electron, which can be written as

$$\mathcal{F} = m^* \frac{d\bar{v}}{dt} \quad (7-31)$$

So the effective mass of the electron m^* is

$$m^* = \hbar^2 \left(\frac{d^2 E}{dk^2} \right)^{-1} \quad (7-32)$$

Equation (7-32) indicates that the narrower the parabola is, the smaller the effective mass. For example, GaAs has a narrow conduction-band parabola, its electron effective mass $m^* = 0.07m$. Silicon has a wider conduction-band parabola, $m^* = 0.19m$. Obviously, m^* depends on the crystal direction, since the $E - k$ relation is direction dependent.

The width of the allowed energy bands depends on the shell to which the electrons belong. The interaction between the K-shells of individual atoms is small because electrons in the K-shell are firmly bound to the nucleus, so

the width of the K-band is extremely small. The less firmly bound the electrons, the wider the allowed energy bands become.

The basic difference between semiconductors and insulators is mainly the difference between their energy band gaps. Figure 7-9 illustrates schematically the basic differences among metals, semiconductors, and insulators. For metals, either the upper band is partially filled or the upper band overlaps the nearly full lower band. For semiconductors, $E_g < 3$ eV, as in Ge, Si, and GaAs, whose E_g are, respectively, 0.67, 1.10, and 1.43 eV. For insulators, $E_g > 3$ eV, as in diamond and SiO₂, whose E_g are, respectively, 5.0 and 9.0 eV. Usually, the larger the energy band gap, the narrower the conduction bandwidth becomes. The energy band gap is temperature dependent. The higher the temperature, the more severe the lattice vibration and the easier it is to break the bond. This is equivalent to saying that E_g decreases with increasing temperature.

In short, band conduction is due to the movement of electrons in the conduction band or holes in the valence band, governed by the following parameters.¹⁵⁻¹⁷

Density of Quantum States

The density of states in the conduction band is the number of states in the conduction band per unit volume per unit energy at E above E_c , which is given by

$$\begin{aligned} N(E) &= \frac{1}{2\pi^2} \left(\frac{2m_n^*}{\hbar^2} \right)^{3/2} (E - E_c)^{1/2} \\ &= 4\pi \left(\frac{2m_n^*}{\hbar^2} \right)^{3/2} (E - E_c)^{1/2} \end{aligned} \quad (7-33)$$

The density of states in the valence band is the number of states in the valence band per unit volume per unit energy at E below E_v , which is given by

$$\begin{aligned} N(E) &= \frac{1}{2\pi^2} \left(\frac{2m_p^*}{\hbar^2} \right)^{3/2} (E_v - E)^{1/2} \\ &= 4\pi \left(\frac{2m_p^*}{\hbar^2} \right)^{3/2} (E_v - E)^{1/2} \end{aligned} \quad (7-34)$$

where m_n^* and m_p^* are, respectively, the effective masses of electron and hole. The density of

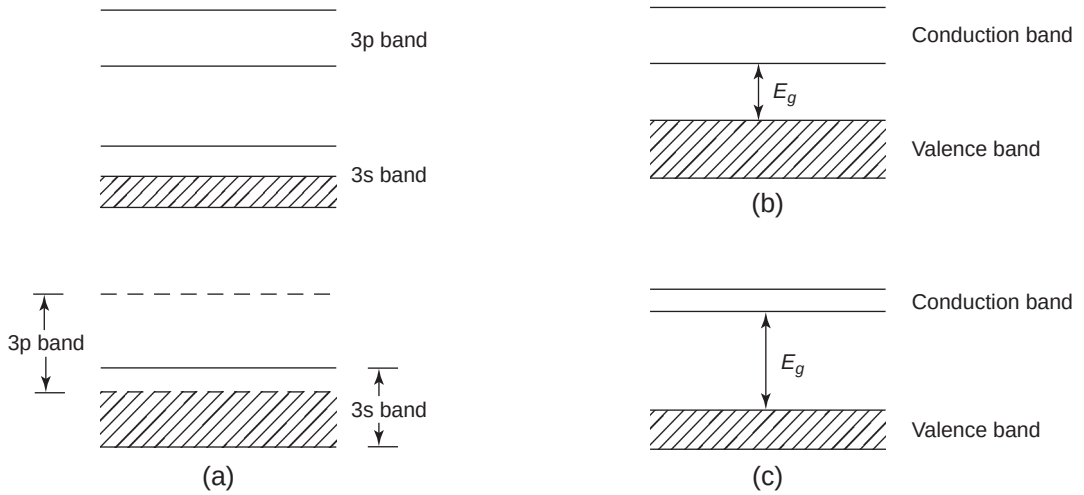


Figure 7-9 Schematic illustration of the basic difference among metals, semiconductors, and insulators: (a) metals such as sodium with a half-full 3s band and an empty 3p band, or such as magnesium with the overlapping of the upper 3p band and the lower 3s band, leaving the 3s band partially unoccupied; (b) intrinsic semiconductors with a wide conduction band and a narrow energy band gap usually less than 3 eV such as E_g of Si = 1.1 eV and E_g of GaAs = 1.43 eV; and (c) insulators with a narrow conduction band and a wide energy band gap usually larger than 3 eV such as E_g of diamond = 5 eV and E_g of SiO₂ = 9 eV.

states as a function of energy is shown in Figure 7-10.

Fermi–Dirac Distribution Function

Electrons in crystalline solids follow the Fermi–Dirac statistics. The probability that an available quantum state at energy level E in the conduction band will be occupied by an electron at temperature T is given by

$$f(E) = \frac{1}{1 + \exp[(E - E_F)/kT]} \quad (7-35)$$

The probability that an occupied state at energy level E in the valence band will be empty (will create a hole) at T is given by

$$f_h(E) = 1 - f(E) = \frac{1}{1 + \exp[(E_F - E)/kT]} \quad (7-36)$$

where E_F is the Fermi level, which is a reference level. This implies that the probability that an empty state at ΔE above E_F will be occupied is equal to the probability that an occupied state at ΔE below E_F will be empty.

Sources of Charge Carriers

There are three ways to supply charge carriers, depending on the type of electronic conduction.

Intrinsic conduction—The carriers—electrons and holes—are generated in the material itself, usually by thermal excitation. The concentrations of electrons and holes are equal. Intrinsic carrier concentration as a function of temperature is given by Equation 7-20. Thus, the intrinsic conductivity can be written as

$$\begin{aligned} \sigma_i &= q(\mu_n n + \mu_p p) = qn_i(\mu_n + \mu_p) \\ &= q(\mu_n + \mu_p)(N_c N_v)^{1/2} \exp(-E_g/2kT) \end{aligned} \quad (7-37)$$

where N_c and N_v are, respectively, the effective densities of states (i.e., the number of states per unit volume) in the conduction and the valence bands, which are given by

$$N_c = 2 \left[\frac{2\pi m_n^* kT}{h^2} \right]^{3/2} \quad (7-38)$$

$$N_v = 2 \left[\frac{2\pi m_p^* kT}{h^2} \right]^{3/2} \quad (7-39)$$

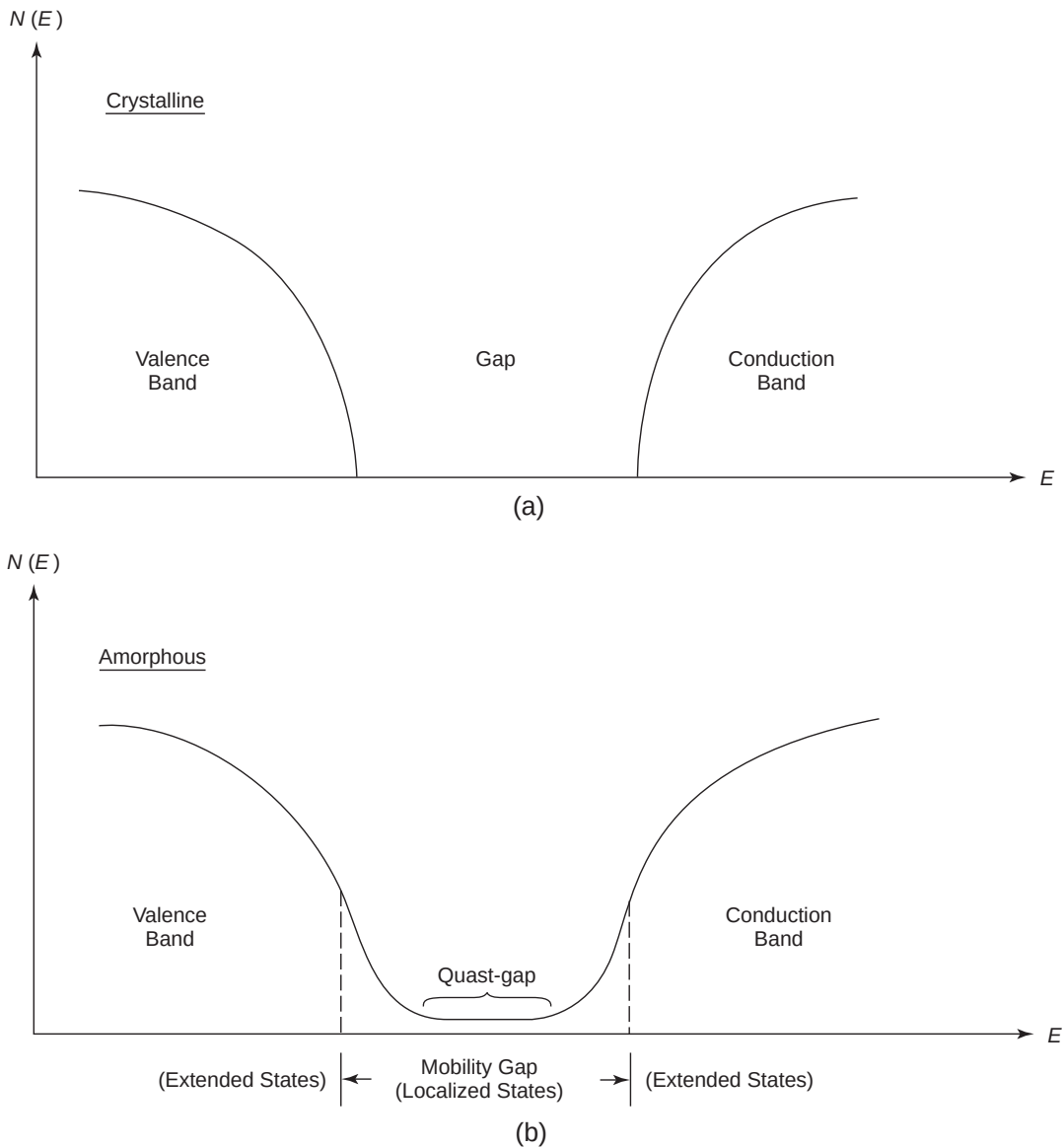


Figure 7-10 Schematic illustration of the density of states as a function of energy for (a) crystalline solid and (b) amorphous solid.

Extrinsic conduction—Even intrinsic silicon, which is supposed to be the material with the highest purity ever achieved by today's technology, still contains unavoidable impurities of a concentration of about 10^{11} cm^{-3} . This so-called intrinsic silicon is usually classified into two types: ν -type silicon (slightly n-type) and π -type silicon (slightly p-type), depending on

whether donor-type or acceptor-type impurities are predominant. At high temperatures, intrinsic conduction is dominant, but at low temperatures the n-type or p-type behavior will appear. Obviously, insulating materials such as polyethylene and SiO_2 would contain various impurities of a concentration higher than 10^{15} cm^{-3} , of which some certainly act as donors or as

acceptors. Since insulating materials have a large band gap, intrinsic conduction can be completely ruled out. If the electrical contacts were ohmic, an extrinsic conduction current would be observed.

For practical applications, semiconductors are usually doped with predetermined donor or acceptor impurities, which can contribute additional electrons or holes, respectively, to the material. A donor atom contains an additional electron beyond those required to bond the impurity to the semiconductor lattice. The bound states of such extra electrons lie in the band gap at the energy level E_d above the middle of the gap that is, $E_d > (E_c + E_v)/2$. The extra electrons are easily excited thermally into the conduction band, leaving behind positively charged ions. Similarly, an acceptor atom is deficient in one electron. The bound states of those missing electrons also lie in the gap at an energy level E_a below $(E_c + E_v)/2$, so electrons are easily excited thermally from the valence band to the bound states of the acceptors, creating holes in the valence band and leaving behind negatively charged ions.

If N_D and N_A are, respectively, the concentrations of donors and acceptors, and N_D^+ and N_A^- are, respectively, the concentrations of ionized donors and acceptors, then for charge neutrality we can write

$$n + N_A^- = p + N_D^+ \quad (7-40)$$

The Fermi level always adjusts itself to maintain charge neutrality in the material. Thus, we can write

$$n = N_c \exp\left[\frac{-(E_c - E_F)}{kT}\right] = n_i \exp\left[\frac{E_F - E_i}{kT}\right] \quad (7-41)$$

$$p = N_v \exp\left[\frac{-(E_F - E_v)}{kT}\right] = n_i \exp\left[\frac{E_i - E_F}{kT}\right] \quad (7-42)$$

$$N_A^- = f(E_a)N_A = \frac{N_A}{1 + \exp[(E_a - E_F)/kT]} \quad (7-43)$$

$$N_D^+ = [1 - f(E_d)]N_D = \frac{N_D}{1 + \exp[(E_F - E_d)/kT]} \quad (7-44)$$

where E_i is the intrinsic Fermi level. If $N_D \gg N_A$, implying that $(E_c - E_F) < (E_F - E_v)$, the semiconductor is of the n-type and its conductivity can be simplified to

$$\sigma = q\mu_n n_i \exp[(E_F - E_i)/kT] \quad (7-45)$$

Similarly, if $N_A \gg N_D$, implying that $(E_F - E_v) < (E_c - E_F)$, the semiconductor is of the p-type and its conductivity can be simplified to

$$\sigma = q\mu_p n_i \exp[(E_i - E_F)/kT] \quad (7-46)$$

For a nondegenerate semiconductor, in which $E_c - E_F > 4kT$ or $E_F - E_v > 4kT$, the product np is always equal to n_i^2 at a fixed temperature independent of the position of the Fermi level.

$$\begin{aligned} np &= n_i^2 = N_c N_v \exp(-E_g/kT) \\ &= 4 \left[\frac{2\pi k}{h^2} \right] (m_n^* m_p^*)^{3/2} T^2 \exp(-E_g/kT) \\ &= BT^3 \exp(-E_g/kT) \end{aligned} \quad (7-47)$$

This is generally referred to as the *mass action law*. It is usual for semiconductors to contain both donor and the acceptor impurities simultaneously. At the time the impurities are doped, the donor electrons immediately fill up any available acceptor bound states, since the crystal must attain the lowest possible energy state consistent with its temperature. For most purposes, the semiconductor can be considered to contain $N_A - N_D$ acceptors or $N_D - N_A$ donors, whichever is larger. Such semiconductors with $N_A \approx N_D$ with intrinsic behavior are said to be *compensated*. The electrical conductivity of the semiconductor containing both electrons and holes can be written as

$$\begin{aligned} \sigma &= q\mu_n n + q\mu_p p \\ &= q\mu_n n + q\mu_p \frac{n_i^2}{n} \end{aligned} \quad (7-48)$$

The electrical conductivity as a function of the ratio p/n is shown in Figure 7-11. The minimum of σ occurs when $d\sigma/dn = 0$, which leads to

$$\frac{p}{n} = \frac{\mu_n}{\mu_p} \quad (7-49)$$

For Si, $\mu_n > \mu_p$, so $\sigma_{\min}$ occurs at $p/n = \mu_n / \mu_p > 1$, but σ_i occurs at $p/n = 1$, which is slightly larger than $\sigma_{\min}$, as shown in Figure 7-11.

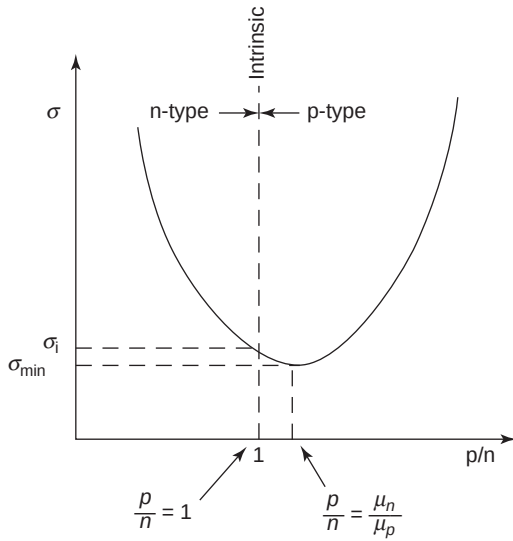


Figure 7-11 Schematic illustration of the conductivity of a semiconductor doped with donors and acceptors as a function of the ratio of p/n for $\mu_n > \mu_p$.

Injection-controlled conduction—The carriers are injected into the material mainly from metallic electrical contacts through a potential barrier at the metal–material interface. This subject has already been discussed in some detail in Charge Carrier Injection through Potential Barriers from Contacts in Chapter 6.

Carrier Drift Mobilities

Electrons in the conduction band or holes in the valence band have three degrees of freedom. Each degree for each electron has thermal energy $kT/2$. Thus, the kinetic energy of an electron may be written as

$$\frac{1}{2} m_n^* v_{th}^2 = \frac{3}{2} kT \quad (7-50)$$

So the electron's thermal velocity is

$$v_{th} = \left(\frac{3kT}{m_n^*} \right)^{1/2} \quad (7-51)$$

For semiconductors, v_{th} is of the order of 10^7 cm s^{-1} at 300K, and the direction of the moving electrons is random. However, under an applied electric field F , the electrons tend to move along the direction of the field, and the

net velocity of the electrons in the direction of the field is called the *drift velocity* v_n . The average distance between collisions is called the *mean free path* $\bar{\ell}$, and the average time between collisions is called the *mean free time* or *collision time* τ . Thus, the momentum of the drifting electron can be written as

$$p_n = m_n^* v_n = qF\tau \quad (7-52)$$

the electron drift velocity as

$$v_n = \frac{q\tau}{m_n^*} F$$

and the electron drift mobility as

$$\mu_n = \frac{v_n}{F} = \frac{q\tau}{m_n^*}$$

Similarly, the hole drift mobility can be written as

$$\mu_p = \frac{q\tau}{m_p^*}$$

The mean free time τ is due mainly to random collisions of the thermally activated electrons, so τ may be expressed as

$$\tau = \frac{\bar{\ell}}{v_{th}}$$

If $\bar{\ell}$ is 10^{-5} cm , $\tau = 10^{-5} / 10^7 = 10^{-12}$ second. The carrier mobility is controlled by $\bar{\ell}$ or τ , which is governed by the scattering mechanisms. The major scattering mechanisms follow.

Lattice scattering—The higher the temperature is, the more severe the lattice vibration and hence the more frequent the scattering (collisions) and the lower the mobility

$$\tau \propto T^{-3/2}$$

Thus,

$$\mu_n \text{ or } \mu_p \propto T^{-3/2}$$

Impurity scattering—The higher the temperature is, the less significant the impurity scattering and hence the higher the carrier mobility

$$\tau \propto T^{3/2}$$

$$\mu_n \text{ or } \mu_p \propto T^{3/2}$$

Also, mobility decreases with increasing concentration of impurities. In general, at high temperatures, lattice scattering is dominant, while at low temperatures impurity scattering becomes dominant. For silicon, lattice scattering is dominant at temperatures around and higher than 300 K.

Defect-Controlled Conduction

Most practical insulating materials are amorphous in nature. Amorphous structure implies the random arrangement of atoms or molecules and the absence of any periodic symmetry. It should be noted that completely random arrangement of atoms or molecules in gases is seldom found in condensed matter, even liquids. The basic difference between an amorphous and a crystalline solid lies in the fact that in the former, the atomic order is restricted to the nearest neighbors so that the atoms exhibit only short-range order because of the ever present binding force between neighbors; in the latter, the atomic order is a long-range one, exhibiting a periodic symmetry.

Short-range order results in the distribution of the density of electronic energy states tailing into the zone that is normally the forbidden zone. The electronic states in the tails are localized; hence, electrons in those states are localized. Beyond the tails are delocalized electronic states. The boundaries between the regions of localized and delocalized states are generally referred to as the conduction-band and valence-band edges, which are denoted by E_c and E_v , respectively, following band theory for crystalline solids. The energy gap between E_c and E_v is called the *mobility gap* and is shown in Figure 7-10(b).⁴

Electron and hole mobilities increase with increasing energy above E_c and below E_v , respectively. Thus, only when the electrons are excited to high electronic energy states in the delocalized region (also called the *extended region*), can appreciable electrical conduction occur. Obviously, electrons in localized states can move only by means of a thermally acti-

vated hopping process from one site across a potential barrier to a neighboring site, or by a tunneling process from one site through the barrier to the next. Both mechanisms may operate simultaneously. The relative importance of these two mechanisms depends on the profile of the potential barrier and the availability of thermal energy.

In general, polymers are different from so-called amorphous materials. A polymeric solid consists of an assembly of molecular chains, each of which also consists of many molecules. Within the chain, molecules are held together by covalent bonds and, in some cases, also by ionic bonds. Between the chains, however, only weak bonding exists, usually of a van der Waals type. A chain can be considered a large molecule (or macromolecule), which is made of many small units called monomers (or mers) repeatedly bonded together. Each unit can be thought of as a separate small molecule with electronic states associated with the molecular orbitals of other molecules. Molecular orbitals of the molecules would overlap, creating the bonding and antibonding states, which lead to the formation of the valence and the conduction bands, respectively. Energy band theory can be used to characterize the electrical properties of polymers.¹⁸

There are two main types of electronic transfer: intramolecular and intermolecular. For intramolecular transfer, electronic movement depends on the intramolecular bond and the bond of the individual monomer. For example, polyethylene has strong bonds between carbon atoms. All monomers are fully saturated and there is no significant overlap in the molecular orbital of each carbon atom with the molecular orbitals of carbon atoms on either side. In this case, we would expect the conduction band to be narrow and the forbidden gap is wide. If each carbon atom along the molecular chain had only one hydrogen atom instead of two, then each carbon atom would have a double bond and the atomic orbital of each would overlap significantly with those of carbon atoms on both sides, forming delocalized molecular orbitals. This type of chain structure is generally called a *conjugated chain*, in which

carbon atoms have alternating single and double bonds. In this case, we would expect the macromolecule to have a wider conduction band and a narrower forbidden gap, and the electronic conduction due to such an intramolecular transfer to be appreciable, as in benzene and anthracene.

While band theory may be used to characterize the electrical properties due to intramolecular transfer along the chains, the intermolecular transfer from one macromolecule to another may be the major stumbling block for electrical conduction. A polymer consists of many macromolecules, whose arrangement can be considered quite random and which are bonded mainly by weak van der Waals bonds. The whole polymer may be thought of as a mixture of crystalline and amorphous domains. The interface between a crystalline domain and an amorphous domain may behave just like a trapping region. Furthermore, molecular motion may also play a role in helping the macromolecules to make intimate contact with each other. It is likely that the boundaries between some macromolecules may behave like grain boundaries in polycrystalline materials.

Polymers have intrinsic defects, due to structural disorders, and extrinsic defects, due to chain end groups and foreign impurities left over from fabrication processes. Based on the arguments presented here, an insulating polymer consists of various kinds of defects which act as traps tending to capture charge carriers. We believe that electronic conduction is mainly a combination of intramolecular band conduction and intermolecular hopping conduction.

Electrical Transport by a Tunneling Process

An electron in a molecule, when excited to a higher energy level, can tunnel through a potential barrier to an unoccupied state in a neighboring molecule with energy conserved by a tunneling process,¹⁹ as shown in Figure 7-12. The electron in the excited state may tunnel to the neighboring molecule or return to its ground state. In general, however, the probability for

the former is much higher than for the latter, depending on the lifetime of the electron in the excited state.

In reality, the tunneling electron would experience a potential which is the sum of the approximate coulomb potential attracting the electron to a positive ion and the potential of electron affinity of the original neutral molecule.²⁰ These potentials vary gradually rather than abruptly and are therefore better approximated by a triangular potential barrier than by a square one (see Figure 7-12). Furthermore, the triangular shape facilitates intermolecular electron transfer, since the barrier width becomes smaller for the excited electron at a higher level.

In the band model for molecular crystals, there is no explicit mention of potential barrier between molecules. However, it can be imagined that an excited electron may tunnel over a distance of several molecules. Thus, the tunneling model may be considered a band model when the potential varies periodically and regularly throughout the crystal and the width of the potential barrier is less than 10 Å. Considering a two-potential well system with a single potential barrier to be the same as a system containing a great number of potential wells, Keller and Rast,²¹ have calculated the bandwidth of anthracene based on the energy level splitting to form a band, with the number of levels equal to the number of wells in the band. Their estimate of the bandwidth for anthracene is 0.029 eV, which is close to the value calculated on the basis of the band model.²²

Insulating or semiconducting films always contain impurities, which may be unavoidably present or may be deliberately doped in the film specimens. These impurities form impurity states in the forbidden energy gap. When the localized electronic wave functions of the impurity states overlap, an electron bound to one impurity state can tunnel to an unoccupied state without involving activation into the conduction band. This tunneling process between impurity sites is referred to as *impurity conduction*.²³ The mobility of an electron moving in the impurity states is very small (since it depends on interaction between widely spaced

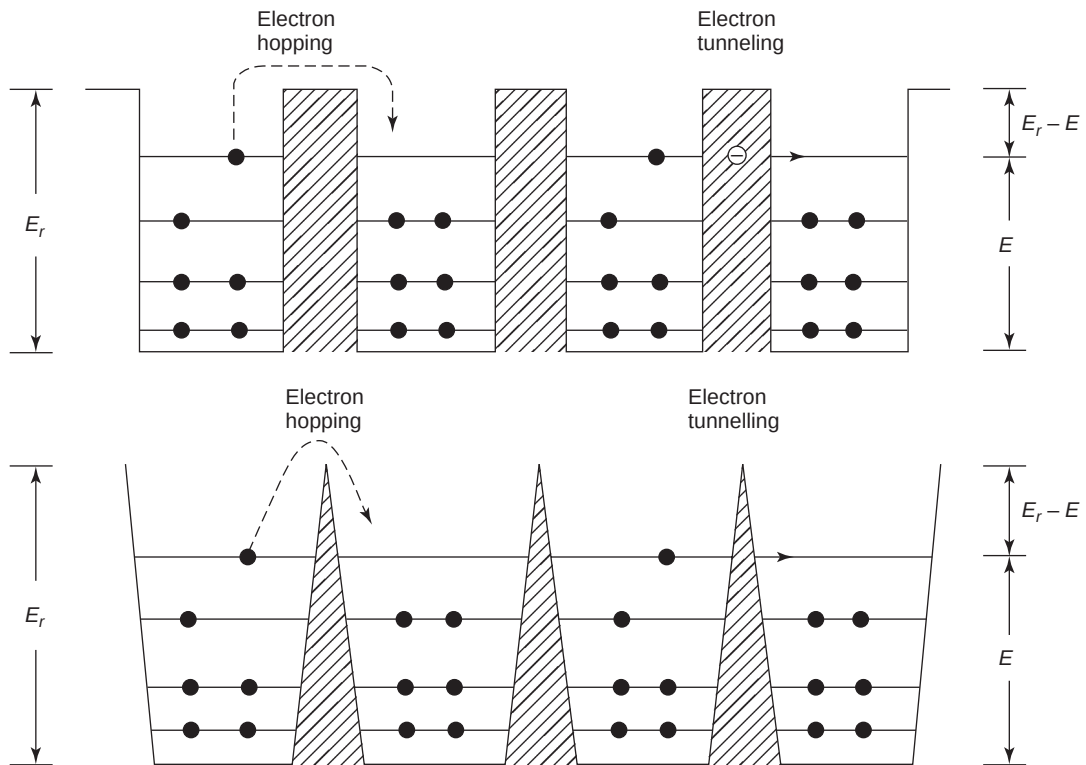


Figure 7-12 Schematic diagrams illustrating an electron hopping across and an electron tunneling through a square and a triangular potential barrier. The electron hopping or tunneling in one direction is equivalent to the hole hopping or tunneling in the opposite direction.

impurities), so this conduction mechanism usually becomes predominant at low temperatures due to low concentration of carriers in the conduction and valence bands. This conduction process depends, however, on impurity concentration and the energy levels of the impurity states, which control the probability of tunneling from impurity site to impurity site and the number of electrons taking part in this tunneling process.

In semiconductors, the impurity conduction process is possible only if the material is compensated (i.e., if the material contains both donor and acceptor impurities). This condition for impurity conduction was put forward by Mott²⁴ and Conwell²⁵ and confirmed experimentally by several investigators.^{26–28} For example, if the donor concentration N_D is larger than the acceptor concentration N_A in a compensated n-type semiconductor, all the accep-

tors will be occupied and become negatively charged; only $N_D - N_A$ donors will remain occupied and neutral at low temperatures. If the impurity state A and the impurity state B are at the same energy level, the overlap of wave functions between these two sites will enable the movement of an electron from an occupied to an empty donor site without involving activation into the conduction band. If the impurity state A is located at a lower energy level than the impurity state B , then thermal energy (phonon) supplied by lattice vibrations of the material is required to assist the electron tunneling from A to B . It should be noted that the field created by the charged acceptors and donors will split the energy levels of donor states. Therefore, an electron can tunnel from one impurity state to another only by exchanging energy with phonons. Also, the applied voltage will alter the energy level difference

between sites, thereby making the tunneling probability higher in one direction than in the other. A similar process can readily be realized in compensated p-type semiconductors ($N_A > N_D$), but in this case electrons tunnel through acceptor impurity sites.

The following are the most significant features due to impurity conduction observed in the resistivity–temperature characteristics:

- Resistivity is strongly dependent on impurity concentration.
- The plot of $\ln \rho$ versus $1/T$ exhibits a finite slope, indicating that a thermal activation energy is required for electron tunneling between sites when the impurity concentration is small.
- Activation energy decreases with increasing impurity concentration and becomes zero when the impurity concentration reaches a certain critical value or higher, indicating that for impurity concentrations higher than that critical value, carriers move freely without involving thermal activation.

Impurity conduction phenomenon has been observed in compensated p-type germanium,²⁸ tantalum oxide films,²⁹ silicon monoxide films,³⁰ nickel oxide,^{31,32} vanadium phosphate glasses,³³ and many other materials.^{4,23}

Electrical Transport by a Hopping Process

A localized electron can drift through a solid specimen by hopping from a molecule (or an atom) to a nonoccupied state of a neighboring molecule if it acquires the energy necessary to overcome the potential barrier. In general, the energy is from thermal excitation of the material. The concept of hopping transport has been familiar for a long time in connection with ionic conduction, since ions move essentially by hopping, whether through interstices or vacancies. This concept has been extended to electrons, particularly for electronic conduction in amorphous and disordered nonmetallic solids.^{4,34}

The probability of a hopping transition may be determined by both the distance between the two sites and the potential barrier that must be

overcome. If the potential barrier width (i.e., the distance between the two sites) is larger than $10 \text{ \AA}$, electrons hop rather than tunnel from one molecule to the neighboring molecule. In this case, the probability of hopping transition can be written as

$$W_H = \nu_j \exp(-\Delta E_j/kT) \quad (7-53)$$

where ΔE_j is the activation energy, which is $E_T - E$ (see Figure 7-12), and ν_j is the attempt-to-escape jump frequency. As with ionic conduction, the barrier height E_T depends on the applied electric field. Hence, the jump probability is higher for the jump across the lower barrier (i.e., the lower $\Delta E_j = E_T - E$).

In fact, the hopping process is similar to the atomic diffusion process, so hopping mobility follows the Einstein relation

$$\mu_H = \frac{q}{kT} D \quad (7-54)$$

where D is the diffusion coefficient. Following the approach for ionic conduction, we can obtain

$$D = W_H a^2 \quad (7-55)$$

where a is the width of the potential barrier. So μ_H can be expressed as

$$\mu_H = \mu_{HO} \exp(-\Delta E_j/kT) \quad (7-56)$$

and hence the hopping conductivity as

$$\sigma_H = \sigma_{HO} \exp(-\Delta E_j/kT) \quad (7-57)$$

where μ_{HO} and σ_{HO} are constants.

The hopping process is illustrated schematically in Figure 7-12. Whether charge transport takes place according to the band model or the hopping model depends on the electron–lattice interaction. For molecular crystals or polymers, this depends on whether the strongest coupling is with the intermolecular (lattice) or the intramolecular (nuclear) vibrations. The vibration periods are typically 10^{-12} sec for intermolecular modes and 10^{-14} sec for intramolecular modes. Denoting electron relaxation time, intermolecular vibration period, and intramolecular vibration period by τ , $\tau_{\nu\ell}$, and $\tau_{\nu n}$ respectively, we have the following two important cases³⁵:

Case 1: $\tau < \tau_{vn} < \tau_{vl}$. In this case, electron motion is so rapid that the vibration motion can be regarded as stationary and a perturbation to the motion of the electrons. The electrons can be thought of as waves traveling over several lattice sites before being scattered. The band model is applicable for this case.

Case 2: $\tau_{vn} < \tau < \tau_{vl}$. In this case, the molecule vibrates (intramolecular vibration) while the electron remains on a particular lattice site. This implies that while the electron remains on the lattice site, the nuclei of the molecule on this particular lattice site move to new equilibrium positions. This gives rise to the formation of a polaron. Polaron theory will be discussed in the next section. The interaction of electrons and phonons in the lattice site may lead to self-trapping, in which the electrons polarize the molecules and are trapped in self-induced potential wells. This case may lead either to random hopping transport or to coherent band transport. For the former, the electron trapped in such a potential well requires an activation energy to surmount a barrier of a height equal to the binding energy of the polaron in order to move to the neighboring site.

Polaron Conduction

The band model for electronic conduction is not always appropriate for some dielectric solids, particularly for those with a low electron mobility. The interaction between a slow electron and the vibration modes of a polar lattice may be so strong that the polarization of the lattice caused by the slow electron will act back on the electron itself, reducing its energy. As the electron moves through the polar lattice, it carries with it the polarization field. Thus, the electron and the accompanying polarization field can be considered a quasi-particle. This quasi-particle is generally referred to as a *polaron*.^{36,37}

The most important effect of lattice polarization is the attendant increase in the effective mass of the electron. The size of a polaron is measured by the extent of the region over which the distortion or deformation of the

lattice due to polarization is introduced. In ionic crystals, electron-phonon coupling arises mainly from long-range, strong coulomb interaction between the electron and optical lattice modes. Therefore, the radius of the distorted region is much larger than a lattice constant. The polarons in ionic crystals are sometimes called *large polarons*. In molecular crystals, electron-phonon coupling is strong but of short range; the distortion may occur predominantly within the order of a lattice constant around the electron. The size of the polarons in this case is small, so such polarons are called *small polarons*.

There is a great deal of work, both theoretical and experimental, on polarons. To review it is beyond the scope of this book. For more details on this subject see some excellent reviews with special emphasis on inorganic materials³⁷⁻⁴⁴ and some mainly on molecular crystals.⁴⁵⁻⁴⁸

Large polarons have a radius larger than several lattice constants. Frohlich³⁷ has derived the Hamiltonian for large polaron behavior. This Frohlich Hamiltonian contains two important characteristic constants, the first of which is of the dimensions of the length given by

$$\ell = \mu^{-1} = (2m^*\omega/\hbar)^{-1/2} \quad (7-58)$$

where ω is the angular frequency of the lattice oscillators and m^* is the rigid-band effective mass of the electron. The electronic polarization of the ions follows the motion of a slow electron adiabatically. This polarization affects the value of the rigid-band effective mass. The length ℓ can be considered a measure of the polaron's size.

The second important characteristic constant is the coupling constant, which is dimensionless and given by

$$\alpha = \frac{2(\text{Deformation Energy})}{\hbar\omega} \quad (7-59)$$

where $\hbar\omega$ is the energy of the longitudinal optical phonon and α is a measure of the strength of the electron-lattice interaction.³⁹ For $\alpha < 1$ the situation corresponds to a weak coupling. For $\alpha > 1$ the coupling is strong. Of

course, the effective mass m^* also plays an important role in determining whether the polaron is large or small. In general, large values of m^* imply small polarons, while small values of m^* are for large and weakly coupled polarons.

Any polaron is generally described by a narrow band of energy whose width decreases with increasing temperature. Polaron motion becomes possible because there are always small overlaps of the wave functions between neighboring positions. Thus, there are two alternatives for polaron motion: one is motion in the band (nonlocalized polarons), and the other is motion by hopping (localized polarons), which is thermally activated.⁴⁹ In a large class of dielectric materials, small polarons prevail, particularly in molecular crystals.

For small polarons, the lifetime of a polaron at any site is long because of strong electron–lattice coupling. The polaron can move either as by tunneling between equivalent localized polaron states centered at different sites or by hopping between two nonequivalent localized states, involving emission and absorption of phonons. Tunneling is analogous to a wavelike motion, which is of the band conduction and in which the vibrational states involve only a few quanta and are well separated. Hopping is a phonon-activated process which is predominant at high temperatures and involves a large number of highly excited vibrational levels, so the polaron motion, which is greatly affected by interactions with vibrations, becomes random and nonwavelike, and cannot be described in terms of a band structure. In general, if the phonon bandwidth is small compared to the polaron bandwidth, band-type conduction may be predominant; if the reverse is true, conduction may be mainly by hopping. Holstein⁴⁵ has found that at $T \leq 0.4\hbar\omega/k$, a band-type conduction may be assumed, and at $T > 0.5\hbar\omega/k$, a hopping-type conduction may be assumed, ω being the optical mode vibrational angular frequency. With increasing temperature, polaron bandwidth decreases, while polaron effective mass increases rapidly.

7.3.2 Lifetime and Relaxation Electrical Conduction

Electrical conduction can be classified into two distinct types, depending on whether the minority lifetime τ_o is greater or smaller than the dielectric relaxation time τ_d . Materials with $\tau_o > \tau_d$ are generally referred to as *lifetime materials*, while materials with $\tau_o < \tau_d$ are generally referred to as *relaxation materials*. In general, most inorganic semiconductors, such as germanium and silicon, are lifetime materials. Semiconductors and insulators with a high resistivity, such as intrinsic GaAs, organic semiconductors, and amorphous materials, are relaxation materials.

Lifetime Regime

Before discussing the relaxation regime, it is important to review briefly the physical concept of the lifetime regime of conventional semiconductors. The condition $\tau_o > \tau_d$ is often not clearly or precisely defined but, in most cases, it is assumed that the neutrality condition prevails in the conventional semiconductors. This implies that dielectric relaxation is so rapid, compared to other time-dependent events, that it can be assumed to be instantaneous. Relaxation time is given by

$$\tau_d = \frac{\epsilon}{\sigma} = \rho\epsilon \quad (7-60)$$

where σ is the electric conductivity, which is a measure of the concentration and the mobility of available mobile carriers, and ϵ is the permittivity, which is a measure of the strength of the coulombic interaction inside the semiconductor (the dielectric medium). Thus, the product $\rho\epsilon$ can be thought of as the RC (R is the resistance and C is the capacitance) time constant of the materials. Obviously, the larger the value of τ_d is, the longer the time required to attain equilibrium. In conventional semiconductors, τ_d is of the order of 10^{-12} sec and τ_o is normally larger than 10^{-9} sec. Thus, it can be imagined that injection of minority carriers (electrons) of concentration Δn into the semi-

conductor will result in a change of the quasi-Fermi level, following the relation

$$n = n_o + \Delta n = N_c \exp[-(E_c - E_{Fn})/kT] \quad (7-61)$$

where n_o is the concentration of electrons in equilibrium. After injection, the majority carriers (holes) will quickly respond to neutralize excess Δn so as to maintain the neutrality condition, as shown in Figure 7-13(a).

The processes for attaining equilibrium are relaxation and recombination processes. The field created by excess minority carriers charge $-q\Delta n$ will attract majority hole carriers Δp from the vicinity until the excess charge is completely neutralized. The characteristic time for this process is the relaxation time τ_d . After this relaxation process, the semiconductor is in the neutrality condition. To reach the equilibrium condition, the recombination process must reduce Δn with time until the law of mass action is restored. The characteristic time for this process is the minority carrier lifetime τ_o . In equilibrium, E_{Fn} and E_{Fp} coincide to form one Fermi level E_{Fo} , which determines the equilibrium electron and hole concentrations n_o and p_o . In nonequilibrium, Δp must increase locally to neutralize Δn , so E_{Fp} follows a similar relation as Equation 7-61.

$$p = p_o + \Delta p = N_v \exp[-(E_{Fp} - E_v)/kT] \quad (7-62)$$

Therefore, both Δn and Δp decrease with time. E_{Fn} and E_{Fp} split rapidly into two levels (within the relaxation time τ_d), then combine slowly to E_{Fo} at equilibrium within the lifetime τ_o , as shown in Figure 7-13(a).

Relaxation Regime

From Equation 7-60, to satisfy the condition $\tau_o < \tau_d$ for the occurrence of the relaxation regime, τ_d must be large. Materials with a low carrier mobility and a low concentration of mobile carriers are most likely of this category. For example, a semiconductor with resistivity of 10^8 ohm-cm has a relaxation time of about 10^{-4} sec, which is normally larger than the minority carrier lifetime, which is generally

lower than 10^{-8} sec for high-resistivity materials. Following this classification, gold-doped silicon and germanium may become relaxation semiconductors at extremely low temperatures, at which τ_o becomes smaller than τ_d . In the lifetime regime, the neutrality condition is retained long before the excess carriers Δn disappear; in the relaxation regime, the situation is reversed. After the injection of Δn minority carriers into a relaxation semiconductor, the law of mass action is quickly restored by reducing the local majority carrier concentration within the minority carrier lifetime τ_o . Thus, at the completion of this process the np product follows the relation

$$pn = (n_o + \Delta n)(p_o - \Delta p) = n_i^2 = p_o n_o \quad (7-63)$$

This leads to

$$\Delta p = p_o \Delta n / (n_o + \Delta n) \quad (7-64)$$

If the injection is so high that $\Delta n > n_o$, then

$$\Delta p \rightarrow p_o \quad (7-65)$$

This implies that under an extreme condition all mobile majority carriers may have disappeared.^{50,51} It is clear that in the lifetime regime, injection of minority carriers tends to decrease the resistivity of the material, while in the relaxation regime, injection of minority carriers tends to increase the resistivity of the material, as can be seen from Equations 7-63 through 7-65 and Figure 7-13(b). This feature of injection is characteristic for the relaxation regime, and it is sometimes called *recombinative space charge injection*. In general, a majority carrier depletion region is adjacent to the minority carrier-injecting contact, followed by a narrow recombination front. This majority carrier depletion region causes a sublinear voltage dependence of currents ($I \propto V^{1/2}$). An increase in voltage enhances the space charge in the depletion region, thus increasing differential resistance and sometimes even creating a negative differential resistance region.^{50,51} After the elapse of τ_o , the space charge slowly relaxes to retain the equilibrium condition, as shown in Figure 7-13(b). Table 7-1 lists the basic differences between lifetime and relaxation semiconductors in their electrical behavior.

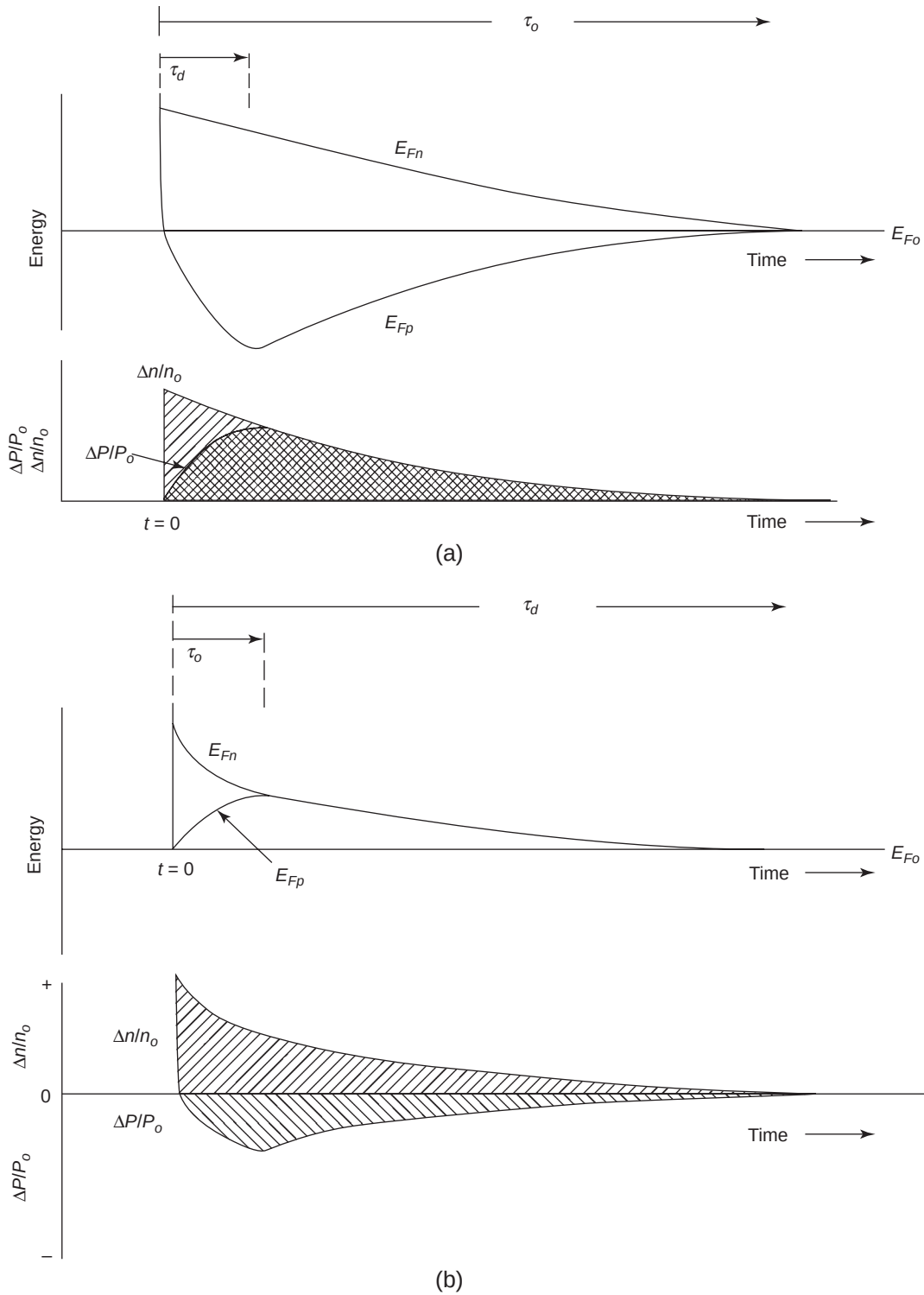


Figure 7-13 Response of quasi-Fermi levels E_{Fn} and E_{Fp} and of the relative carrier concentrations to an applied pulse of excess minority carriers (electrons) Δn (a) in the lifetime regime $\tau_0 > \tau_d$ and (b) in the relaxation regime $\tau_0 < \tau_d$.

Table 7-1 Comparison between lifetime materials and relaxation materials.

<i>Item</i> \ <i>Type</i>	<i>Lifetime Semiconductors</i>	<i>Relaxation Semiconductors</i>
t_o, t_d	$t_o > t_d$	$t_o < t_d$
I - V characteristics (forward biased)	$I \propto V^n$ or $I \propto \exp(qV/nkT)$ with $n \geq 1$ (super-linear)	$I \propto V^n$ with $n < 1$ or $n = 1/2$ (sublinear)
E_{Fn}, E_{Fp} (after carrier injection)	$E_{Fn} \neq E_{Fp}$ Local space charge neutrality, local nonequilibrium	$E_{Fn} = E_{Fp}$ Local space charge enhancement-space charge effect, local equilibrium
The region near the contact of minority carrier injection	Majority carrier enhancement (resistivity decreasing)	Majority carrier depletion (resistivity increasing)
General properties	Low resistivity, high mobility, and narrow energy band gap (e.g., conventional semiconductors, Si and Ge)	High resistivity, low mobility, and wide energy band gap (e.g., organic and amorphous semiconductors or conventional semiconductors at low temperatures)
Electrical transport (electronic, photoelectric, and galvanomagnetic properties)	Extensively studied (well understood)	Not yet explored (not fully understood)

A field domain (or a potential disturbance) created by an injected pulse of minority carriers will move under the influence of an applied electric field. In the relaxation regime, the disturbance will move in the same direction as the majority carriers, because the excess minority carriers are quickly reduced by the recombination process. The disturbance in majority carrier concentration moves slowly to relax toward neutralization. In the lifetime regime, the disturbance moves in the direction of minority carriers. Unless otherwise stated, this book deals with solids in the lifetime regime.

7.4 Bulk-Limited Electrical Conduction

To achieve bulk-limited electrical conduction, the electrical contacts (electrodes) must be ohmic (see Charge Carrier Injection through Potential Barriers from Contacts in Chapter 6). For electrical conduction involving mainly one type of charge carrier, say electrons, the cathode region can be considered a carrier reservoir that supplies charge carriers to the

anode region as demanded by applied voltage conditions. Thus, the bulk electrical conduction becomes space-charge limited (SCL).

7.4.1 Basic Concepts Relevant to Space-Charge Limited Electrical Conduction

Space charge is generally referred to as the space filled with a net positive or negative charge, and it appears in a great variety of situations associated with semiconductors and insulators. This section is concerned mainly with the limit such a space charge imposes on the current or the number of charge carriers per second passing from one electrode to the other. For example, if the cathode emits more electrons per second than the space can accept, the remainder will form a negative space charge, which creates a field to reduce the rate of electron emission from the cathode. Thus, the current is controlled not by the electron-injecting electrode but by the bulk of the semiconductor or the insulator—in other words, by the carrier mobility in the space inside the material. In general, the emitted electrons have a distribution of energies, the material has traps

of various distributions, and there exist various high-field effects. Actually, the situation is quite complicated; in order to find a solution that is not too burdensome for the current–voltage characteristics, we must resort to simplifying assumptions.

Space-charge limited dark conduction occurs when the contacting electrodes are capable of injecting either electrons into the conduction band or holes into the valance band of a semiconductor or an insulator, and when the initial rate of such charge-carrier injection is higher than the rate of recombination, so the injected carriers will form a space charge to limit the current flow. Therefore, the SCL current is bulk limited.

Once the carrier-injecting contact can provide a reservoir of carriers, the behavior of the injected carriers and hence the current is controlled by the properties of the material in which the carriers are flowing. In molecular crystals, the bandwidth is narrow and the forbidden energy gap is wide; hence, carrier mobility is low, so the intrinsic resistivity of these materials is high. Thus, the SCL current is easily observed even though the carrier-injecting contact may not be perfect because the intrinsic resistance of the material is usually much larger than the contact resistance. As there are no perfect crystals existing in this world, traps created by all types of imperfections are always present in the crystals and interact with injected carriers from ohmic contacts, thus controlling the carrier flow and determining the current–voltage (J – V) characteristics. In molecular crystals, two types of carrier trap distributions have been reported^{52,53}:

- Traps confined in discrete energy levels in the forbidden energy gap
- Traps with a quasi-continuous distribution of energy levels (normally following an exponential form or a Gaussian form) with a maximum trap density near the band edges

Several methods can be used to determine experimentally the energetic and kinetic parameters (energy levels and distributions) of carrier

traps, such as the space-charge limited current (SCLC) method,⁵⁴ the thermally stimulated current (TSC) method,⁵⁵ and the photo-emission method.⁵⁶ However, these methods do not provide any information about the possible physical nature of traps. Some general considerations have been suggested to relate the discrete trap levels to chemical impurities introduced into the lattice (chemical traps),^{54,57} and to relate quasi-continuous trap distribution to the imperfection of the crystal structure (structural traps).^{53,58,59}

It should be noted, however, that the surroundings of a given type of trapping are not uniquely defined. It can be imagined that there always exist differences in configuration between nearest neighbors and in character between trapping centers, so a discrete trap level can be considered “smeared out.” Furthermore, the surroundings of an impurity entity are generally inhomogeneous. Several investigators^{60,61–63} have proposed that some types of traps are better described by a Gaussian distribution function, such as the quasi-continuous trap distribution associated with statistical dispersion of the charge-carrier polarization energy caused by fluctuational structural irregularities of the lattice.^{64,65} The current–voltage (J – V) characteristics have been analyzed for solids with traps distributed in a Gaussian manner in energy but uniformly distributed in space.^{66–68}

The spatial distribution of traps can never be homogeneous because there always exist discontinuities between the material and the electrodes. The thinner the material specimen used for experimental studies is, the greater the influence of the form of spatial distribution of traps on J – V characteristics. The effect of nonuniform spatial trap distribution is important for thin films. In fact, this effect has been observed in thin films,^{69–71} possibly due to surface topography, grain boundaries, nonuniform doping, microcrystalline defects, etc.

The probability that a trap will capture an electron follows the Fermi–Dirac statistics

$$f_n(E) = \frac{1}{1 + g_n^{-1} \exp[(E - E_{Fn})/kT]} \quad (7-66)$$

The probability that a trap will capture a hole follows

$$f_p(E) = \frac{1}{1 + g_p \exp[(E_{Fp} - E)/kT]} \quad (7-67)$$

On the basis of their energy levels, the traps can be classified as shallow or deep. The so-called *shallow traps* refer to traps whose energy levels $E = E_m$ are located above the quasi-Fermi level E_{Fn} for electron traps, and to traps whose energy levels $E = E_p$ are located below the quasi-Fermi level E_{Fp} for hole traps. It can be seen from Equations 7-66 and 7-67 that $f_n(E) \ll 1$ or $f_p(E) \ll 1$ if $(E_m - E_{Fn})$ or $(E_{Fp} - E_p)$ is much greater than kT . This means that most of the traps may be empty. Conversely, if E_m is below E_{Fn} or E_p above E_{Fp} , the traps are called *deep traps*, in which $f_n(E) \rightarrow 1$ or $f_p(E) \rightarrow 1$ if $(E_{Fn} - E_m)$ or $(E_p - E_{Fp})$ is much greater than kT . This implies that most of the traps are filled with trapped carriers (trapped electrons or trapped holes). Figure 7-14 shows schematically these two cases.

Carrier injection into a solid is generally classified as single or double injection. *Single injection* means that the current flow is due mainly to one type of carrier (electrons or holes) injected from a contacting electrode into the solid. These injected carriers gradually establish a space charge leading to the well known single-carrier SCL current. *Double injection* means that the current flow involves two types of carriers: electrons injected from the cathode and holes from the anode. In double injection, recombination kinetics control all the electrical properties. The recombination process may either be bimolecular (i.e., band-to-band electron-hole recombination) or may occur through one or more sets of localized recombination centers. The J - V characteristics are strongly dependent on the concentration and the distribution function of traps inside the specimen and other boundary conditions.

7.4.2 SCL Electrical Conduction: One-Carrier (Single) Planar Injection

This section deals mainly with the theoretical analyses of the SCL electrical conduction pro-

cesses under various trappings conditions, the scaling rule and the effect of carrier diffusion.

Theoretical Analysis

In single crystals the trap energy levels, if any, are generally discrete. In amorphous and polycrystalline materials they are distributed in accordance with certain distribution functions.⁶² The latter has been attributed to the intrinsic disorder of the lattice, which is possibly due to the variation of the nearest neighbor distances. Material specimens in film form, produced by vacuum deposition or other means, are likely to be polycrystalline. Therefore, traps created by defects are generally distributed and their density is rather high, even if the material itself is very pure chemically. Furthermore, material specimens always have boundaries, such as their surfaces with metallic contacts. Trap distribution near such boundaries would be different from that in the bulk. In our theoretical analysis, we will confine our discussion to steady-state DC one-dimensional planar current flow and make the following assumptions, but the treatment is general and therefore can be applied to thick or thin specimens in the crystal or film form of any material.

- The energy band model can be used to treat the behavior of injected carriers.
- Only injected hole carriers are considered, and the ohmic contact to inject them is perfect. (Similar treatment can easily be extended to a case of only injected electron carriers.) This implies that there is no electrode limitation to the current.
- The mobility of the free holes (or free electrons) is independent of electric field and not affected by the presence of traps.
- The free hole (or electron) density follows the Maxwell-Boltzmann statistics; the trapped hole (or trapped electron) density follows the Fermi-Dirac statistics.
- The electric field is so large that the current components due to diffusion and to carriers generated thermally in the specimen can be neglected. The former is justified if the applied voltage is larger than several kT/q so

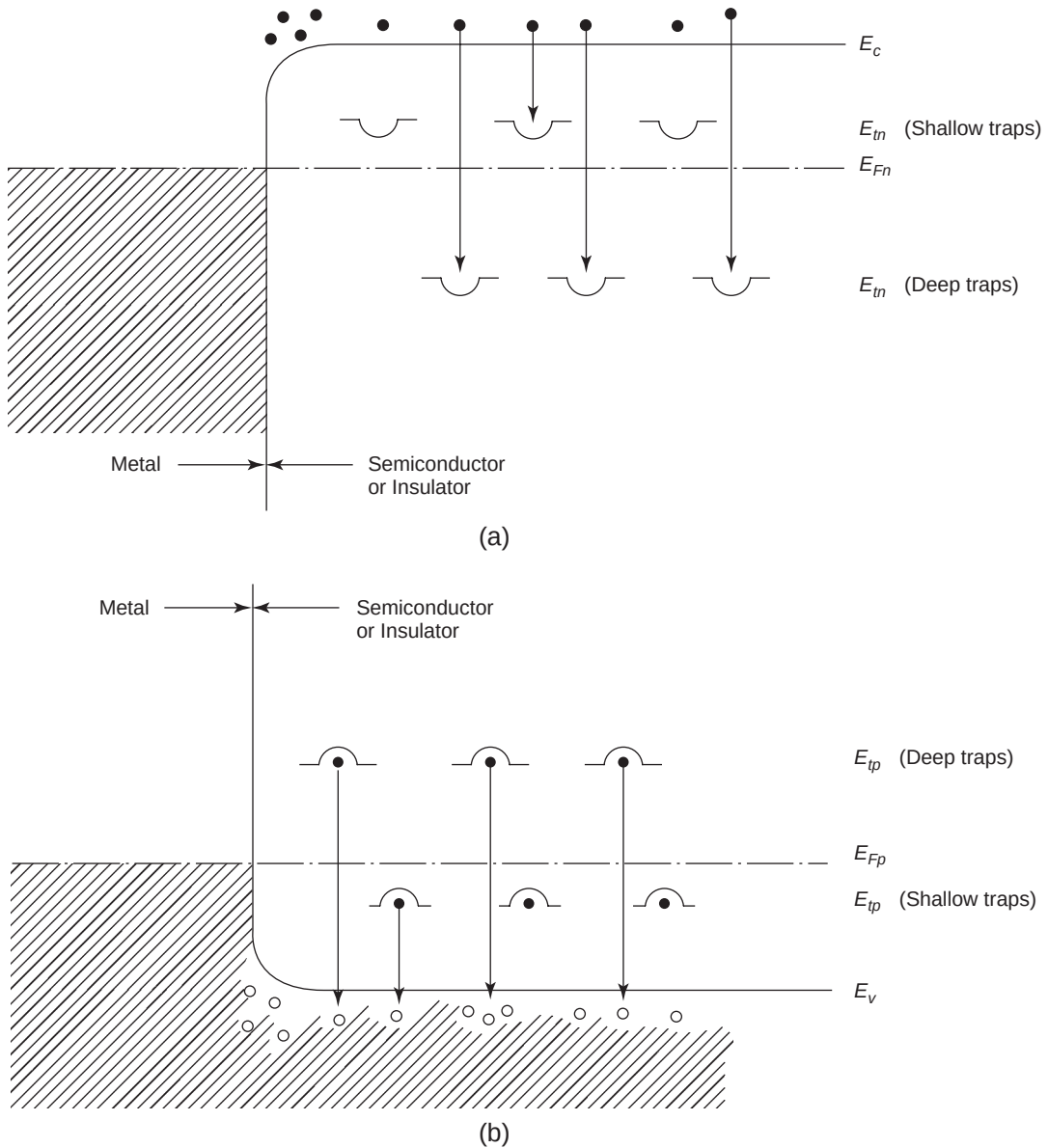


Figure 7-14 Schematic energy level diagrams for (a) electrons injecting from an electron ohmic contact to a semiconductor or an insulator with shallow and deep electron traps and (b) holes injecting from a hole ohmic contact to a semiconductor or an insulator with shallow and deep hole traps.

that the drift term becomes predominant and the diffusion term may be neglected without causing serious error. The latter is justified if the contact resistance is much less than the intrinsic resistance of the specimen, as in molecular crystals.

- High-field effects, such as the Poole-Frenkel effect, impact ionization, and field-dependent mobility, are ignored.
- The treatment is one-dimensional with the plane at $x = 0$ as the hole injecting contact

(anode) and that at $x = d$ as the collecting contact, the specimen thickness being d . This implies that $F(x = 0) = 0$, and that the distance W_a between the actual electrode surface and the virtual anode ($-dV/dx = F = 0$) is so small that we can assume $F(x = W_a \rightarrow 0) = 0$ for simplicity.

The distribution function for trap density as a function of energy level E above the edge of the valence band and distance x from the injecting contact for hole carriers can be written as

$$h(E, x) = N_t(E)S(x) \quad (7-68)$$

where $N_t(E)$ and $S(x)$ represent, respectively, the energy and spatial distribution functions of traps. If the traps capture only holes, the electric field $F(x)$ inside the specimen follows Poisson's equation

$$\frac{dF(x)}{dx} = \frac{q[p(x) + p_t(x)]}{\epsilon} \quad (7-69)$$

and the current density may be written as

$$J = q\mu_p p(x)F(x) \quad (7-70)$$

where $p(x)$ and $p_t(x)$ are, respectively, the densities of injected free and trapped holes, which are given by

$$p_t(x) = \int_{E_t}^{E_u} h(E, x) f_p(E) dE \quad (7-71)$$

and

$$p(x) = N_v \exp(-E_{fp}/kT) \quad (7-72)$$

and $f_p(E)$ is the Fermi-Dirac distribution function, which is given by Equation 7-67. In the following sections, we shall consider six general cases.

Without Traps (Trap-Free Solids—Ideal Case)

For this case, $p_t(x) = 0$. Multiplying both sides of Equation 7-69 by $2F(x)$ and substituting Equation 7-70 into it, we obtain

$$2F(x) \frac{dF(x)}{dx} = \frac{d[F(x)]^2}{dx} = \frac{2J}{\epsilon\mu_p} \quad (7-73)$$

Integration of Equation 7-73 and use of the boundary condition

$$V = \int_0^d F(x) dx$$

yield

$$J = \frac{9}{8} \epsilon\mu_p \frac{V^2}{d^3} \quad (7-74)$$

This is the well known Mott-Gurney equation⁷² and is sometimes referred to as the square law for trap-free SCL currents.

We have ignored the effect of thermally generated carriers. At low applied voltages, the J - V characteristics may follow Ohm's law if the density of thermally generated free carriers p_o inside the specimen (we are considering holes only) is predominant such that

$$qp_o\mu_p \frac{V}{d} \gg \frac{9}{8} \epsilon\mu_p \frac{V^2}{d^3}$$

The onset of the departure from Ohm's law or the onset of SCL conduction takes place when the inequality becomes equal. The applied voltage for this condition to occur is

$$V_\Omega = \frac{8}{9} \frac{qp_o d^2}{\epsilon} \quad (7-75)$$

as shown schematically in Figure 7-15. By rearranging Equation 7-75 in the form

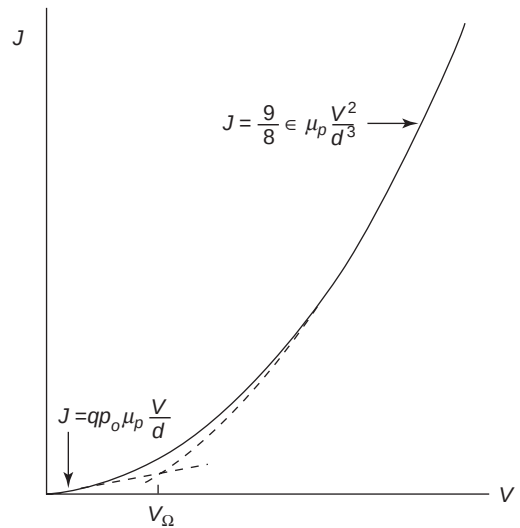


Figure 7-15 Schematic diagram showing the transition from ohmic to space charge limited conduction for one carrier (hole) injection in a trap-free solid.

$$\frac{d^2}{\mu_p V_\Omega} = \frac{9}{8} \frac{\epsilon}{qp_o \mu_p}$$

we have

$$t_i \approx \tau_d \quad (7-76)$$

This means that when the transition from the ohmic to the SCL regime takes place, the carrier transit time $t_i = d^2/\mu_p V_\Omega$ at V_Ω (the minimum voltage required for the transition) is approximately equal to the dielectric (or ohmic) relaxation time $\tau_d = \epsilon/qp_o \mu_p$. If the applied voltage V is less than V_Ω , then $t_i > \tau_d$, implying that the injected carrier density p is small in comparison with p_o , that the injected carriers will redistribute themselves with a tendency to maintain electric-charge neutrality internally in a time comparable to τ_d , and that they have no chance to travel across the specimen. The redistribution of the charge is known as *dielectric relaxation*. This means that the injected carriers under this condition do not alter the density of carriers p_o because the injection of p would be accompanied by an unbalanced space charge which, according to Gauss's law, would give rise to an electric field, thus exerting a force on adjacent electrons so that they move in to neutralize the space charge. The net result is that holes in all parts of the specimen start to drift in such a way that the injected holes flow into the specimen from the injecting contact to replace the holes flowing out at the collecting electrode, so no appreciable change in hole density occurs anywhere within the specimen. This can be easily understood by solving the following continuity equation

$$q \frac{\partial(p + p_o)}{\partial t} = -\nabla \cdot J \quad (7-77)$$

where current J is given by

$$J = q(p + p_o)\mu_p F - qD_p \nabla(p + p_p) \quad (7-78)$$

in which F produced by injected p follows Poisson's equation. Substituting Equation 7-69 into Equation 7-78 and then into Equation 7-77 and neglecting second-order terms involving pF , we obtain

$$\frac{\partial p}{\partial t} = -\left(\frac{qp\mu_p}{\epsilon}\right)p + D_p \nabla^2 p \quad (7-79)$$

If $p < p_o$ and p spreads uniformly over the specimen in a time comparable to the dielectric relaxation time τ_d , the second term on the right side of Equation 7-79 can be ignored. Thus, the solution of Equation 7-79 yields

$$p(t) = p(t=0) \exp\left(-\frac{t}{\tau_d}\right) \quad (7-80)$$

τ_d is a measure of the time required for the injected carrier to reestablish equilibrium. Since $t_i > \tau_d$, negligible space charge would appear in the bulk. In most inorganic semiconductors, τ_d is small (e.g., τ_d in Ge is about 10^{-12} sec) and the dielectric relaxation is not easy to observe. However, in molecular crystals, such as anthracene whose τ_d is about 10^{-5} sec, the relaxation becomes important.

Equation 7-80 is no longer valid if $t_i \approx \tau_d$ because $D_p \nabla^2 p$ in Equation 7-79 can no longer be neglected. When $V > V_\Omega$ and $t_i \approx \tau_d$ (for $V < V_\Omega$, t_i increases with decreasing V but τ_d remains practically constant, while for $V > V_\Omega$, t_i decreases with increasing V , and τ_d also decreases with increasing V because the increase in V causes an increase in free carrier density in the bulk) or $t_i < \tau_d$, the injected excess carriers dominate the thermally generated carriers because the injected carrier transit time is too short for their charge to be relaxed by the thermally generated carriers.^{54,73} By rearranging Equation 7-75, we get

$$qp_o d = \frac{9}{8} V_\Omega \frac{\epsilon}{d}$$

The physical meaning of this equation is that the total charge of free carriers is approximately equal to the condenser charge (the product of capacitance and voltage). This is equivalent to saying that the SCL current is due to the condenser charge driven from one plate through the bulk of the specimen and then into the opposite plate by the applied voltage.

Through qualitative reasoning, we can imagine that there are two parallel current paths in the specimen, one due to the ohmic process and the other to the SCL process. Figure 7-15 shows the qualitative physical picture of these two processes and the net current. At $V = V_\Omega$, the net current can be qualitatively expressed as

$$\begin{aligned}
J &= qp_o\mu_p \frac{V_\Omega}{d} + \frac{9}{8}\epsilon\mu_p \frac{V_\Omega^2}{d^3} \\
&= qp_o\mu_p \frac{V_\Omega}{d} + qp_o\mu_p \frac{V_\Omega}{d} \quad (7-81) \\
&= q(2p_o)\mu_p \frac{V_\Omega}{d} = qp\mu_p \frac{V_\Omega}{d}
\end{aligned}$$

This implies that the onset of $t_i = \tau_d$ is also the condition for doubling the total free carrier density in the bulk of the specimen, as shown in Equation 7-81. It should be noted that Equation 7-81 serves only to explain the physical picture. It is not logical to consider the two processes existing simultaneously in the same space because the potential distributions are different for these two types of current. It is much better to say that when $t_i > \tau_d$, the ohmic process is predominant and the effect of injected space charge is suppressed, while when $t_i < \tau_d$, SCL conduction is predominant and the ohmic process is suppressed; the potential distribution will adjust itself to suit the dominant process. However, the transition from ohmic to SCL conduction is not an abrupt change but a gradual one.

Traps Confined in Single or Multiple Discrete Energy Levels

For this case, Equation 7-68 can be expressed as

$$h(E, x) = H_a \delta(E - E_t) S(x) \quad (7-82)$$

where H_a is the density of traps, E_t is the trap energy level above the edge of the valence band, and $\delta(E - E_t)$ is the Dirac delta function. From Equations 7-71 and 7-73 we obtain

$$\begin{aligned}
p_i(x) &= \int_{E_t}^{E_u} \frac{H_a \delta(E - E_t) S(x) dE}{1 + g_p \exp[(E_{FP} - E)/kT]} \\
&\simeq \frac{H_a S(x)}{1 + [H_a \theta_a / p(x)]} \quad (7-83)
\end{aligned}$$

in which

$$\theta_a = \frac{g_p N_v}{H_a} \exp(-E_t/kT) \quad (7-84)$$

Substitution of Equation 7-84 into Equation 7-69 gives

$$\frac{dF(x)}{dx} = \frac{q}{\epsilon} \left[p(x) + \frac{H_a S(x)}{1 + [H_a \theta_a / p(x)]} \right] \quad (7-85)$$

An analytical solution of Equation 7-85 for J as a function of applied voltage is not possible, although a numerical solution can be obtained for all possible cases separately. For simplicity, we will assume that E_t is a shallow trap level located below E_{FP} . This implies that $H_a \theta_a > p(x)$. On the basis of this assumption and by multiplying both sides of Equation 7-85 by $2F(x)$ and substituting Equation 7-70 into it, we obtain

$$2F(x) \frac{dF(x)}{dx} = \frac{d[F(x)]^2}{dx} = \frac{2J}{\epsilon\mu_p \theta_a} [\theta_a + S(x)] \quad (7-86)$$

Integration of Equation 7-86 and use of the boundary condition

$$\begin{aligned}
V &= \int_0^d F(x) dx \\
J &= \frac{9}{8} \epsilon\mu_p \theta_a \frac{V^2}{d_{eff}^3} \text{ give} \quad (7-87)
\end{aligned}$$

in which V is the applied voltage and

$$d_{eff} = \left\{ \frac{3}{2} \int_0^d \left(\int_0^x [\theta_a + S(x)] dx \right)^{1/2} dt \right\}^{2/3} \quad (7-88)$$

Equation 7-87 is similar in form to that derived by Lampert,⁷⁴ except that d has been replaced with d_{eff} , which can be considered as effective thickness. The difference between d_{eff} and d can be attributed to the inhomogeneous spatial distribution of free and trapped carriers.

We shall discuss some important parameters related to the effects of traps. In what follows, we will ignore the effect of nonuniform spatial distribution of traps for simplicity. This means that we will use d instead of d_{eff} .

θ_a is, in fact, the ratio of free carrier density to total carrier (free and trapped) density

$$\theta_a = \frac{p}{p + p_t} \quad (7-89)$$

Thus, for the trap-free case, $p_t = 0$, $\theta_a = 1$. With traps, θ_a is always less than unity and could be as small as 10^{-7} .

When the density of thermally generated free carriers p_o inside the specimen (we are considering holes only) is larger than the density of injected carriers p , ohmic conduction is predominant. The onset of the transition from ohmic to SCL conduction, following the same

principle used in the previous section, occurs when the applied voltage reaches

$$V_{\Omega} = \frac{8}{9} \frac{qp_o d^2}{\theta_a \epsilon} \quad (7-90)$$

This equation indicates the following:

- The voltage for the transition V_{Ω} increases with increasing density of thermally generated carriers in the specimen p_o .
- The higher the concentration of traps (this means the smaller the value of θ_a), the higher the value of V_{Ω} for the transition.
- When the free carrier density is changed by injection from p_o to a new value p , then in the steady state (when the trapping and detrapping reach a quasi-thermal equilibrium) the density of trapped carriers is p_t . Thus, the total density of injected carriers becomes $p_T = p + p_t$. Since μ_p is the mobility of free carriers, we define the effective mobility as

$$\mu_{p\text{eff}} = \left(\frac{p}{p + p_t} \right) \mu_p = \theta_a \mu_p \quad (7-91)$$

with the understanding that the effective carrier density for electric conduction is p_T rather than p . On the basis of this definition, the effective carrier transit time can be expressed in terms of free carrier transit time t_t and V_{Ω} as

$$\begin{aligned} t_{t\text{eff}} &= \frac{t_t}{\theta_a} = \frac{d^2}{\theta_a \mu_p V_{\Omega}} \\ &= \frac{d^2}{\mu_{p\text{eff}} V_{\Omega}} \end{aligned} \quad (7-92)$$

- The transition from ohmic to SCL conduction occurs when $t_{t\text{eff}}$ is approximately equal to τ_d .

Even with $V < V_{\Omega}$, in which ohmic conduction is predominant in the steady state, there is a transient supply of injected carriers when a voltage is applied across the specimen. The ohmic behavior can be observed only after these space-charge carriers become trapped. This phenomenon has been observed in CdS crystals.^{75,76}

Increase of applied voltage may increase the density of free carriers resulting from injection to such a value that the quasi-Fermi level E_{Fp}

moves below the shallow hole trapping level E_t . Then, most traps are filled: For hole traps, a filled trap means that it has given up an electron to the valence band, while for electron traps a filled trap means that it has captured an electron or it is occupied, so when most electron traps are filled, the quasi-Fermi level E_{Fn} moves above the electron trapping level E_t . The *traps-filled limit* (TFL) is the condition for the transition from the trapped J - V characteristics to the trap-free J - V characteristics. It can be imagined that after all traps are filled, the subsequently injected carriers will be free to move in the specimen. So, at the threshold voltage V_{TFL} that brings on this transition, the current will rapidly jump from its low, trap-limited value to a high, trap-free SCL current. V_{TFL} is defined as the voltage required to fill up the traps—in other words, the voltage at which E_{Fp} passes through E_t . In thermal equilibrium (i.e., in the absence of external perturbation, and for present consideration, in the absence of applied voltage) the density of trapped holes is

$$\begin{aligned} p_{to} &= \int_{E_u}^{E_t} \frac{H_a \delta(E - E_t) dE}{1 + g_p \exp[(E_{Fpo} - E)/kT]} \\ &= \frac{H_a}{1 + g_p \exp[(E_{Fpo} - E_t)/kT]} \end{aligned} \quad (7-93)$$

and the density of unfilled traps is

$$H_a - p_{to} = \frac{H_a}{1 + g_p^{-1} \exp[(E_t - E_{Fpo})/kT]} \quad (7-94)$$

where E_{Fpo} is the quasi-Fermi level in thermal equilibrium (i.e., in the absence of applied voltage).

Shallow-traps—In this case, $E_t < E_{Fpo}$, Equation 7-94 can be approximated to

$$H_a - p_{to} \approx H_a \quad (7-95)$$

V_{TFL} can be interpreted in such a way that when the unfilled traps are completely filled, the applied voltage reaches the value of V_{TFL} . On the assumption that $H_a \gg p$, then at V_{TFL} we have

$$\frac{dF_{TFL}}{dx} = \frac{qH_a}{\epsilon} \quad (7-96)$$

Integration of Equation 7-96 gives

$$V_{TFL} = \int_0^d F_{TFL} dx = \frac{qH_a d^2}{2\epsilon} \quad (7-97)$$

For cases in which θ_a is not too small, so that $H_a/p_o > \frac{8}{9}\left(\frac{2}{\theta_a}\right)$, we have

$$V_{TFL} > V_\Omega \quad (7-98)$$

Figure 7-16 shows schematically the variation of V_Ω and V_{TFL} with θ_a and H_a .

Deep-traps—In this case, $E_t > E_{Fpo}$, Equation 7-94 can be written as

$$H_a - p_{to} = g_p H_a \exp[(E_{Fpo} - E_t)/kT] \quad (7-99)$$

Following the same method used for shallow traps and assuming that $H_a - p_{to} \gg p$, then at V_{TFL} we have

$$V_{TFL} = \frac{q(H_a - p_{to})d^2}{2\epsilon} \quad (7-100)$$

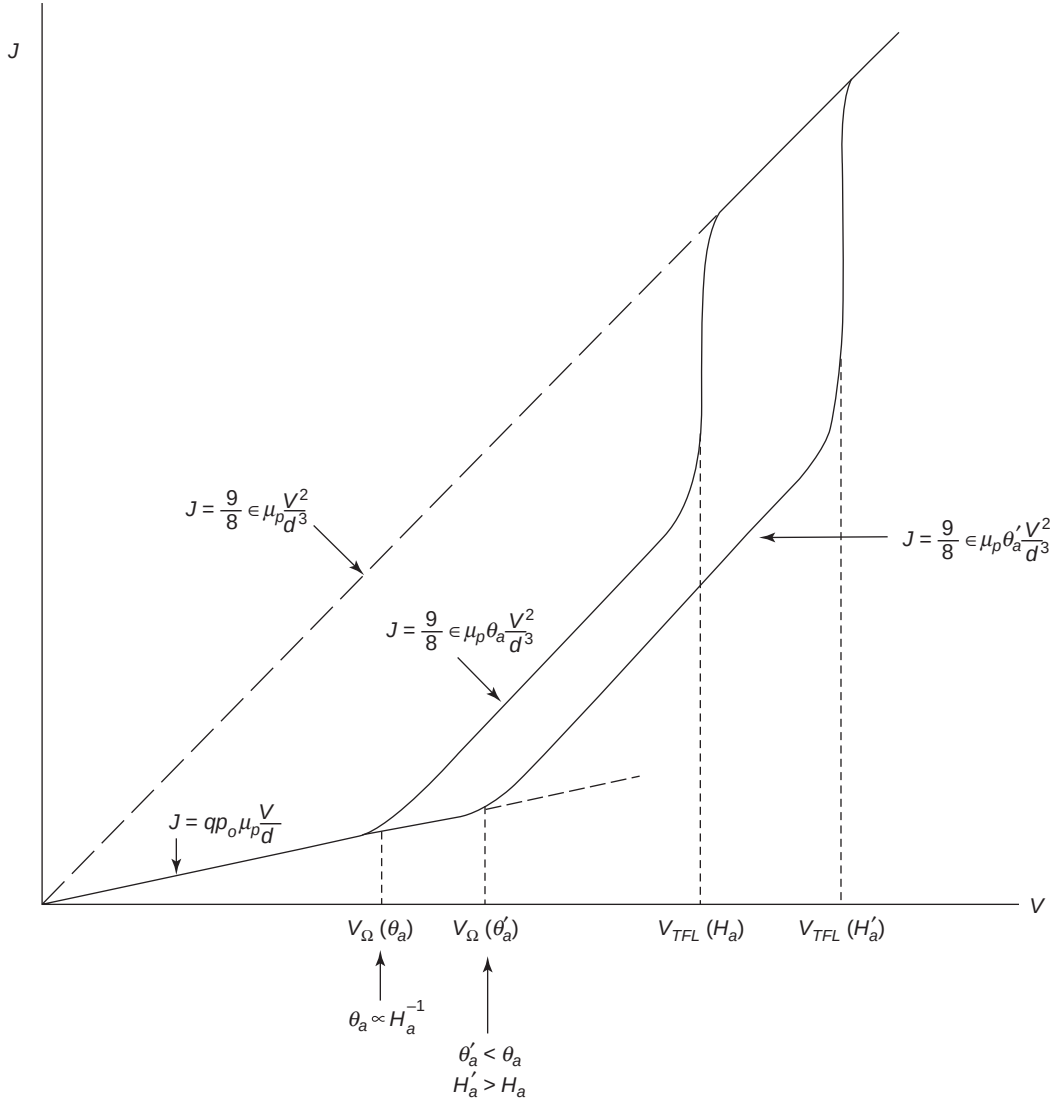


Figure 7-16 Schematic log-log plot of J - V characteristics, showing the variation of V_Ω and V_{TFL} with θ_a or H_a for a solid with shallow traps confined in a trapping level E_t ($E_t < E_{Fpo}$).

Since $E_t > E_{Fp0}$, all injected carriers will be used first to fill the traps. At V_{TFL} all traps will be filled, so V_{TFL} can also be considered the voltage to begin the transition from ohmic to SCL conduction, that is, V_Ω . For this deep-trap case, however, the transition is from ohmic to trap-free SCLC because all traps have been filled. Recalling that

$$p(|J|_{V=V_{TFL}}) \approx 2p_o \quad (7-101)$$

(see Equation 7-81), it is now interesting to see the change of p when V is increased to $2V_{TFL}$. Based on the total charge in the specimen, $Q = q(H_a - p_{to})d$ at $V = V_{TFL}$ and $Q = 2q(H_a - p_{to})d$ at $V = 2V_{TFL}$, if the capacitance of the specimen is assumed to be unchanged. At $V = V_{TFL}$, the total injected carriers $H_a - p_{to}$ are trapped carriers, but at $V = 2V_{TFL}$, one-half of the injected carriers are trapped and the other half are free carriers, so

$$p(|J|_{V=2V_{TFL}}) \approx H_a - p_{to} \quad (7-102)$$

From Equations 7-101 and 7-102, we have

$$A = \frac{J(2V_{TFL})}{J(V_{TFL})} \approx \frac{H_a - p_{to}}{p_o} \quad (7-103)$$

Figure 7-17 shows the schematic log-log plot of the current-voltage characteristics. The triangle of this plot is sometimes referred to as the *Lampert triangle*.^{74,77} A more analytic treatment of this problem, taking into account the thermally generated carriers in the specimen, has been reported by Lampert and Mark.⁵⁴

If the traps are not confined to a single discrete energy level but are in several discrete energy levels (as in a solid that contains more than two kinds of impurities), Equations 7-87 and 7-88 are still applicable, provided that θ_a is given by

$$\theta_a^{-1} = \sum_i \theta_i^{-1} \quad (7-104)$$

where

$$\theta_i = \frac{g_{pi} N_v}{H_{ai}} \exp(-E_{ti}/kT) \quad (7-105)$$

in which g_{pi} , H_{ai} , and E_{ti} refer to g_p , H_a , and E_t in the i th single discrete energy level.

Traps Distributed Exponentially within the Forbidden Energy Gap

In the analyses remaining in this section, we shall not repeat the detailed mathematical treatment. The formulation and the solution of the problem will be given, because the analytical processes are similar to those given in the analyses for cases without traps or with traps confined in single or multiple discrete energy levels. When traps distributed exponentially in the forbidden energy gap, the distribution function for trap density as a function of energy level E above the edge of the valence band and distance x from the injecting contact for holes can be written, following Equation 7-68, as

$$h(E, x) = \frac{H_b}{kT_c} \exp\left(-\frac{E}{kT_c}\right) S(x) \quad (7-106)$$

where H_b is the density of traps and T_c is a characteristic constant of the distribution. If $T_c > T$, we can assume that $f_p(E) = 1$ for $E_{Fp} < E < \infty$ and $f_p(E) = 0$ for $E < E_{Fp}$ as if we take $T = 0$. This is a good approximation, particularly when T_c is much larger than T . With this assumption, we obtain

$$\begin{aligned} p_t(x) &= \int_{E_{Fp}}^{\infty} \frac{H_b}{kT_c} \exp\left(-\frac{E}{kT_c}\right) S(x) dE \\ &= H_b \exp\left(-\frac{E_{Fp}}{kT_c}\right) S(x) \\ &= H_b \left(\frac{p}{N_v}\right)^{T/T_c} S(x) \end{aligned} \quad (7-107)$$

The upper limit of the integral has been extended to infinity. This is permissible if $E_{Fp}(x)$ is far removed from the Fermi level of the neutral region. By substituting Equation 7-107 into Equation 7-69, letting $T_c/T = \ell$ and multiplying both sides by $\left(\frac{\ell+1}{\ell}\right) [F(x)]^{1/\ell}$, we obtain

$$\begin{aligned} \left(\frac{\ell+1}{\ell}\right) [F(x)]^{1/\ell} \frac{dF(x)}{dx} &= \frac{d[F(x)]^{(\ell+1)/\ell}}{dx} \\ &= \left(\frac{\ell+1}{\ell}\right) \frac{q}{\varepsilon} [pF(x)]^{1/\ell} \\ &\quad \times [p^{(1-\ell)/\ell} + H_b N_v^{-1/\ell} S(x)] \\ &= \left(\frac{\ell+1}{\ell}\right) \frac{qH_b}{\varepsilon} \left(\frac{J}{q\mu_p N_v}\right)^{1/\ell} [\theta_b + S(x)] \end{aligned} \quad (7-108)$$

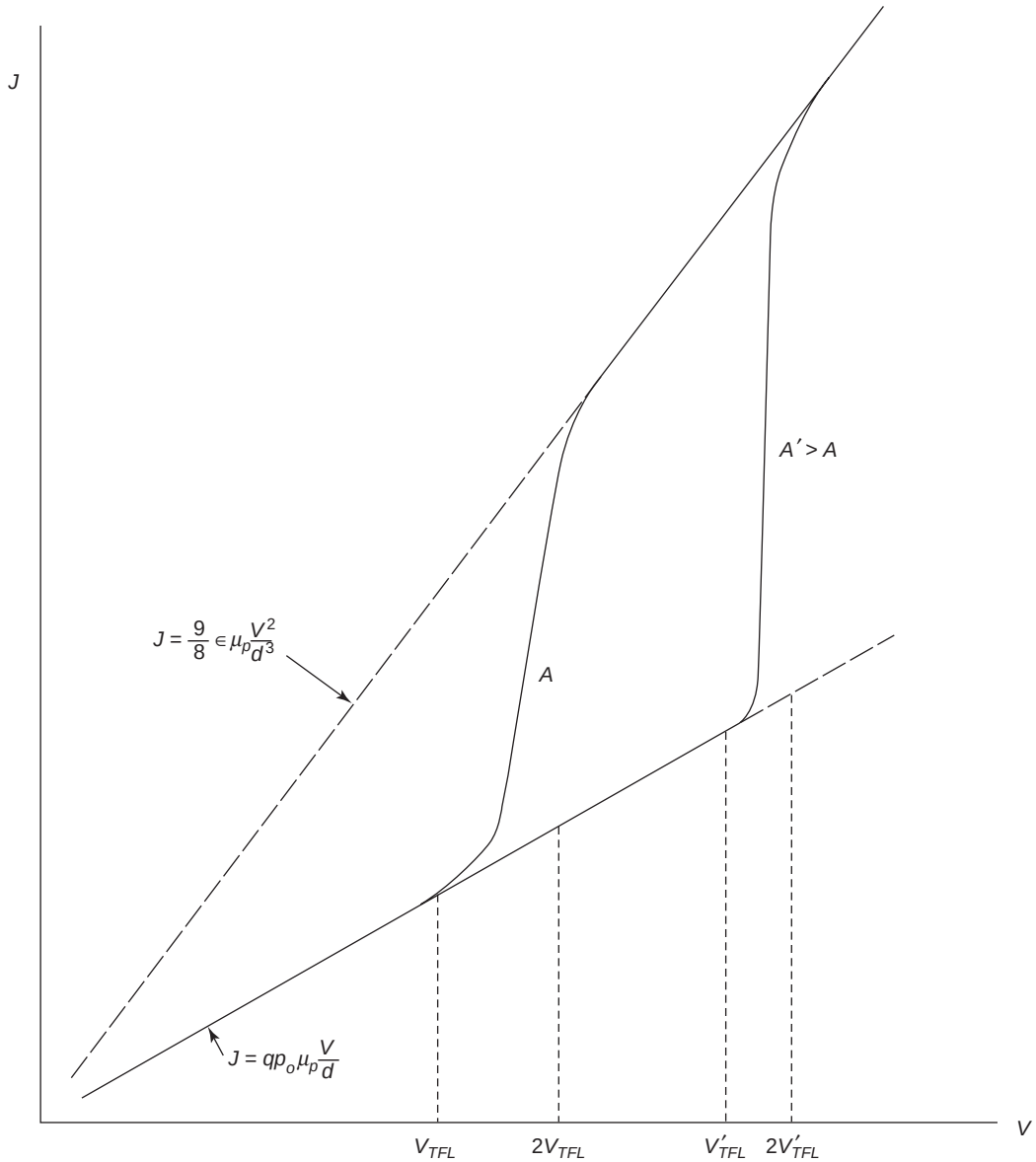


Figure 7-17 Schematic log-log plot of J - V characteristics, showing the variation of V_{TFL} with the ratio of $A = (H_a - p_o)/p_o$ for a solid with deep traps confined in a trapping level E_t ($E_t > E_{Fpo}$).

in which

$$\theta_b = \frac{N_v}{H_b} \exp\left[-\frac{E_{Fp}}{kT} \left(\frac{\ell-1}{\ell}\right)\right] \quad (7-109)$$

If $p_t \gg p$, θ_b is very small and can be neglected. Integration of Equation 7-108 and use of the boundary condition $V = \int_0^d F(x) dx$ gives

$$J = q^{1-\ell} \mu_p N_v \left(\frac{2\ell+1}{\ell+1}\right)^{\ell+1} \left(\frac{\ell}{\ell+1} \frac{\varepsilon}{H_b}\right)^\ell \times \frac{V^{\ell+1}}{d_{\text{eff}}^{2\ell+1}} \quad (7-110)$$

in which

$$d_{\text{eff}} = \left\{ \frac{2\ell+1}{\ell+1} \times \int_0^d \left(\int_0^x S(x) dx \right)^{1/(\ell+1)} dt \right\}^{(\ell+1)/(2\ell+1)} \quad (7-111)$$

for $p_i > p$.

Equation 7-110 is similar in form to that derived by Mark and Helfrich,⁷⁸ except that d has been replaced with d_{eff} . Again, the difference between d_{eff} and d is caused by the inhomogeneous spatial distribution of free and trapped carriers.

Several investigators have reported that some experimental results reveal the exponentially distributed traps scanned at levels above a certain discrete level E_{te} .⁷⁹⁻⁸² In this case, Equation 7-110 becomes

$$J = q^{1-\ell} \mu_p N_v \left(\frac{2\ell+1}{\ell+1} \right)^{\ell+1} \left(\frac{\ell}{\ell+1} \frac{\varepsilon}{H'_b} \right)^\ell \frac{V^{\ell+1}}{d_{\text{eff}}^{2\ell+1}} \quad (7-112)$$

The only difference is that H_b has been replaced with H'_b , which is given by

$$H'_b = H_b \exp(E_{te}/kT_c) \quad (7-113)$$

This implies that the distribution function for trap density is written as

$$h(E, x) = \frac{H_b}{kT_c} \exp[-(E - E_{te})/kT_c] S(x) \quad (7-114)$$

It is obvious that Equation 7-112 reduces to Equation 7-110 if $E_{te} = 0$, that is, the highest trap concentration is at the edge of the valence band.

Following the same procedure given for traps confined in single or multiple discrete energy levels, and using d instead of d_{eff} for simplicity, we obtain V_Ω and V_{TFL} for the present case.

When the density of thermally generated free carriers p_o inside the specimen (we are considering holes only) is larger than that of the injected carriers p , ohmic conduction is predominant. By setting $J = qp_o \mu_p \frac{V}{d}$ equal to Equation 7-112, we obtain the applied voltage required for the onset of the transition from ohmic to SCL conduction, which is given by

$$V_\Omega = \frac{qd^2 H'_b}{\varepsilon} \left(\frac{p_o}{N_v} \right)^{1/\ell} \left(\frac{\ell+1}{\ell} \right) \left(\frac{\ell+1}{2\ell+1} \right)^{\ell+1} \quad (7-115)$$

When all traps are filled, a transition from the trapped SCLC to a trap-free SCLC will take place. By setting Equation 7-74 equal to Equation 7-112, we obtain the TFL threshold voltage⁷⁸:

$$V_{TFL} = \frac{qd^2}{\varepsilon} \left[\frac{9}{8} \frac{H'_b}{N_v} \left(\frac{\ell+1}{\ell} \right)^\ell \left(\frac{\ell+1}{2\ell+1} \right)^{\ell+1} \right]^{\frac{1}{(\ell-1)}} \quad (7-116)$$

Traps Distributed in a Gaussian Manner within the Forbidden Energy Gap

In this case, Equation 7-68 becomes

$$h(E, x) = \frac{H_d}{(2\pi)^{1/2} \sigma_t} \exp \left[-\frac{(E - E_{tm})^2}{2\sigma_t^2} \right] S(x) \quad (7-117)$$

where $(2\pi)^{1/2}$ is the normalizing factor, E_{tm} is the hole-trapping energy level with a maximum trap density, and σ_t is the standard deviation of the Gaussian function.

Shallow traps—Hole traps are considered shallow if $E_{tm} < E_{fp}$. In this case, Equation 7-71 becomes⁸³

$$\begin{aligned} p_i &= \int_{E_t}^{E_\mu} h(E, x) g_p^{-1} \exp[(E - E_{fp})/kT] dE \\ &= H_d g_p^{-1} \exp(-E_{fp}/kT) \\ &\quad \times \exp \left[E_{tm}/kT - \frac{1}{2} (\sigma_t/kT)^2 \right] S(x) \\ &= pS(x)/\theta_d \end{aligned} \quad (7-118)$$

in which

$$\theta_d = \frac{g_p N_v}{H_d} \exp \left[-\frac{E_{tm}}{kT} + \frac{1}{2} \left(\frac{\sigma_t}{kT} \right)^2 \right] \quad (7-119)$$

Substitution of Equations 7-70 and 7-118 into Equation 7-69 gives

$$\frac{dF}{dx} = \frac{J}{\varepsilon \mu_p F} \cdot \frac{1}{\theta_d} [\theta_d + S(x)] \quad (7-120)$$

Integration of Equation 7-120 and use of the boundary condition

$$V = \int_0^d F(x) dx \quad (7-121)$$

give

$$J = \frac{9}{8} \epsilon \mu_p \theta_d \frac{V^2}{d_{\text{eff}}^3} \quad (7-122)$$

in which

$$d_{\text{eff}} = \left[\frac{3}{2} \int_0^d \left(\int_0^t [\theta_d + S(x)] dx \right)^{1/2} dt \right]^{2/3} \quad (7-123)$$

Equation 7-122 is similar in form to that for traps confined in a single discrete energy level see (Equation 7-87), except that θ_d has been replaced with θ_d . It is interesting to note that as $\sigma_t \rightarrow 0$, the case for Gaussian trap distribution approaches the case for traps confined in a simple discrete energy level, and that in the former case, the plot of $\ell n J$ as a function of $1/T$ may not be linear.

Deep traps—Hole traps are considered deep if $E_{tm} > E_{FP}$. By letting $z = E - E_{tm}$ and using appropriate approximations, Equation 7-71 becomes⁸³

$$\begin{aligned} p_t &\approx \frac{H_d S(x)}{(2\pi)^{1/2} \sigma_t} \int_0^\infty \frac{\exp(-z^2/2\sigma_t^2) dz}{1 + g_p \exp[(E_{FP} - E_{tm} - z)/kT]} \\ &= H'_d (p/N_v)^{1/m} S(x) \end{aligned} \quad (7-124)$$

in which

$$H'_d = (H_d/2) g_p^{-1} \exp(E_m/mkT) \quad (7-125)$$

and

$$m = [1 + 2\pi\sigma_t^2/16k^2T^2]^{1/2} \quad (7-126)$$

Substituting Equations 7-70 and 7-124 into Equation 7-69 and multiplying both sides by $(m+1)[F(x)]^{1/m/m}$, we obtain

$$\begin{aligned} \left(\frac{m+1}{m} \right) [F(x)]^{1/m} \frac{dF(x)}{dx} &= \frac{d[F(x)]^{(m+1)/m}}{dx} \\ &= \left(\frac{m+1}{m} \right) \frac{qH'_d}{\epsilon} \left(\frac{J}{q\mu_p N_v} \right)^{1/m} [\theta'_d + S(x)] \end{aligned} \quad (7-127)$$

in which

$$\theta'_d = \frac{g_p N_v}{H'_d} \exp \left[-\frac{E_{FP}}{kT} \left(\frac{m-1}{m} \right) \right]$$

If $p_t \gg p'_d$, θ'_d is very small and can be neglected. Integration of Equation 7-127 and use of the boundary condition given in Equation 7-121 give

$$J = q^{1-m} \mu_p N_v \left(\frac{2m+1}{m+1} \right)^{m+1} \left(\frac{m}{m+1} \right) \left(\frac{\epsilon}{H'_d} \right)^m \frac{V^{m+1}}{d_{\text{eff}}^{2m+1}} \quad (7-128)$$

in which

$$d_{\text{eff}} = \left\{ \frac{2m+1}{m+1} \int_0^d \left(\int_0^t S(x) dx \right)^{\frac{m}{m+1}} dt \right\}^{\frac{m+1}{2m+1}} \quad (7-129)$$

for $p_t \gg p$.

Equation 7-128 is similar in form to that for traps distributed exponentially within the forbidden gap (see Equation 7-110), except that ℓ has been replaced with m . To distinguish between the traps distributed in a Gaussian manner and those distributed exponentially, the technique of measuring thermally stimulated currents as functions of temperature and applied voltage can be used.^{62,84}

Following the same procedure given previously for traps confined in single or multiple discrete energy levels, and using d instead of d_{eff} for simplicity, we obtain V_Ω and V_{TFL} in this case as follows:

By setting $J = qp_o \mu_p \frac{V}{d}$ equal to Equation 7-122 and equal to Equation 7-128, we obtain

$$V_\Omega = \frac{8}{9} \frac{qp_o d^2}{\theta_d \epsilon} \quad (7-130)$$

for shallow traps and

$$V_\Omega = \frac{qd^2 H'_d}{\epsilon} \left(\frac{p_o}{N_v} \right)^{1/m} \left(\frac{m+1}{m} \right) \left(\frac{m+1}{2m+1} \right)^{\frac{m+1}{m}} \quad (7-131)$$

for deep traps.

To obtain V_{TFL} for shallow traps, we follow Equation 7-93, and the density of trapped holes in thermal equilibrium (in the absence of applied voltage) is

$$\begin{aligned} p_{t0} &= \int_{E_t}^{E_u} h(E) g_p^{-1} \exp[(E - E_{Fp0})/kT] dE \\ &= H_d g_p^{-1} \exp(-E_{Fp0}/kT) \\ &\quad \times \exp \left[E_m/kT - \frac{1}{2} (\sigma_t/kT)^2 \right] \end{aligned} \quad (7-132)$$

The density of unfilled traps is

$$\begin{aligned}
N_{t(\text{unfilled})} &= \int_{E_\ell}^{E_u} h(E) dE - p_{t0} \\
&= \frac{H_d}{2} [\text{erf}(E_u) - \text{erf}(E_\ell)] - p_{t0}
\end{aligned} \quad (7-133)$$

If $N_{t(\text{unfilled})} > p$ at V_{TFL} , Poisson's equation is

$$\frac{dF}{dx} = \frac{qN_{t(\text{unfilled})}}{\epsilon}$$

Thus

$$V_{TFL} = \int_0^d F dx = \frac{qN_{t(\text{unfilled})}d^2}{2\epsilon} \quad (7-134)$$

To obtain V_{TFL} for deep traps, we set Equation 7-74 equal to Equation 7-128 and obtain

$$V_{TFL} = \frac{qd^2}{\epsilon} \left[\frac{9}{8} \frac{H_d'^m}{N_v} \left(\frac{m+1}{m} \right)^m \left(\frac{m+1}{2m+1} \right)^{m+1} \right]^{\frac{1}{(m-1)}} \quad (7-135)$$

Traps Confined in Smeared Discrete Energy Levels

Earlier, we derived an expression for J as a function of V in the case of traps confined in a discrete energy level. However, this case could also be considered physically as a case with a Gaussian distribution having a very narrow trap energy deviation σ_t (i.e., $\sigma_t \ll kT$). Suppose that the trap energy level E_{tm} is located below E_{Fp} , which we considered shallow traps; the J - V characteristics can be easily derived from Equations 7-122 and 7-123 for $\sigma_t \ll kT$.

If the traps are confined not in a single smeared discrete energy level, but in multiple smeared discrete energy levels (such as a solid containing more than two kinds of impurities), Equations 7-122 and 7-123 are still valid, provided that θ_d is given by

$$\theta_d^{-1} = \sum_i \theta_i^{-1} \quad (7-136)$$

where

$$\theta_i = \frac{g_{pi}N_v}{H_{di}} \exp \left[-\frac{E_{ti}}{kT} + \frac{1}{2} \left(\frac{\sigma_{ti}}{kT} \right)^2 \right] \quad (7-137)$$

in which g_{pi} , H_{di} , σ_{ti} , and E_{ti} refer to g_p , H_d , σ_t , and E_t in the i th single smeared discrete energy level.

If the trap energy level E_{tm} is located above E_{Fp} in the case of deep traps, it is not possible to obtain an exact solution for the J - V characteristics. However, from Equations 7-126, 7-128, and 7-129, we can obtain an approximate solution for $\sigma_t \ll kT$. This is given by

$$J = \frac{9}{8} \epsilon \mu_p \left(\frac{N_v}{H_d'} \right) \bullet \frac{V^2}{d_{\text{eff}}^3} \quad (7-138)$$

Traps Distributed Uniformly within the Forbidden Energy Gap

This type of trap distribution was first investigated by Rose.⁷⁵ Although physically, the uniform distribution of traps within the forbidden energy gap is unlikely to occur in a solid, it is possible that traps due to impurities may not be confined in a single discrete energy level, but rather within a narrow band from E_ℓ to E_u in the forbidden energy gap. In such a case, Equation 7-68 can be written as

$$h(E, x) = H_c U(E - E_\ell) U(E_u - E) S(x) \quad (7-139)$$

where U is the Heaviside step-function. $U(E - E_\ell) = 0$ if $E < E_\ell$; $U(E_u - E) = 0$ if $E > E_u$; $U(E - E_\ell) = 1$ if $E > E_\ell$; $U(E_u - E) = 1$ if $E < E_u$; and H_c is the density of traps per unit energy interval. From Equations 7-71 and 7-139, we obtain

$$\begin{aligned}
p_t(x) &= \int_{E_\ell}^{E_u} \frac{H_c U(E - E_\ell) U(E_u - E) S(x) dE}{1 + g_p \exp[(E_{Fp} - E)/kT]} \\
&= H_c \left\{ \frac{1 + g_p \exp[(E_{Fp} - E_u)/kT]}{E_g + kT \ln \frac{1 + g_p \exp[(E_{Fp} - E_u)/kT]}{1 + g_p \exp[E_{Fp}/kT]}} \right\} S(x) \\
&= H_c kT \left(\frac{E_u - E_\ell - E_{Fp}}{kT} - \ln g_p \right) S(x)
\end{aligned} \quad (7-140)$$

By assuming $p_t \gg p$ and substituting Equation 7-140 into Equation 7-69, we obtain

$$\begin{aligned}
\frac{dF}{dx} &= \frac{q}{\epsilon} \left\{ p + H_c kT \left(\frac{E_u - E_\ell - E_{Fp}}{kT} - \ln g_p \right) \times \right. \\
&\quad \left. S(x) \right\} = \frac{qH_c kT}{\epsilon} S(x) \\
&\quad \times \ln \left\{ \frac{q\mu_p N_v g_p \exp[-(E_u - E_\ell)/kT] F}{J} \right\}
\end{aligned} \quad (7-141)$$

Integration of Equation 7-141 and use of the boundary condition

$$V = \int_0^d F(x) dx$$

give⁶⁹

$$J = 2q\mu_p N_v g_p \frac{V}{d_{\text{eff}}} \exp\left[-\frac{E_u - E_\ell}{kT}\right] \times \exp\left(\frac{2\varepsilon V}{qH_c k T d_{\text{eff}}^2}\right) \quad (7-142)$$

in which

$$d_{\text{eff}} = \left\{ 2 \int_0^d \int_0^\ell S(x) dx dt \right\}^{1/2} \quad (7-143)$$

Again, the difference between d_{eff} and d is caused by the inhomogeneous spatial distribution of free and trapped carriers. Obviously, if the traps are uniformly distributed from E_v to E_c in the forbidden energy gap, $E_u - E_\ell = E_g$.

Following the same procedure given for traps confined within single or multiple discrete energy levels, and using d instead of d_{eff} for simplicity, we obtain V_Ω and V_{TFL} for this case as follows:

By setting $J = qp_o\mu_p \frac{V}{d}$ equal to Equation 7-142, we obtain

$$V_\Omega = \frac{qH_c k T d^2}{2\varepsilon} \left\{ \ell n \left(\frac{p_o}{2g_p N_v} \right) + \frac{E_u - E_\ell}{kT} \right\} \quad (7-144)$$

For an accurate approach, we can obtain V_{TFL} by setting Equation 7-74 equal to Equation 7-142 and using graphical or computer techniques to calculate V_{TFL} . However, using the method of Muller,⁸⁵ which states that exponential trap distribution approaches uniform trap distribution when $\ell \rightarrow \infty$. Thus, by letting $\ell \rightarrow \infty$ in Equation 7-116, we obtain

$$V_{\text{TFL}} = \lim_{\ell \rightarrow \infty} \frac{qd^2}{\varepsilon} \left[\frac{9}{8} \frac{H_b^\ell}{N_v} \left(\frac{\ell+1}{\ell} \right)^\ell \left(\frac{\ell+1}{2\ell+1} \right)^{\ell+1} \right]^{\frac{1}{(\ell-1)}} \approx \frac{qH_b d^2}{2\varepsilon} = \frac{qH_c (E_u - E_\ell) d^2}{2\varepsilon} \quad (7-145)$$

since H_b is equivalent to $H_c(E_u - E_\ell)$ when $\ell \rightarrow \infty$.

The Scaling Rule

It can be seen from all expressions for J - V characteristics given in the previous cases that the general scaling rule^{54,86} for one-carrier SCL conduction in any material with any trap distributions can be expressed in the form of

$$\frac{J}{d_{\text{eff}}} = f\left(\frac{V}{d_{\text{eff}}^2}\right) \quad (7-146)$$

This equation is universally valid, provided that the carrier mobility is field independent and the effect of carrier diffusion is ignored. It is also valid for two-carrier (double-injection) space charge conduction if mobilities of both types of carriers are field independent and the effect of carrier diffusion is ignored. (High-field effects and the effect of carrier diffusion will be discussed later.) In Equation 7-146, the use of d_{eff} (effective specimen thickness) instead of d (true specimen thickness) in all expressions for J - V characteristics can be thought of as taking into account the effect of nonuniform spatial distribution of traps.

By writing Equation 7-146 in the following form

$$\frac{J}{d} \left(\frac{d}{d_{\text{eff}}} \right) = f \left[\left(\frac{F_{\text{av}}}{d} \right) \left(\frac{d}{d_{\text{eff}}} \right)^2 \right] \quad (7-147)$$

we derive the following features:

The factor $\left(\frac{d}{d_{\text{eff}}} \right)$ means that a solid with a

nonuniform spatial distribution of traps is equivalent to a solid with a uniform spatial distribution of traps if its true thickness d is replaced with an effective thickness d_{eff} .

J is directly related to the average field $F_{\text{av}} = V/d$, because J can always be written as

$$J = q\bar{p}\mu_p F_{\text{av}} \quad (7-148)$$

where $\bar{p}$ is the average density of free carriers. It is this $\bar{p}$ that produces different forms of J - V characteristics. It is obvious that $\bar{p}$ depends on the distribution of space charge, which in turn depends not only on the type of trap distribution, but also significantly on the interaction between the traps and the field, which is volume dependent.

$\frac{J}{d}$ can be interpreted as the flow of one unit volume of free charge carriers per second, while F_{av}/d can be written as $D/\epsilon d$, in which D is the average charge per unit area on the electrode and can be interpreted as the average charge density in the specimen having a dielectric constant ϵ . Thus, the flow of charge carriers per unit volume per second is directly related to the average charge density, which includes the free and trapped charged carriers in the specimen.

With $\frac{J}{d} \left(\frac{d}{d_{eff}} \right)$ expressed in terms of $\left[\left(\frac{F_{av}}{d} \right) \left(\frac{d}{d_{eff}} \right)^2 \right]^n$, the value of n will reflect the type of trap distribution. Table 7-2 sum-

marizes the values of n for all cases, and Table 7-3 summarizes all expressions for J - V characteristics for all cases discussed in this chapter.

The general expressions for the current-voltage characteristics in a solid with traps uniformly and nonuniformly distributed in space and in energy have been derived using a unified mathematical approach. The analytical technique discussed in this chapter may, in principle, be used to analyze any distribution of traps in space and energy. However, it should be noted that in the derivation, both permittivity and carrier mobility have been assumed to be constant. For a more rigorous treatment, these two physical parameters may have to be considered altered by the charge exchange in traps⁸⁷ and by high-field effects.

Table 7-2 The values of n in the factor $\left[\left(\frac{F_{av}}{d} \right) \left(\frac{d}{d_{eff}} \right)^2 \right]^n$ for single-injection J - V characteristics.

n	Description	References
$\frac{1}{2}$	Minority carrier electron injection into a p-type semiconductor Experimental results (e.g., single crystal trigonal selenium with indium electrodes)	Roberts ⁸⁸
1	Traps distributed uniformly within the forbidden energy gap Experimental results (e.g., amorphous Se films)	Touraine and Carles ⁸⁹
2	Without traps Experimental results (e.g., CdS)	Smith and Rose ⁷⁶
2	Traps confined in single discrete energy levels or in smeared discrete energy levels Experimental results (e.g., single crystal anthracene with silver electrodes, β -phthalocyanine with gold electrodes)	Helfrich ⁵³ Schadt and Williams ⁹⁰ Barbe and Westgate ⁹¹ Hwang and Kao ⁶⁹
2	Shallow traps distributed in a Gaussian manner within the forbidden energy gap Experimental results (e.g., amorphous Se films)	Lanyon ⁶⁰ Hwang and Kao ⁸³
$\ell + 1$	Traps distributed exponentially within the forbidden energy gap Experimental results (e.g., anthracene crystals, tetracene crystals)	Helfrich ⁵³ Mark and Helfrich ⁷⁸ Reucroft and Mullins ⁸¹ Baessler, Herrmann, Riehl, and Vaubel ⁹²
$m + 1$	Deep traps distributed in a Gaussian manner within the forbidden energy gap Experimental results (e.g., copper phthalocyanine films with gold electrodes)	Sussman ⁶² Hwang and Kuo ⁸³

Table 7-3 The expressions for single injection J - V characteristics with and without traps.

Case	Single Injection in Solids (for hole injection)	References
Trap-free	$J = \frac{9}{8} \epsilon \mu_p \frac{V^2}{d^3}$	Mott and Gurney ⁷²
Traps confined in a single discrete energy level	$J = \frac{9}{8} \epsilon \mu_p \theta_a \frac{V^2}{d_{\text{eff}}^3}$	Helfrich ⁵³ Lampert and Mark ⁵⁴ Hwang and Kao ⁵⁹
Traps distributed exponentially within the forbidden energy gap	$J = q^{1-\ell} \mu_p N_v \left(\frac{2\ell+1}{\ell+1} \right)^{\ell+1} \left(\frac{\ell}{\ell+1} \frac{\epsilon}{H'_b} \right)^{\ell} \frac{V^{\ell+1}}{d_{\text{eff}}^{\ell+1}}$	Mark and Helfrich ⁷⁸ Hwang and Kao ^{69,83} Reucroft and Mullins ⁸¹
Traps distributed in a Gaussian manner within the forbidden energy gap	$J = \frac{9}{8} \epsilon \mu_p \theta_d \frac{V^2}{d_{\text{eff}}^3}$ (for shallow traps) $J = q^{1-m} \mu_p N_v \left(\frac{2m+1}{m+1} \right)^{m+1} \left(\frac{m}{m+1} \frac{\epsilon}{H'_b} \right)^m \frac{V^{m+1}}{d_{\text{eff}}^{2m+1}}$ (for deep traps)	Bonham ⁶⁷ Hwang and Kao ⁸³
Traps distributed uniformly within the forbidden energy gap	$J = 2q\mu_p N_v g_p \frac{V}{d_{\text{eff}}} \exp\left[-\frac{E_u - E_t}{kT}\right] \exp\left[\frac{2eV}{qH_c k T d_{\text{eff}}^2}\right]$	Muller ⁸⁵ Rose ⁷⁵ Hwang and Kao ^{69,83}

The Effect of Carrier Diffusion

Previously, we assumed that our theoretical analysis may neglect the diffusion current component for mathematical simplicity. In this section, we shall examine the validity of this assumption and the condition under which such an assumption can be applied without causing a serious error.

The total current density for hole injection from the injecting contact (anode) at $x = 0$ is the sum of drift and diffusion current densities and is given by

$$J = J_{dr} + J_{diff} = q\mu_p \left(pF + V_T \frac{dp}{dx} \right) \quad (7-149)$$

where

$$V_T = \frac{D_p}{\mu_p} = \frac{kT}{q} \quad (7-150)$$

based on Einstein relation. From Equation 7-149, it is obvious that J_{diff} may be neglected if and only if $J_{dr} \gg J_{diff}$. This means

$$|pF| \gg \left| V_T \frac{dp}{dx} \right| \quad (7-151)$$

Supposing that the drift current component must be at least 10 times the diffusion current component before we can justify neglecting the diffusion current component, then we have

$$\left| \int_0^d F dx \right| = V = F_{av} d \geq 10 V_T \left| \int_0^d \frac{dp}{p} \right| \geq 10 V_T \ell n \frac{p(x=0)}{p(x=d)} \quad (7-152)$$

This equation implies that only when $p(x=0) = p(x=d)$ may we completely neglect the diffusion contribution, and that when $p(x=0) > p(x=d)$, the voltage V must be larger than $10 V_T \ell n[p(x=0)/p(x=d)]$ before we may neglect the diffusion term.

For example, if $p(x=0)/p(x=d) = 10^3$, the applied voltage must be larger than about $100 V_T$, that is, larger than 2.5 volts at 300 K. The magnitude of applied voltage alone does not give a clear criterion; it must go with specimen thickness. If we want to apply an average

field of 3 kV/cm across the specimen, specimen thickness must be larger than $2.5/3 \times 10^3 \approx 10^{-3}$ cm, or 10 μm . This simple example indicates that for a given applied average field F_{av} , the larger the specimen thickness is, the less important the diffusion contribution.

This is why the diffusion current component may sometimes be ignored for long specimens but not for thin specimens. For thin films, the diffusion current may become large enough to produce various effects. For example, diffusion tends to move the virtual anode (the plane of which $F = 0$), away from the hole injecting contact, thus shortening the effective specimen thickness and enhancing the field toward the cathode. The same argument is applicable to the case for only injected electron carriers.

However, for most insulators, the energy band gap is large; the carrier mobility, the carrier diffusion coefficient, and the thermally generated carrier density are all very small. Thus, for these materials, it is not difficult to satisfy the condition given by Equation 7-152 for neglecting the diffusion term, provided that the specimen thickness is not too thin.

7.5 Bulk-Limited Electrical Conduction Involving Two Types of Carriers

Similar to bulk-limited electrical conduction involving one type of carrier, in cases involving two types of carriers, one of the electrical contacts must be ohmic, acting as a carrier reservoir for one type of carrier (say electrons), and the other electrical contact must also be ohmic, acting as a reservoir for the other type of carrier (say holes). Thus, bulk-limited electrical conduction is governed by the bulk properties of the materials.

7.5.1 Physical Concepts of Carrier Trapping and Recombination

Because no perfect crystals exist in this world, there are always traps associated with various defects present in the solids. Sudden application of a step-function voltage across a specimen will force the electrons to inject from the

ohmic contact into the conduction band (or holes into the valence band) of the specimen, giving rise to a large current burst. If there are no traps, the space charge created by the injected carriers will remain in the conduction band (or the valence band), and the peak value of the transient current will not decay but will continue as a steady current. Such an ideal case, however, never happens. In actual specimens, there are always traps. After carrier injection, the free carriers will be captured by traps, thus causing the current to decay gradually to a steady-state value.

In general, the trap concentration (usually of the order of 10^{15} cm^{-3}) is higher than the free-carrier concentration. Smith and Rose⁷⁶ have used the experimental arrangement shown in Figure 7-18 to demonstrate that charge carriers injected into the specimen are trapped and remain in the specimen, even when both electrodes are grounded before the specimen is released and dropped into the electrometer pan.

However, the trap-free ideal case, based on the well known Mott-Gurney equation given by Equation 7-74, can only be observed when $V > V_{TFL}$, as shown theoretically in Figures

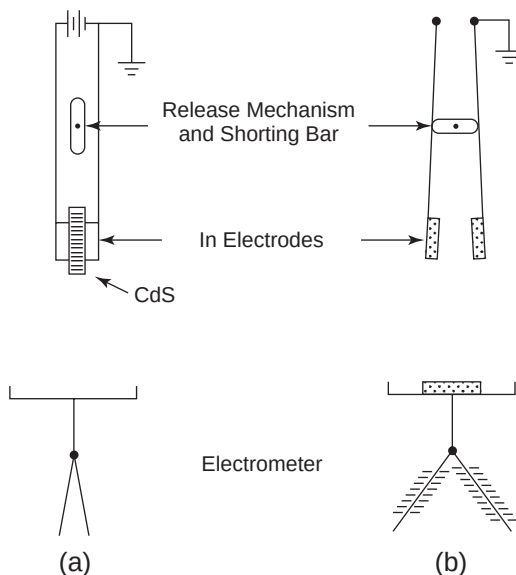


Figure 7-18 Schematic diagram illustrating the experimental arrangement for detecting injected trapped charges in an insulator.

7-16 and 7-17 and experimentally in Figure 7-19. For the deep traps discussed previously, the log-log plot of the J - V characteristics would form a Lampert triangle. The experimental result of Henderson, Ashley, and Shen,⁷⁷ shown in Figure 7-19, provides good experimental evidence of the Lampert triangle in the J - V characteristics of neutron-irradiated silicon, which contains radiation-created deep traps. It should be noted that the slope in the TFL region increases in verticality with increasing ratio of the unfilled equilibrium trap density to the equilibrium free-carrier density, and that TFL behavior is greatly affected by high fields and possibly by double injection.

Charge carriers injected into an insulator or a semiconductor through electron or hole emission (or both) from the contacting electrodes, or through absorption of light, will contribute to the increase in electrical conductivity. Under

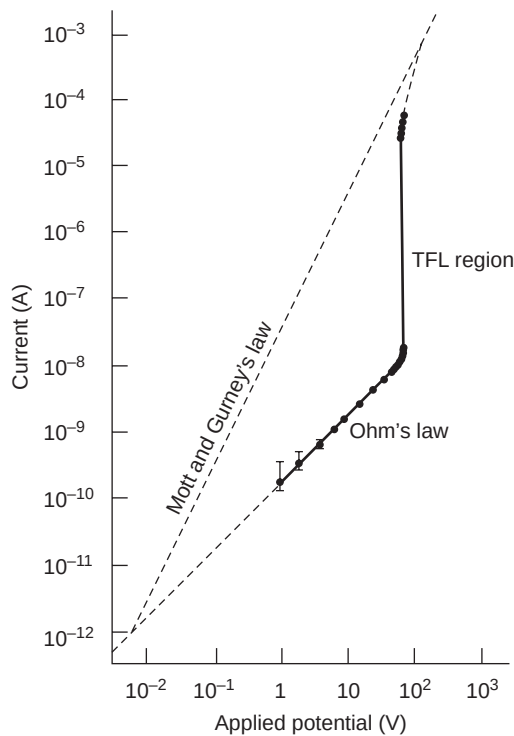


Figure 7-19 Experimental evidence of the Lampert triangle in silicon irradiated to 1.1×10^{16} neutrons cm^{-2} (> 0.1 MeV) with a steep transition at V_{TFL} . (After Henderson et al., 1972).

such a condition, $np > n_i^2$, implying that there are excess carriers in the materials, where $n_i^2 = n_o p_o$ under thermal equilibrium and n_o and p_o are the thermally generated electron and hole densities inside the material. Contrarily, it is also possible that $np < n_i^2$, but under this condition the carriers are extracted from the material. These two conditions are generally referred to as *nonequilibrium conditions*. However, when carrier density distributions are disturbed from their thermal equilibrium values, they tend to return to equilibrium through a recombination process (in the case of carrier injection) or through a generation process (in the case of carrier extraction).

Injected carriers in the material will be either temporarily captured at trapping centers or lost permanently through recombination centers. As there are no perfect crystals, completely free of defects or imperfections, materials always contain localized states, which may be confined in smeared discrete levels or distributed in the forbidden energy gap. These localized states form so-called trapping and recombination centers. Therefore, trapping and recombination processes play one of the most important roles in single or double injection, in photoconduction and luminescence in solids (insulators and semiconductors), and in all solid-state electron devices. This section is devoted to clarifying some physical concepts related to these processes.

For single injection, presented in Section 7.4.2, we were concerned mainly with traps acting as trapping centers, but for double injection, the so-called traps, in the case of single injection, may act as trapping centers (or simply traps) or as recombination centers. To avoid confusion, we will start with a clear definition of these two centers. A *trapping center* (or simply a *trap*) is a center that captures a free carrier; after a while, this captured (or trapped) carrier has a greater probability of being thermally reexcited to the nearest allowed band to become a free carrier again than of recombining with a carrier of the opposite sign at the center. Trapping centers that capture electrons only are called *electron traps*, and those capturing holes only are *hole traps*. The occupancy

of such centers is determined by the thermal equilibrium interchange of the carriers of one particular sign between the centers and the nearest allowed band.

A *recombination center* is a center that also captures a free carrier, but the captured carrier has a greater probability of recombining with a carrier of the opposite sign, resulting in the annihilation of both, than of being thermally reexcited to the nearest allowed band. Those recombination centers in which the localized states are normally empty capture electrons first and then recombine with holes; those in which the localized states are normally filled, capture holes (or, in other words, give up electrons to the valence band) first and then recombine with electrons. The occupancy of such centers is governed by the kinetic recombination processes.

A localized state may act as a trapping or a recombination center, depending on its location in the forbidden energy gap (governed by the nature of impurities, defects, and temperature), the concentrations of free electrons and holes, and the capture cross sections of electrons and holes. Thus, the distinction between a trapping and a recombination center is a quantitative rather than a qualitative one.

In the steady state, the rate of generation of electrons and holes must be equal to the rate of recombination, and the rate of trapping must be equal to the rate of detrapping (or reexcitation). To analyze carrier transport problems involving these processes, it is important to set a quantitative criterion to separate trapping and recombination centers. Rose^{73,75} has used *demarcation levels* to separate them. The demarcation level for electron traps, E_{Dn} , is defined as the level at which a captured electron has an equal probability of being excited into the conduction band and of recombining with a hole from the valence band. Similarly, the demarcation level for hole traps, E_{Dp} , is defined as the level at which a captured hole has an equal probability of being excited into the valence band and of recombining with an electron from the conduction band. The localized states located between E_C and E_{Dn} act predominantly as electron traps; those located between E_v and E_{Dp} act predominantly as hole traps; and those located between

E_{Dp} and E_{Dn} act predominantly as recombination centers.

Capture Rates and Capture Cross-Sections

The electron capture rate is defined as the rate at which electrons are captured from the conduction band by traps following the equation

$$\frac{dn}{dt} = -C_n n N_n \quad (7-153)$$

where n is the free (or conduction) electron density in the conduction band, N_n is the density of empty electron traps, and C_n is the electron capture rate constant (or simply the electron capture coefficient). The capture cross-section of an electron trapping center σ_n is defined as a cross-section through which a moving electron must come to the center to be captured. Assuming that all the electrons have the same energy E and the same velocity v , then within a time Δt the volume of space through which electrons pass and where they will be captured is $\sigma_n v \Delta t n$, and the number of electrons per unit volume to be captured within a time Δt is

$$\text{and } \left. \begin{aligned} \Delta n &= v \Delta t \sigma_n N_n \\ v &= (2E/m_e^*)^{1/2} \end{aligned} \right\} \quad (7-154)$$

From Equations 7-153 and 7-154, we obtain

$$C_n = v \sigma_n \quad (7-155)$$

Experimentally, we do not measure σ_n directly but measure C_n . The measured value of C_n is an average value of $C_n(E)$, taking into account the actual energy distribution of electrons. Thus, for a thermal equilibrium distribution, we may write

$$C_n = \langle v \rangle \sigma_n = \langle v \sigma_n \rangle \quad (7-156)$$

where $\langle v \rangle$ is the average velocity of the carriers, which is given by

$$\begin{aligned} \langle v \rangle &= \frac{\int_0^\infty \left(\frac{2E}{m_e^*} \right)^{1/2} f(E) g(E) dE}{\int_0^\infty f(E) g(E) dE} \\ &= \left(\frac{2}{m_e^*} \right)^{1/2} \frac{\int_0^\infty E \exp(-E/kT) dE}{\int_0^\infty E^{1/2} \exp(-E/kT) dE} \\ &= (4kT/\pi m_e^*)^{1/2} \end{aligned} \quad (7-157)$$

and

$$\begin{aligned}\sigma_n &= \frac{\langle v\sigma_n(E) \rangle}{\langle v \rangle} \\ &= \frac{\int_0^\infty v\sigma_n(E)f(E)g(E)dE}{\int_0^\infty vf(E)g(E)dE} \bigg/ \frac{\int_0^\infty f(E)g(E)dE}{\int_0^\infty f(E)g(E)dE} \\ &= \frac{\int_0^\infty E\sigma_n(E)\exp(-E/kT)dE}{\int_0^\infty E\exp(-E/kT)dE}\end{aligned}\quad (7-158)$$

in which the electron energy E is measured from E_c (the edge of the conduction band).

The thermal velocity of the carriers is given by

$$v_{th} = (3kT/m_e^*)^{1/2} \quad (7-159)$$

The electron capture cross-section most frequently quoted in the literature is the root mean square cross-section.⁹³

$$\sigma_{n(r,m,s)} = C_n/(3kT/m_e^*)^{1/2} \quad (7-160)$$

Similarly, the hole capture rate constant (or simply the hole capture coefficient) C_p and the hole capture cross-section σ_p can be expressed as

$$C_p = \langle v \rangle \sigma_p = \langle v\sigma_p \rangle \quad (7-161)$$

$$\sigma_p = \frac{\langle v\sigma_p(E) \rangle}{\langle v \rangle} \quad (7-162)$$

where

$$\langle v \rangle = (4kT/\pi m_h^*)^{1/2} \quad (7-163)$$

Recombination Rates and Recombination Cross-Sections

Recombination occurs by

Direct band-to-band recombination of free electrons and free holes not involving recombination centers

Indirect recombination through recombination centers as a stepping—stone: free carriers of one type being captured first at the centers and then recombined with free carriers of opposite sign

Theoretically, both recombination mechanisms exist simultaneously, but in most cases indirect recombination is predominant. Direct band-to-

band recombination becomes important only when both electron and hole densities are high.

The direct band-to-band recombination rate R can be defined by the following equation:

$$\frac{dn}{dt} = \frac{dp}{dt} = -R = -C_r np \quad (7-164)$$

As with the expressions for C_n and C_p , we can write the direct band-to-band recombination rate constant C_r as

$$C_r = \langle v\sigma_R \rangle \quad (7-165)$$

where v , in this case, is the microscopic relative velocity of an electron and a hole, and σ_R is their recombination cross-section. Thus, the measured value of C_r is the average value of $v\sigma_R$ over the two velocity distributions.

For indirect recombination through a set of acceptor-type recombination centers, the centers will capture electrons first and then recombine with holes. The rate of capturing electrons at the centers must be equal to the rate of capturing holes at the centers for recombination there. Thus, the recombination rate is

$$\begin{aligned}R_a &= \langle v\sigma_n \rangle n(N_{ra} - n_{ra}) \\ &= \langle v\sigma_p \rangle p n_{ra}\end{aligned}\quad (7-166)$$

where N_{ra} and n_{ra} are, respectively, the densities of total acceptor-type recombination centers, including occupied (or filled) and unoccupied (or empty) localized states, and captured electrons (filled localized states).

If the recombination centers are of donor type, the centers will capture holes first and then recombine with electrons there. In this case, the recombination rate is

$$\begin{aligned}R_d &= \langle v\sigma_n \rangle n n_{rd} \\ &= \langle v\sigma_p \rangle p (N_{rd} - n_{rd})\end{aligned}\quad (7-167)$$

where N_{rd} and n_{rd} are, respectively, the densities of total donor-type recombination centers, including occupied (or filled) and unoccupied (or empty) localized states, and captured holes (empty localized states).

In Equation 7-164 we ignored the effect of thermally generated carriers $n_o p_o = n_i^2$, and in Equations 7-166 and 7-167 we ignored the probability of thermal reexcitation of captured carriers in recombination centers to the nearest

allowed band (instead of recombination with carriers of the opposite sign). However, for large energy-gap materials, such as organic crystals, n_o and p_o are small and can be ignored without causing a great error in most cases. If the recombination centers are far away from E_{Fn} for the acceptor type or far away from E_{Fp} for the donor type, so that the thermal reexcitation may be ignored, then Equations 7-166 and 7-167 are valid.

Demarcation Levels

First we will consider acceptor-type centers of density N_n located at an energy level E_t in the forbidden energy gap. In thermal equilibrium, if there are no other sinks to take the free electrons away, the rate of capturing free electrons by the empty centers is equal to the rate of thermally reexciting the captured electrons from the occupied centers to the conduction band. Thus, we can write

$$n\langle v\sigma_n \rangle (N_n - n_t) = n_t v_n \exp[-(E_c - E_t)/kT] \quad (7-168)$$

where n_t is the trapped electron density and v_n is the attempt-to-escape frequency which, in a classical physical concept, represents the number of times per second a captured electron attempts to absorb sufficient energy from the lattice vibration and surmount the potential barrier of the trap. By expressing n , n_t , and $(N_n - n_t)$ as

$$n = N_c \exp[-(E_c - E_{Fn})/kT] \quad (7-169)$$

$$n_t = N_n \{1 + g_n^{-1} \exp[(E_t - E_{Fn})/kT]\}^{-1} \quad (7-170)$$

$$N_n - n_t = N_n \{1 + g_n \exp[(E_{Fn} - E_t)/kT]\}^{-1} \quad (7-171)$$

Substitution of Equations 4-169 through 7-71 into Equation 7-168 yields

$$v_n = g_n^{-1} N_c \langle v\sigma_n \rangle \quad (7-172)$$

The rate of thermal reexcitation of captured electrons is

$$g_n^{-1} n_t N_c \langle v\sigma_n \rangle \exp[-(E_c - E_t)/kT]$$

To derive an expression for the electron demarcation level E_{Dn} , we assume $g_n = 1$ for simplic-

ity and replace E_t with E_{Dn} for $E_t = E_{Dn}$, then set the rate of thermal reexcitation of captured electrons to the conduction band equal to the rate of recombination of these captured electrons with free holes from the valence band. Thus, we can write

$$n_t N_c \langle v\sigma_n \rangle \exp[-(E_c - E_{Dn})/kT] = n_i p \langle v\sigma_p \rangle \quad (7-173)$$

From Equations 7-169 and 7-173, we obtain

$$E_{Dn} = E_{Fn} + kT \ln \left(\frac{p\sigma_p}{n\sigma_n} \right) \quad (7-174)$$

Using the same argument, if the same localized states are located below the middle of the forbidden energy gap, then these states may act as hole traps and we can set the rate of thermal reexcitation of captured holes to the valence band equal to the rate of recombination of these captured holes ($N_n - n_t$) with free electrons from the conduction band.

$$(N_n - n_t) N_v \langle v\sigma_p \rangle \exp[-(E_{Dp} - E_v)/kT] = (N_n - n_t) n \langle v\sigma_n \rangle \quad (7-175)$$

Thus, we obtain

$$E_{Dp} = E_{Fp} + kT \ln \left(\frac{p\sigma_p}{n\sigma_n} \right) \quad (7-176)$$

Derivation of Equations 7-174 and 7-176 is based on the assumption that $\langle v\sigma_n \rangle = v\sigma_n$ and $\langle v\sigma_p \rangle = v\sigma_p$.

There is one set of demarcation levels (E_{Dn} and E_{Dp}) for a particular type of imperfection (or for one set of localized states), characterized by a particular set of capture cross-sections (σ_n and σ_p), as shown in Figure 7-20. From Equations 7-174 and 7-176 we have

$$E_{Dn} - E_{Fn} = E_{Dp} - E_{Fp} \quad (7-177)$$

Suppose that electron traps are distributed exponentially between E_{te1} and E_{te2} within the forbidden energy gap, following the equation

$$h(E) = \frac{H_b}{kT_c} \exp[-(E_{te1} - E)/kT_c] \quad (7-178)$$

as shown in Figure 7-20(a). These traps will act as electron traps for single injection, but their behavior will be quite different for double injection. The following list shows their

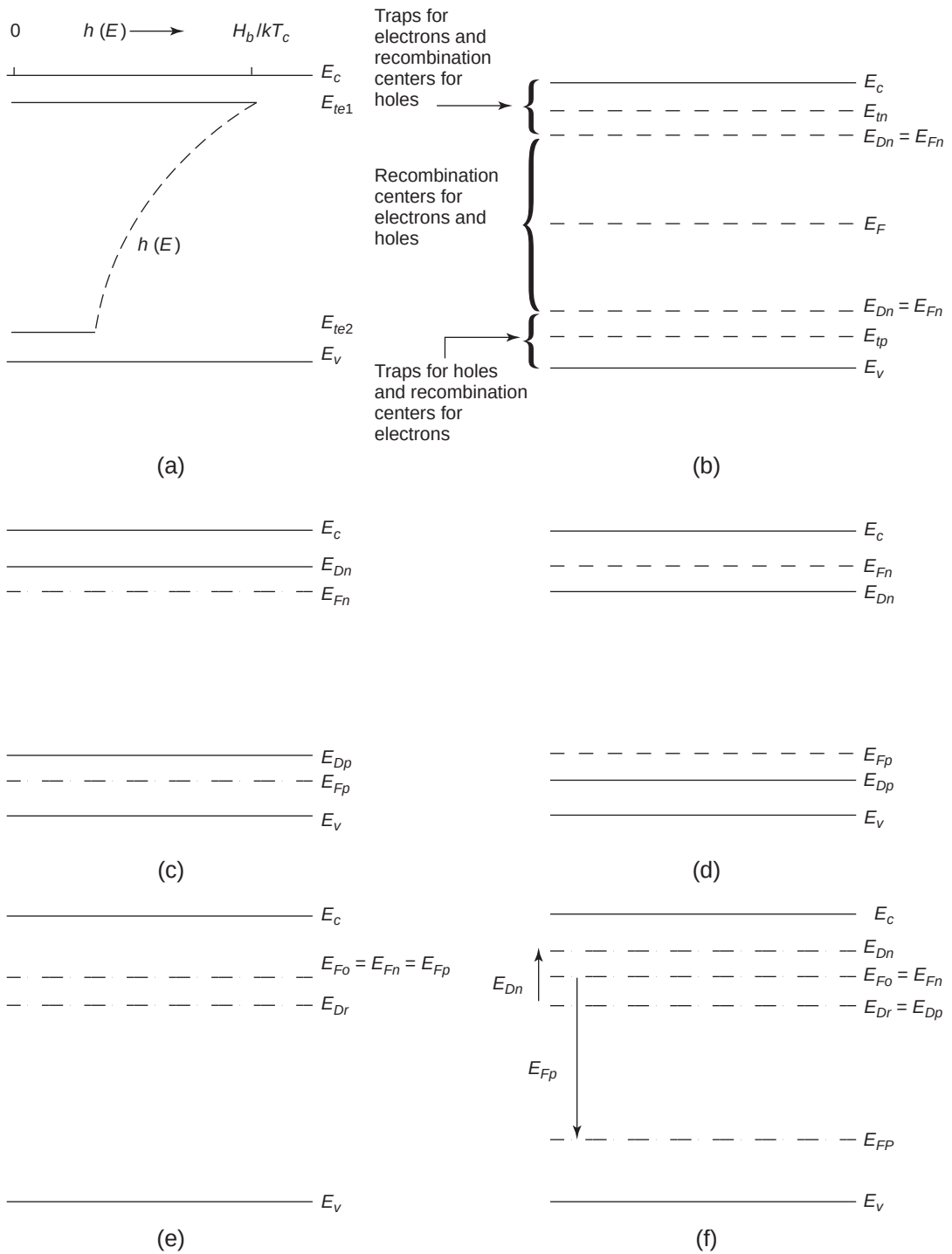


Figure 7-20 (a) Exponential electron trap density distribution function; (b) demarcation levels, Fermi levels, energy levels for trapping and recombination centers for $n\sigma_n = p\sigma_p$; (c) for $n\sigma_n < p\sigma_p$; (d) for $n\sigma_n > p\sigma_p$; (e) for an n-type semiconductor with $n_o > \Delta n$, $p_o > \Delta p$, and $n_o > p_o$; (f) for an n-type semiconductor with $n_o > \Delta n$, $p_o < \Delta p$, and $n_o > p_o$.

behavior under various conditions, based on Equations 7-174, 7-176, and 7-177.

- When the injected carrier densities (injected by either optical or electrical means) exceed the thermally generated carrier densities, as in most insulators ($n > n_o$ and $p > p_o$), the carrier densities n and p are generally expressed in terms of quasi-Fermi levels E_{Fn} and E_{Fp} under such nonequilibrium conditions. If the injected carrier densities are of such values that $n\sigma_n = p\sigma_p$, then the electron demarcation level E_{Dn} coincides with E_{Fn} , and the hole demarcation level E_{Dp} coincides with E_{Fp} , as shown in Figure 7-20(b). If the injected carrier densities are of such values that $n\sigma_n < p\sigma_p$, then E_{Dn} separates from E_{Fn} and is located above it by an amount of $kT\ln(p\sigma_p/n\sigma_n)$, and from Equation 7-177, E_{Dp} is located above E_{Fp} by the same amount, as shown in Figure 7-20(c). Similarly, if $n\sigma_n > p\sigma_p$, E_{Dn} and E_{Dp} will be, respectively, located below E_{Fn} and E_{Fp} by an amount of $kT\ln(n\sigma_n/p\sigma_p)$, as shown in Figure 7-20(d). In the interval between E_{Dn} and E_{Dp} , the occupancy of the localized states—the so-called *traps* in Figure 7-20(a)—is determined by kinetic recombination processes. Therefore, the localized states within this interval act as recombination centers. The occupancy of the localized states above E_{Dn} is determined by E_{Fn} , and these states act as electron trapping centers. The occupancy of the localized states below E_{Dp} is determined by E_{Fp} , and these states act as hole traps. Each set of recombination centers has its own set of demarcation levels, and when there are two different types of imperfections (each characterized by its own pair of capture cross-sections σ_n and σ_p), the demarcation levels are displaced independently for each set from the common quasi-Fermi levels E_{Fn} and E_{Fp} .⁷³ It can also be seen that in single injection, the electron traps (e.g., acceptor-type) distributed as shown in Figure 7-20(a) act as electron traps, but in double injection only the traps located above E_{Dn} can be considered electron traps. The rest act as recom-

bination centers between E_{Dn} and E_{Dp} , and as hole traps between E_{Dp} and E_v .

- With an increase in carrier injection, either by increasing the light intensity in optical excitation or by increasing the applied field in double electrical contact injection, both E_{Fn} and E_{Fp} and E_{Dn} and E_{Dp} will be shifted toward the band edges. Therefore, some localized states acting as traps will be transformed into recombination centers. This process in producing recombination centers is sometimes referred to as *electronic doping* because it involves electronic excitation or injection.
- An increase in temperature will shift the quasi-Fermi levels, and hence the demarcation levels, away from the band edges. So some localized states acting as recombination centers will be transformed into trapping centers.
- When injected carrier densities are smaller than thermally generated carrier densities, then n_o and p_o are predominant in the conduction and valence bands. In this case, the quasi-Fermi levels coincide with each other, $E_{Fn} = E_{Fp} = E_{Fo}$, and the demarcation levels also coincide with each other, $E_{Dn} = E_{Dp} = E_{Dr}$, as shown in Figure 7-20(e) for n-type semiconductors. When E_{Dr} is considered E_{Dp} , all electron-occupied states located between E_{Dr} and E_{Fo} act mainly as recombination centers for holes, and those above E_{Fo} as electron traps and below E_{Dr} as hole traps. But when E_{Dr} is considered E_{Dn} , the unoccupied localized states located above E_{Dr} act as electron traps and those below E_{Dr} as recombination centers for electrons.
- If the injected electron density is still smaller than the thermally generated electron density n_o but the injected hole density is larger than the thermally generated hole density p_o , then E_{Fn} remains practically at E_{Fo} but E_{Fp} separates from E_{Fo} and moves toward E_v . At the same time, E_{Dp} remains practically at E_{Dr} but E_{Dn} moves toward E_c at the same rate E_{Fp} moves toward E_v , as shown in Figure 7-20(f).
- If the injected electron density is much larger than n_o and p_o and there is no hole

injection, such as single injection (electron injection), into an insulator, then the term $kT\ln(n\sigma_n/p\sigma_p)$ becomes very large, much larger than E_g and E_{Dn} , and E_{Dp} and E_{Fp} will disappear in the forbidden energy gap. In this case, the traps located above E_{Fn} are shallow traps and those below E_{Fn} are deep traps. There are no recombination centers.

- It should be noted that E_{Fn} , E_{Fp} , E_{Dn} , and E_{Dp} are functions of distance from the injecting contacts, which are not shown in Figure 7-20 for clarity.

Coulombic Traps

Trapping can be considered a process of energy storage by spatially localizing electrons and holes at certain sites so as to hinder their free movement, that is, to stop them from contributing to electrical conduction. These captured electrons and holes may be released to be free again by absorbing sufficient thermal or optical energy, or they may be lost through recombination by giving up their stored energy.

Any centers formed by localized states capable of capturing carriers are called *traps* or *trapping centers*. After the capture of a carrier, the subsequent action determines whether the trap acts as a trapping center or as a recombination center. This has been discussed earlier in Section 7.5.1.

A trap can be considered an entity with a certain charge—positive, neutral or negative—when empty or unoccupied. With M to represent one trap, e to represent an electron, h to represent a hole, and the superscript to denote the charge—(–) negative, (o) neutral, and (+) positive—Table 7-4 shows that the behavior of the traps can be grouped into three types: coulombic attractive centers, coulombic neutral centers, and coulombic repulsive centers. The variation of potential energy of these three types of coulombic traps is shown schematically in Figure 7-21. In general, the trap densities in solids range from about 10^{12}cm^{-3} in highly pure single crystals to about 10^{19}cm^{-3} in imperfect large bandgap insulators; the capture cross-sections range from about 10^{-11}cm^2 for

Table 7-4 The trapping and detrapping processes for three types of traps.

Coulombic trapping center	Type of carrier to be trapped	Trapping and detrapping processes		Remarks
		Charge before trapping (trap unoccupied)	Charge after trapping (trap filled)	
Attractive	Electron (e)	$M^+ + e$	M^o	Deep donors with compensation for electron traps. The detrapping process is known as the Poole–Frenkel effect.
Attractive	Hole (h)	$M^- + h$	M^o	Deep acceptors with compensation for hole traps. The detrapping process is known as the Poole–Frenkel effect.
Neutral	Electron (e)	$M^o + e$	M^-	The field dependence of the detrapping process for electrons or for holes is small.
Neutral	Hole (h)	$M^o + h$	M^+	
Repulsive	Electron (e)	$M^- + e$	M^{--}	Double donors or double acceptors with one level compensated.
Repulsive	Hole (h)	$M^+ + h$	M^{++}	

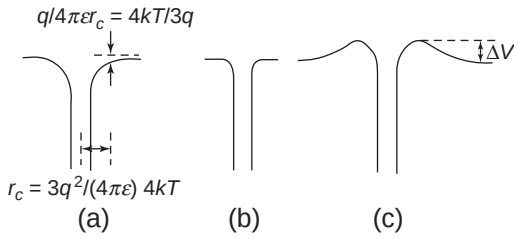


Figure 7-21 Schematic diagrams illustrating the variation of potential energy of (a) coulombic attractive center, (b) coulombic neutral center, and (c) coulombic repulsive center.

coulombic attractive centers to about or less than 10^{-21} cm^2 for coulombic repulsive centers; and the carrier lifetimes based on a carrier thermal velocity of 10^7 cm/sec could range from 10^2 sec to 10^{-14} sec .⁹⁴

The capture cross-section of a center is determined by the variation of potential energy in the vicinity of the center. For attractive centers, we can assume that a free carrier will be captured when it approaches the center at a distance r from the center having a drift velocity

$$v_d = F\mu = \frac{q\mu}{4\pi\epsilon r^2} \quad (7-179)$$

The average diffusion velocity of a particle executing Brownian motion at a distance r from the center is given by

$$\bar{v} = \frac{2\lambda}{3r} v \quad (7-180)$$

By equating v_d and $\bar{v}$, we obtain the radius of the attractive center⁷³

$$r_c = \frac{3q\mu}{(4\pi\epsilon)2\lambda v} \quad (7-181)$$

Since

$$\left. \begin{aligned} v &= \lambda/\tau \\ \mu &= q\tau/m^* \\ kT &= m^*v^2/2 \end{aligned} \right\} \quad (7-182)$$

the capture cross-section can be written as

$$\begin{aligned} (\sigma)_{\text{attractive}} &= \pi r_c^2 \\ &= \frac{9\pi q^4}{(4\pi\epsilon)^2 16(kT)^2} \end{aligned} \quad (7-183)$$

where λ is the carrier mean free path between collisions for energy loss, τ is the carrier lifetime, and v is the carrier thermal velocity. If $\lambda < r_c$, many collisions can occur within the critical radius and the interaction between the center and the carrier is diffusion limited.⁹³ Equations 7-181 and 7-183 are derived on the basis of this assumption. For this condition $(\sigma)_{\text{attractive}} \propto (\epsilon)^{-2}(T)^{-2}$ and in most materials with a large band gap and a lower carrier mobility $\lambda < r_c$, as in organic semiconductors or dielectric materials. However, for silicon and most inorganic semiconductors $\lambda > r_c$, then Equation 7-183 is not applicable but can be used as a guide by multiplying it by the ratio of $2r_c/\lambda$.⁷³ Thus, we have

$$(\sigma)_{\text{attractive}}^{(\lambda > r_c)} \approx \frac{2r_c}{\lambda} (\pi r_c^2) \quad (7-184)$$

This implies that the capture cross-section is roughly proportional to r_c^3 and T^{-2} if λ can be assumed to be proportional to μ , which is proportional to $T^{-3/2}$. For $\lambda > r_c$, it can be imagined that a free carrier may pass through the center a number of times before being captured. This implies that the effective capture cross-section is reduced. If such an attractive center is not deformed by the field, the radius of the center is approximately proportional to $F^{-1/2}$, so its capture cross-section should decrease with field as $F^{-3/2}$.⁹⁵

It should be noted that the rapid increase in capture cross-section with decreasing temperature may be associated with a large capture rate into highly excited states, followed by a cascade process in which a certain fraction of the captured carriers reaches the ground state.⁹³

For neutral centers, the polarizability of the center provides a quasi-long range interaction with the carrier. A charge q at a distance r from the center with polarizability α will produce a dipole moment of $\alpha q/4\pi\epsilon r^2$. This dipole will in turn produce an attractive force of $2(\alpha q/4\pi\epsilon r^2)(q/4\pi\epsilon r^3)$ on the charge, so an attractive potential is

$$V(r) = -A/r^4 \quad (7-185)$$

where A is a constant equal to $\alpha q^2/2(4\pi\epsilon)^2$. Different centers (different impurities) in the

same host lattice or the same impurities in different host lattices will have different values of α . Since the radius r at which the potential energy is kT varies with $T^{-1/4}$, the capture cross-section will be less sensitive to temperature. Since the capture cross-section is much smaller than that of the attractive center, it is more sensitive to the potential profile at short ranges, so it is more sensitive to the chemical nature of the center.⁹³

For repulsive centers, either $M^- + e \rightarrow M^-$ or $M^+ + h \rightarrow M^{++}$, the capturing action is repulsive. An electron approaching a center that has already captured an electron will see a repulsive potential barrier that it must surmount thermally or tunnel through before being captured. Figure 7-21(c) shows that when the free electron reaches the potential top (ΔV), it sees an attractive potential field. Such a trapping process indicates that the capture cross-section of the repulsive center would be very small and very sensitive to temperature. The increase in electric field causes an increase in the number of high-energy electrons to overcome the repulsive potential barrier and be trapped. This phenomenon is generally referred to as *field-enhanced trapping*, which produces a negative differential resistance region and current oscillations in n-type GaAs.⁹⁶⁻⁹⁸

In general, $\sigma_{\text{attractive}} > \sigma_{\text{neutral}} > \sigma_{\text{repulsive}}$. When an electron is captured, it must lose its energy, which is carried away by one of the following:

- A photonradiative capture or recombination
- Phonons (optical and acoustic phonons), that is, nonradiative capture or recombination
- Another electron or hole, that is, Auger recombination

Characteristic Times

There are four characteristic times commonly used in the literature to describe various quantities. It is therefore necessary to have a clear definition and physical concept of them.

Carrier Lifetime τ

The carrier lifetime (or simply the lifetime) is generally referred to as the time during which a charge carrier is free to move and so to con-

tribute to electric conduction. We can define τ_n as the time that an excited electron spends in the conduction band and τ_p as the time that an excited hole spends in the valence band. Supposing that a uniform excitation generates G electron-hole pairs per second per unit volume in a solid, then the generated electron and hole densities in the conduction band and valence band are, respectively

$$\Delta n = G\tau_n \quad (7-186)$$

and

$$\Delta p = G\tau_p \quad (7-187)$$

If these carriers are trapped and then thermally reexcited, the time spent in the traps is not included in τ_n and τ_p . In the steady state, the rate of generation is equal to the rate of trapping. Thus, we have

$$\tau_n = \frac{1}{\langle v\sigma_n \rangle (N_r - n_r)} \quad (7-188)$$

$$\tau_p = \frac{1}{\langle v\sigma_p \rangle n_r} \quad (7-189)$$

where N_r and n_r are, respectively, the total (occupied and unoccupied) and the occupied trapping or recombination centers. These equations are valid provided that the carrier mean free path is larger than the diameter of the capture cross-section ($2\sqrt{\sigma_n/\pi}$ or $2\sqrt{\sigma_p/\pi}$). For small mean free paths, τ_n or τ_p will be increased by a factor of the order of the ratio of half the spacing between capturing centers to the mean free path.⁷³

If $\Delta n \propto G$ or $\Delta p \propto G$, τ_n or τ_p is constant. This implies that n and p are smaller than the density of trapping or recombination centers. This is true for most insulators, in which the density of localized states is usually greater than 10^{15} cm^{-3} and n or p less than 10^{15} cm^{-3} . In this case, photoconductivity is linear with light intensity. However, in some materials and under certain conditions Δn or Δp may vary with G^a . Then the photoconductivity is said to be superlinear if $a > 1$ and sublinear if $a < 1$. In such cases, τ_n and τ_p are no longer constant but depend also on G .

Several *lifetime* terms that are frequently used in the literature are defined as follows^{99,100}:

Free lifetime: The lifetime of a free carrier excluding any time spent by the carrier in the traps

Excited lifetime: The total lifetime of an excited carrier, including both the free lifetime and the time spent in the traps (trapping time or capturing time)—in other words, the total time between the action of excitation and the action of recombination

Minority carrier lifetime: The free lifetime of a minority carrier, electron or hole, present in lower density

Majority carrier lifetime: The free lifetime of a majority carrier, electron or hole, present in higher density. If free-carrier densities n and p are much larger than the density of recombination centers, the majority carrier lifetime is equal to the minority carrier lifetime. If n and p are smaller than the density of recombination centers, as in most insulators, the majority carrier lifetime is much larger than the minority carrier lifetime.

Electron–hole pair lifetime: The free lifetime of the carrier (usually the minority carrier) first captured in traps

Diffusion-length lifetime (or recombination lifetime), τ_o : Based on the relation $\tau_o = L_o^2/D_o$ where D_o is the ambipolar diffusion coefficient and L_o is the diffusion length

Dielectric Relaxation Time τ_d

The dielectric relaxation time is defined as the time necessary for the reestablishment of quasi-neutrality after an injection of carriers into the solid. The dielectric relaxation time for electrons is $\tau_d = \epsilon/q\mu_n n$; for holes, it is $\tau_d = \epsilon/q\mu_p p$.

Carrier Transit Time t_t

Carrier transit time is defined as the time required for a carrier to travel across a specimen, which includes the total time spent as a free carrier and the total time spent as a trapped carrier in the traps during the transit. It is obvious that, only for $\tau_d < t_t$ and $\tau_d < \tau_o$, the condition of local space charge neutrality can be used as a good approximation for electric transport calculations.

Response Time τ_r

Response time is defined as the time required for a transient current to reach a steady-state value (or an appropriate fraction of the steady-state value, such as $1 - 1/e$) after light excitation is switched on. This is also the time required for the photocurrent to decay to the fraction $(1/e)$ of its steady-state value after light excitation is switched off. Response times are given by

$$\tau_{rn} = \left(1 + \frac{n_t}{n}\right) \tau_n \quad (7-190)$$

$$\tau_{rp} = \left(1 + \frac{p_t}{p}\right) \tau_p \quad (7-191)$$

Two things can be seen:

If there are no trapping or recombination centers, or if the free-carrier densities n and p are much larger than the density of trapping or recombination centers, the response time is equal to the carrier lifetime.

If n and p are comparable or less than the density of trapping or recombination centers, $\tau_{rn} > \tau_n$ and $\tau_{rp} > \tau_p$.

Therefore, a temperature rise or an increase of light intensity may reduce the difference between the response time and the carrier lifetime. $\tau_r > \tau$ because carrier injection must supply not only carriers into the bands for electric conduction, but also pour carriers into trapping centers. When the light excitation is switched off, time is required not only for the free carriers to be recombined in the recombination centers, but also for trapped carriers to be detrapped and then recombined in the recombination centers via thermal excitation and subsequent capture and recombination processes.

7.5.2 Kinetics of Recombination Processes

In Radiative and Nonradiative Transition Process in Chapter 3, we discussed the radiative and nonradiative transitions during recombination processes. Here, we shall confine ourselves to a discussion of the kinetics of recombination processes in general cases only.

Band-to-Band Recombination without Involving Recombination Centers or Traps

An intrinsic semiconductor or insulator may have excess carriers due to thermal excitation or external stimulation (for example, optical or electrical injections). Such excess carriers will disappear through either recombination or carrier flow to the collecting contacts. If there are no collecting contacts, the only way to limit the excess carriers is through recombination. Thus, the rates of change of carrier densities due to optical excitation with a generation rate G can be written as

$$\begin{aligned}\frac{dn}{dt} = \frac{dp}{dt} &= G - R \\ &= G - \langle v\sigma_R \rangle np \\ &= G - C_r np\end{aligned}\quad (7-192)$$

where

$$\begin{aligned}n &= n_o + \Delta n \\ p &= p_o + \Delta p \\ \Delta n &= \Delta p\end{aligned}\quad (7-193)$$

Δn and Δp are the excess carrier densities generated by optical excitation. We shall consider three cases.

Case 1

If $\Delta n \gg n_o$ and $\Delta p \gg p_o$, then the thermally generated carriers may be neglected and $n = p$. Thus, we have

In equilibrium:

$$\frac{dn}{dt} = 0 \quad (7-194)$$

$$(\Delta n)_o = (G/C_r)^{1/2} \quad (7-195)$$

Rate of growth:

$$\frac{dn}{dt} = G - C_r n^2 \quad (7-196)$$

$$\Delta n = (\Delta n)_o \tanh[(GC_r)^{1/2} t] \quad (7-197)$$

Rate of decay:

$$\frac{dn}{dt} = -C_r n^2 \quad (7-198)$$

$$\Delta n = (\Delta n)_o [1 + (\Delta n)_o C_r t]^{-1} \quad (7-199)$$

This case is a typical example of bimolecular recombination. If n_o and p_o are taken into account, the bimolecular recombination rate is not determined by Equations 7-195, 7-197, and 7-199, but by equations given in Cases 2 and 3 following.

Case 2

If $\Delta n \ll n_o$ and $\Delta p \ll p_o$, then the thermally generated carriers are predominant and $n \approx n_o$ and $p \approx p_o \approx n_i^2/n_o$. In this case, we have

In equilibrium:

$$\begin{aligned}\frac{dn}{dt} &= G - C_r (n_o + \Delta n)(p_o + \Delta p) = 0 \\ (\Delta n)_o &= \frac{G - C_r n_i^2}{C_r (n_o + p_o)}\end{aligned}\quad (7-200)$$

Rate of growth:

$$\begin{aligned}\frac{dn}{dt} &= G - C_r (n_o + \Delta n)(p_o + \Delta p) \\ \frac{d(\Delta n)}{dt} &= G - C_r [n_i^2 + \Delta n(n_o + p_o)], \\ \Delta n &= (\Delta n)_o [1 - \exp(-t/\tau)]\end{aligned}\quad (7-201)$$

and

$$\tau = [C_r (n_o + p_o)]^{-1}$$

Rate of decay:

$$\begin{aligned}\frac{d(\Delta n)}{dt} &= -C_r \Delta n (n_o + p_o) \\ \Delta n &= (\Delta n)_o \exp(-t/\tau)\end{aligned}\quad (7-202)$$

Case 3

Both Δn and n_o and Δp and p_o play equally important roles in the recombination processes. In this case, we have

In equilibrium:

$$\begin{aligned}\frac{dn}{dt} &= G - C_r (n_o + \Delta n)(p_o + \Delta p) \\ &= G - C_r [n_o p_o + \Delta n (n_o + p_o) + (\Delta n)^2] \\ &= 0 \\ (\Delta n)_o &= \frac{2(G - C_r n_i^2)}{C_r (n_o + p_o) + A}\end{aligned}\quad (7-203)$$

$$\begin{aligned}\text{or } (\Delta n)_o &= \frac{G - C_r n_i^2}{C_r (n_o + p_o)} \\ \text{if } C_r^2 (n_o + p_o)^2 &\gg 4C_r (G - C_r n_i^2)\end{aligned}$$

$$(\Delta n)_o = \left(\frac{G - C_r n_i^2}{C_r} \right)^{1/2}$$

if $C_r^2(n_o + p_o)^2 \ll 4C_r(G - C_r n_i^2)$.

Rate of growth:

$$\frac{d(\Delta n)}{dt} = (G - C_r n_i^2) - C_r(n_o + p_o)\Delta n - C_r(\Delta n)^2$$

$$\Delta n = (\Delta n)_o \left\{ \frac{C_r(n_o + p_o) + A}{C_r(n_o + p_o) + A \coth(At/2)} \right\} \quad (7-204)$$

Rate of decay:

$$\frac{d(\Delta n)}{dt} = -C_r[n_i^2 + (n_o + p_o)\Delta n + (\Delta n)^2]$$

$$\Delta n = (\Delta n)_o \left\{ \frac{(n_o + p_o) \exp(-t/\tau)}{n_o + p_o + (\Delta n)_o \times [1 - \exp(-t/\tau)]} \right\} \quad (7-205)$$

where

$$A = [C_r^2(n_o + p_o)^2 + 4C_r(G - C_r n_i^2)]^{1/2}$$

$$\tau = [C_r(n_o + p_o)]^{-1}$$

From Equation 7-205, it can be seen that when $(\Delta n)_o < n_o + p_o$, the decay is virtually exponential throughout its course. Even if $(\Delta n)_o > n_o + p_o$, only the initial decay is hyperbolic when $t < \tau$ and it eventually becomes exponential for $t > \tau$. No matter how large $(\Delta n)_o$ may be, hyperbolic decay will rapidly bring (Δn) down to a value less than $(n_o + p_o)$ within the time interval τ .

With a Single Set of Recombination Centers but without Traps

Shockley and Read¹⁰¹ were the first to analyze recombination kinetics in detail for semiconductors with one type of recombination center but without traps, based on the processes shown in Figure 7-22. The center is assumed to be neutral when empty and negatively charged when occupied by an electron. The four processes are

The rate of electron capture by neutral centers of density N_r with capture coefficient C_n is $C_n n(N_r - n_r)$.

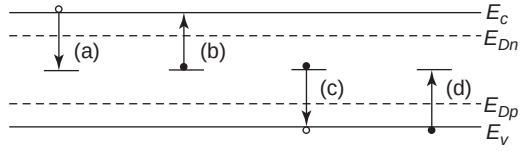


Figure 7-22 Shockley-Read recombination and generation model. The recombination centers are neutral when empty and negatively charged when filled with a captured electron. (a) Capture of an electron from the conduction band by a neutral center, (b) thermal excitation of the captured electron from the center to the conduction band, (c) capture of a hole from the valence band by a filled center (recombination), and (d) thermal excitation of a hole to the valence band from a neutral center (capture of an electron from the valence band by a neutral center, i.e. generation).

The rate of thermal reexcitation of captured electrons from the centers to the conduction band is $C_n n_r N_c \exp[-(E_c - E_r)/kT]$, based on Equations 7-168 and 7-172.

The rate of hole capture by the electron-occupied centers with capture coefficient C_p leading to recombination is $C_p p n_r$.

The rate of thermal excitation of holes from the neutral centers to the valence band (thermal excitation of electrons from the valence band to the unoccupied centers) is $C_p(N_r - n_r)N_v \exp[-(E_r - E_v)/kT]$.

Thus, the net rate of electron capture by recombination centers is

$$\frac{dn_r}{dt} = C_n \{n(N_r - n_r) - n_r N_c \exp[-(E_c - E_r)/kT]\} \quad (7-206)$$

Similarly, the net rate of hole capture by recombination centers is

$$\frac{d(N_r - n_r)}{dt} = C_p \{pn_r - (N_r - n_r)N_v \times \exp[-(E_r - E_v)/kT]\} \quad (7-207)$$

where n_r is the density of electron-occupied centers and E_r is the energy level of the centers. In the steady state, Equations 7-206 and 7-207 must be equal. By denoting f_r as the probability that a recombination center will be occupied, we can write

$$n_r = N_r f_r \quad (7-208)$$

$$(N_r - n_r) = N_r(1 - f_r) \quad (7-209)$$

$$n_i = N_c \exp[-(E_c - E_r)/kT] \quad (7-210)$$

$$p_i = N_v \exp[-(E_r - E_v)/kT] \quad (7-211)$$

Substituting Equations 7-208 through 7-211 into Equations 7-206 through 7-207, setting them equal to each other, and then solving them, we obtain

$$f_r = \frac{C_n n_i + C_p p_i}{C_n(n + n_i) + C_p(p + p_i)} \quad (7-212)$$

Thus, the recombination rate for these acceptor-type recombination centers is

$$R_a = \frac{C_n C_p N_r (np - n_i p_i)}{C_n(n + n_i) + C_p(p + p_i)} \quad (7-213)$$

By introducing

$$\tau_{no} = \frac{1}{C_n N_r} \quad (7-214)$$

$$\tau_{po} = \frac{1}{C_p N_r} \quad (7-215)$$

and since

$$\begin{aligned} n_i p_i &= N_c N_v \exp[-(E_c - E_r)/kT] \\ &= n_o p_o = n_i^2 \end{aligned} \quad (7-216)$$

Equation 7-213 can be expressed as

$$R_a = [\tau_{po}(n + n_i) + \tau_{no}(p + p_i)]^{-1} (np - n_i^2) \quad (7-217)$$

This is the Shockley–Read equation. It can be seen that the driving force for recombination is $np - n_i^2$ which is, in fact, the deviation from the equilibrium condition.

To obtain exact equations for the carrier lifetimes τ_n and τ_p for arbitrary values of excess carrier densities Δn and Δp and recombination center density N_r is difficult, but equations that are useful for most cases with a very good approximation have been derived by Blake-more¹⁰² and they are:

$$\begin{aligned} \tau_n &= \frac{\tau_{po}(n_o + n_i + \Delta n) + \tau_{no}(p_o + p_i + \Delta p) + \tau_{no} N_r \left[\frac{p_i(n_o + n_i + \Delta n) + 2p_o \Delta n}{(p_o + p_i)(n_o + n_i + \Delta n)} \right]}{(n_o + p_o + \Delta n)} \\ &\quad + N_r \left[\frac{p_o(n_o + \Delta n)}{(p_o + p_i)(n_o + n_i + \Delta n)} \right] \end{aligned} \quad (7-218)$$

$$\begin{aligned} \tau_p &= \frac{\tau_{po}(n_o + n_i + \Delta n) + \tau_{no}(p_o + p_i + \Delta p) + \tau_{po} N_r \left[\frac{p_o(p_o + p_i + \Delta p) + 2p_i \Delta p}{(p_o + p_i)(p_o + p_i + \Delta p)} \right]}{(n_o + p_o + \Delta p)} \\ &\quad + N_r \left[\frac{p_i(p_o + \Delta p)}{(p_o + p_i)(p_o + p_i + \Delta p)} \right] \end{aligned} \quad (7-219)$$

Case 1

Δn and Δp are small compared to n_o and p_o ($n = n_o + \Delta n$, $p = p_o + \Delta p$), but $\tau_n \neq \tau_p$. In this case, we can obtain τ_n and τ_p by setting $\Delta n \rightarrow 0$ and $\Delta p \rightarrow 0$ in Equations 7-218 and 7-219. Thus, we have

$$\tau_n = \frac{\tau_{po}(n_o + n_i) + \tau_{no}[p_o + p_i + N_r p_i/(p_o + p_i)]}{n_o + p_o + N_r[n_o p_o/(n_o + n_i)(p_o + p_i)]} \quad (7-220)$$

$$\tau_p = \frac{\tau_{no}(p_o + p_i) + \tau_{po}[n_o + n_i + N_r p_o/(p_o + p_i)]}{n_o + p_o + N_r[p_o p_i/(p_o + p_i)^2]} \quad (7-221)$$

and

$$R_a = \frac{n_o \Delta p + p_o \Delta n}{\tau_{po}(n_o + n_i) + \tau_{no}(p_o + p_i)} \quad (7-222)$$

Case 2

Δn and Δp are small compared to n_o and p_o , but $\Delta n = \Delta p$ and $\tau_n = \tau_p$. In this case, we can obtain $\tau_n = \tau_p$ by setting the terms in N_r to zero in Equation 7-220 or 7-221. Thus, we have

$$\tau_n = \tau_p = \tau_{po}(n_o + n_i)/(n_o + p_o) + \tau_{no}(p_o + p_i)/(n_o + p_o) \quad (7-223)$$

and

$$R_a = \frac{\Delta n(n_o + p_o)}{\tau_{po}(n_o + n_i) + \tau_{no}(p_o + p_i)} \quad (7-224)$$

Case 3

Δn and Δp are much larger than n_o or p_o , but still $\Delta n = \Delta p$ and $\tau_n = \tau_p$. In this case, we can obtain $\tau_n = \tau_p$ by setting the terms in N_r to zero in Equation 7-218 or 7-219. Thus, we have

$$\begin{aligned} \tau_n = \tau_p &= \tau_{po}(n_o + n_i + \Delta n)/(n_o + p_o + \Delta n) \\ &\quad + \tau_{no}(p_o + p_i + \Delta n)/(n_o + p_o + \Delta n) \end{aligned} \quad (7-225)$$

and

$$R_a = \frac{\Delta n(n_o + p_o + \Delta n)}{\tau_{po}(n_o + n_i + \Delta n) + \tau_{no}(p_o + p_i + \Delta n)} \quad (7-226)$$

Letting τ' denote the lifetime for vanishingly small values of Δn as given by Equation 7-223, then Equation 7-225 can be written in the form

$$\tau = \tau' \frac{1 + a\Delta n}{1 + b\Delta n} \quad (7-227)$$

where

$$a = \frac{\tau_{po} + \tau_{no}}{\tau_{po}(n_o + n_i) + \tau_{po}(p_o + p_i)} \quad (7-228)$$

$$b = (n_o + p_o)^{-1} \quad (7-229)$$

If $a > b$, then τ increases monotonically with increasing Δn , giving rise to superlinear photoconductivity; if $a < b$, then τ decreases monotonically with increasing Δn , giving rise to sublinear photoconductivity. The limiting value for τ as Δn approaches infinity is¹⁰¹

$$\tau_\infty = \tau_{po} + \tau_{no} \quad (7-230)$$

In fact, Case 3 is close to most cases for insulators or organic semiconductors in which the energy band gap is large and the mobility is small. Supposing that $\Delta n = \Delta p$, $\Delta n > n_o$, and $\Delta n > p_o$, then we can write

$$\begin{aligned} \frac{d(\Delta n)}{dt} &= G - R \\ &= G - \frac{\Delta n}{\tau} \end{aligned} \quad (7-231)$$

In the steady state

$$G = \frac{(\Delta n)_o}{\tau'} \left[\frac{1 + b(\Delta n)_o}{1 + a(\Delta n)_o} \right] \quad (7-232)$$

After Δn has reached its steady-state value $(\Delta n)_o$, if the optical excitation is switched off, Δn will decay following the expression

$$\Delta n = (\Delta n)_o \exp(-t/\tau) \quad (7-233)$$

where τ is given by Equation 7-227. For this case, the response time is approximately equal to the lifetime.

By comparing Equation 7-233 with Equation 7-199, it is clear that for $\Delta n = \Delta p \gg n_o$ and

$\gg p_o$, the excess carrier density or photocurrent decays exponentially with time if the recombination is monomolecular (indirect recombination through recombination centers), and decays hyperbolically with t for $t \ll \tau$ if the recombination is bimolecular (direct band-to-band recombination). This is generally used as a criterion to distinguish experimentally these two types of recombination processes.

7.5.3 Space-Charge Limited Electrical Conduction: Two-Carrier (Double) Planar Injection

In double injection, charge carriers of both types (electrons and holes) are present, and the problem becomes much more complicated because the current-voltage (J - V) characteristics are controlled by recombination, which may be either direct band-to-band recombination or indirect recombination, occurring through one or more sets of localized states. This section deals only with the steady-state DC one-dimensional planar current flow for two general cases, making the following assumptions:

- The energy band model can be used to treat the behavior of injected carriers.
- The anode for injecting holes and the cathode for injecting electrons are both perfectly ohmic injecting contacts located at $z = 0$ and $z = d$, respectively, the specimen thickness being d .
- The electric field is so large that the current components due to diffusion and to carriers generated thermally in the specimen can be neglected.
- The free hole and electron densities and trapped hole and electron densities in shallow traps follow the Maxwell-Boltzmann statistics; the trapped hole and electron densities in deep traps follow the Fermi-Dirac statistics.
- The mobilities of the free holes and electrons are independent of field and are not affected by the presence of traps and recombination centers.
- The planes perpendicular to the z -axis, at which the field is zero, are located at $z = w_a$

and $z = w_c$ which are, respectively, very close to the injecting contacts at $z = o$ and $z = d$, so that

$$F(z = w_o \approx o) = F(z = w_c \approx d) = 0 \quad (7-234)$$

The behavior of double injection in a solid is governed by the current flow equations

$$J_n = q\mu_n nF \quad (7-235)$$

$$J_p = q\mu_p pF \quad (7-236)$$

$$J = J_n + J_p \quad (7-237)$$

the continuity equations

$$\frac{1}{q} \frac{dJ_n}{dz} = R \quad (7-238)$$

$$-\frac{1}{q} \frac{dJ_p}{dz} = R \quad (7-239)$$

and Poisson's equation

$$\frac{dF}{dz} = \frac{q}{\epsilon} [p(z) + p_t - n(z) - n_t] = \frac{\rho}{\epsilon} \quad (7-240)$$

where n and p are given by

$$n = N_c \exp[-(E_c - E_{Fn})/kT] \quad (7-241)$$

$$p = N_v \exp[-(E_{Fp} - E_v)/kT] \quad (7-242)$$

and n_t and p_t are given by

$$n_t = \int_{E_t}^{E_u} h_n(E, z) \{1 + g_n^{-1} \times \exp[(E - E_{Fn})/kT]\}^{-1} dE \quad (7-243)$$

$$p_t = \int_{E_t}^{E_u} h_p(E, z) \{1 + g_p \times \exp[(E_{Fp} - E)/kT]\}^{-1} dE \quad (7-244)$$

in which $h_n(E, z)$ and $h_p(E, z)$ are the distribution function for electron and hole trap densities, respectively.

We shall consider two general cases and state any additional assumptions that must be made when dealing with individual cases.

Without Recombination Centers and without Traps

For this case, $n_t = p_t = 0$, $R = n_p \langle v\sigma_R \rangle$. To simplify the mathematical treatment, we introduce the following parameters

$$\left. \begin{aligned} S_o &= q\mu_n nF/J \\ T_o &= q\mu_p pF/J \\ U_o &= qF^2/J \\ \alpha &= 2q/\epsilon \\ \beta_o &= \langle v\sigma_R \rangle \end{aligned} \right\} \quad (7-245)$$

Using these parameters, Equations 7-237 through 7-240 can be written as

$$S_o + T_o = 1 \quad (7-246)$$

$$\frac{dS_o}{dz} = \beta_o S_o T_o / \mu_n \mu_p U_o \quad (7-247)$$

$$\frac{dT_o}{dz} = -\beta_o S_o T_o / \mu_n \mu_p U_o \quad (7-248)$$

$$\frac{dU_o}{dz} = \alpha [T_o / \mu_p - S_o / \mu_n] \quad (7-249)$$

Solving these equations gives

$$U_o = C_o S_o^{\alpha\mu_n/\beta_o} (1 - S_o)^{\alpha\mu_p/\beta_o} \quad (7-250)$$

where C_o is the integration constant. The entire current at the anode is carried by holes, so $S_o = 0$, and the entire current at the cathode is carried by electrons, so $S_o = 1$. Substituting Equation 7-250 into Equation 7-247 and then integrating it, we obtain

$$C_o = \beta_o d \left[\mu_n \mu_p \int_0^1 S_o^{\alpha\mu_n/\beta_o-1} (1 - S_o)^{\alpha\mu_p/\beta_o-1} dS_o \right]^{-1} \quad (7-251)$$

Using the boundary condition

$$V = \int_0^d F dz \quad (7-252)$$

and from Equations 7-245 and 7-251, it can easily be shown that the relation between J and V is¹⁰³

$$J = \frac{9}{8} \epsilon \mu_{\text{eff}} \frac{V^2}{d^3} \quad (7-253)$$

where

$$\mu_{\text{eff}} = \frac{8}{9} \frac{q}{\epsilon} \frac{\mu_n \mu_p}{\langle v\sigma_R \rangle} \times \frac{\left[\int_0^1 S_o^{\alpha\mu_n/\beta_o-1} (1 - S_o)^{\alpha\mu_p/\beta_o-1} dS_o \right]^3}{\left[\int_0^1 S_o^{3\alpha\mu_n/2\beta_o-1} (1 - S_o)^{3\alpha\mu_p/2\beta_o-1} dS_o \right]^2}$$

$$\begin{aligned}
&= \frac{8 q \mu_n \mu_p}{9 \varepsilon \langle v \sigma_R \rangle} \left[\frac{B\left(\frac{\alpha \mu_n}{\beta_o}, \frac{\alpha \mu_p}{\beta_o}\right)}{B\left(\frac{3\alpha \mu_n}{2\beta_o}, \frac{3\alpha \mu_p}{2\beta_o}\right)} \right]^3 \\
&= \frac{8 q \mu_n \mu_p}{9 \varepsilon \langle v \sigma_R \rangle} \left\{ \frac{\left(\frac{\alpha \mu_n}{\beta_o} - 1\right)! \left(\frac{\alpha \mu_p}{\beta_o} - 1\right)!}{\left[\frac{\alpha}{\beta_o} (\mu_n + \mu_p) - 1\right]!} \right\}^3 \\
&\quad \times \left\{ \frac{\left[\frac{3\alpha}{2\beta_o} (\mu_n + \mu_p) - 1\right]!}{\left(\frac{3\alpha \mu_n}{2\beta_o} - 1\right)! \left(\frac{3\alpha \mu_p}{2\beta_o} - 1\right)!} \right\}^2 \quad (7-254)
\end{aligned}$$

in which $B(m, n)$ is the Beta function. This is the general equation for solids without recombination centers and without traps. This equation for μ_{eff} can be simplified for some simple cases, as shown in the following sections.

Injected Plasma

In this case, $\langle v \sigma_R \rangle$ is small, so $\frac{\alpha \mu_n}{\beta_o} \gg 1$ and $\frac{\alpha \mu_p}{\beta_o} \gg 1$. Thus, with the aid of Stirling's formula

$$(m-1)! \approx m! \approx (m/e)^m (2\pi m)^{1/2} \quad (7-255)$$

Equation 7-254 can be approximated to

$$\mu_{\text{eff}} = \frac{2}{3} \left[\frac{4\pi q \mu_n \mu_p (\mu_n + \mu_p)}{\varepsilon \langle v \sigma_R \rangle} \right]^{1/2} \quad (7-256)$$

If $\langle v \sigma_R \rangle$ is small, the recombination cross-section σ_R is small. This implies that recombination is not a barrier that hinders both electrons and holes from completing their penetration of the material specimen, and that the charge neutrality condition prevails throughout the bulk. This can be seen in Figure 7-23(a), in which $n \approx p$ throughout the bulk.

Space-Charge Limited Currents

In this case, $\langle v \sigma_R \rangle$ is very large, so $\alpha \mu_n / \beta_o \ll 1$ and $\alpha \mu_p / \beta_o \ll 1$. Using the relation

$$(m-1)! \approx 1/m \quad \text{for } 0 < m \ll 1 \quad (7-257)$$

Equation 7-254 can be approximated to

$$\mu_{\text{eff}} = \mu_n + \mu_p \quad (7-258)$$

This equation and Equation 7-253 indicate that the total current is simply the sum of the two separate one-carrier SCL currents.

When $\langle v \sigma_R \rangle$ is very large and approaches infinity, σ_R also approaches infinity. This implies that there is no region in which electrons and holes can overlap, since there would be an infinite recombination current if such an overlap region occurred. Under such a condition, the electron current would exist only on

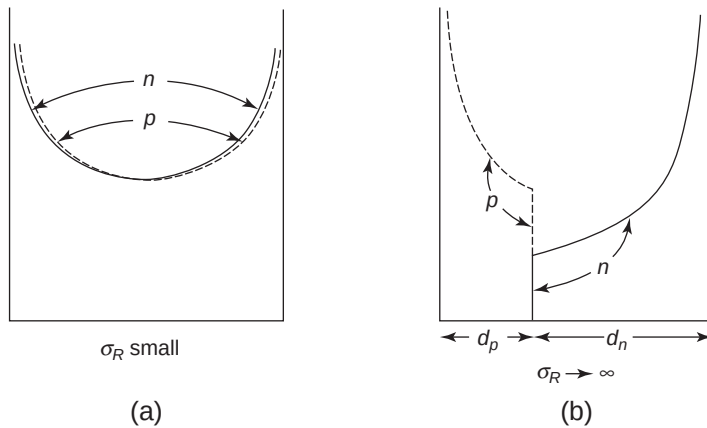


Figure 7-23 Spatial distributions of the electron and hole densities for double injection into a trap-free dielectric specimen for two limiting cases: (a) the small σ_R limit corresponding to an injected plasma and (b) the $\sigma_R \rightarrow \infty$ limit corresponding to back-to-back single carrier SCL currents. The particular case shown corresponds to the choice of $\mu_n/\mu_p = 2$.

the cathode side and the hole current on the anode side, and they would meet and annihilate at a certain plane dividing these two regions, as shown in Figure 7-23(b).

One-Carrier Space-Charge Limited Currents

In this case, $\alpha\mu_n/\beta_o \gg 1$, but $\alpha\mu_p/\beta_o \ll 1$. Using the same technique for simplifying Equation 7-254, we obtain

$$\mu_{\text{eff}} \approx \mu_n \quad (7-259)$$

Thus, the total current will be mainly electron SCL current, the hole current being negligible because the holes are so sluggish that most of them will be annihilated by recombination at the anode.

With Recombination Centers but without Traps

Several investigators have used a model based on a perfectly compensated semiconductor for the analysis of double injection with recombination centers.^{54,103-108} In this model, the deep acceptorlike traps act as recombination centers and the shallow donors provide compensating electrons for these deep acceptors to preserve local electrical neutrality, as shown in Figure 7-24. A recombination center is nega-

tively charged when it is occupied by a compensating electron, but this negative charge is balanced by the positive charge of the ionized shallow donor, maintaining local neutrality in thermal equilibrium. Similarly, the existence of a neutral, unoccupied recombination center implies the existence of an un-ionized shallow donor in thermal equilibrium. Double injection will disturb this thermal-equilibrium local neutrality condition.

The shallow donors can be assumed to provide only compensating electrons in thermal equilibrium and to play no role in the recombination and conduction processes. The recombination centers play the major role in double injection. Under such a condition, the capture cross-section for capturing holes is greater than the capture cross-section for capturing electrons in the recombination centers. This implies that the electron lifetime is greater than the hole lifetime. In this section, we shall use this model to describe and to explain some important features of double injection.

In the steady state, the behavior of double injection in a solid is governed by the current flow equations

$$J_n = q\mu_n nF + qD_n \frac{dn}{dz} \quad (7-260)$$

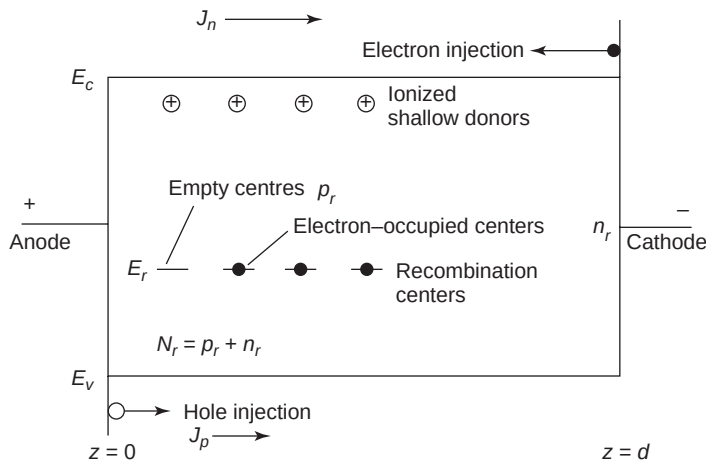


Figure 7-24 A single set of recombination centers in a perfectly compensated semiconductor. N_r denotes the density of total recombination centers; n_r and p_r denote, respectively, the densities of electron-occupied and empty recombination centers.

$$J_p = q\mu_p pF - qD_p \frac{dp}{dz} \quad (7-261)$$

$$J = J_n + J_p \quad (7-262)$$

the continuity equations

$$\frac{1}{q} \frac{dJ_n}{dz} = R \quad (7-263)$$

$$-\frac{1}{q} \frac{dJ_p}{dz} = R \quad (7-264)$$

Poisson's equation

$$\frac{dF}{dz} = \frac{q}{\epsilon} [(p - p_o) - (n - n_o) + (p_r - p_{ro})] \quad (7-265)$$

and

$$V = \int_0^d F dz \quad (7-266)$$

These equations include the current components due to diffusion and to carriers generated thermally in the specimen. In thermal equilibrium, $n = n_o$, $p = p_o$, $p_r = p_{ro}$ and the whole specimen is electrically neutral (there is no net space charge). Double injection changes these thermal equilibrium quantities by an amount of

$$\begin{aligned} \Delta n &= n - n_o \\ \Delta p &= p - p_o \\ \Delta p_r &= p_r - p_{ro} \end{aligned} \quad (7-267)$$

Thus, Δn and Δp are densities of injected free electron and hole carriers, respectively, and Δp_r is the density of injected trapped holes. If n_o , p_o , and p_{ro} , which are thermally generated, are ignored, then automatically $\Delta n = n$, $\Delta p = p$, and $\Delta p_r = p_r$.

Equations 7-260 through 7-266 are completely general and apply to all electrical transport problems in semiconductors and insulators. If monomolecular recombination is predominant, and if bimolecular recombination can be ignored, then the rate of recombination may be written based on Equation 7-206 or 7-207 as

$$\begin{aligned} R &= C_n[(n_o + \Delta n)(p_{ro} + \Delta p_r) - (n_{ro} - \Delta p_r)n_i] \\ &= C_n[\Delta n p_{ro} + \Delta p_r(\Delta n + n_o + n_i)] \\ &= C_p[(p_o + \Delta p)(n_{ro} - \Delta p_r) - (p_{ro} + \Delta p_r)p_i] \\ &= C_p[\Delta p n_{ro} - \Delta p_r(\Delta p + p_o + p_i)] \end{aligned} \quad (7-268)$$

Analytical solutions of Equations 7-260 through 7-266 are not possible without involving restrictive assumptions. However, analyses can be many, depending on the assumptions made for certain conditions in order to simplify the problem. The carrier lifetimes τ_n and τ_p depend on the injection level and hence on p_r , and the changes to these lifetimes are responsible for various features of the J - V characteristics. The following analysis shows the solution for one case with a large density of deep recombination centers.

To begin, we will make the following assumptions:

- The current is volume controlled and no constraints on the current are imposed by the carrier-injecting contacts.
- The structure is a $p^+ - i - n^+$ long structure with the i -region several ambipolar diffusion lengths long, so the diffusion current component may be neglected.
- The carrier mobilities are field independent.
- The recombination centers are acceptorlike deep traps.

Based on these assumptions, the behavior of the double-injection current can be described by the following equations:

$$J = q(\mu_n n + \mu_p p)F \quad (7-269)$$

$$\mu_n \frac{d(nF)}{dz} = R \quad (7-270)$$

$$\frac{dF}{dz} = \frac{q}{\epsilon} [(p - p_o) - (n - n_o) + (p_r - p_{ro})] \quad (7-271)$$

where

$$R = \frac{N_r C_n C_p (pn - n_i^2)}{C_n(n + n_i) + C_p(p + p_i)} \quad (7-272)$$

$$N_r - P_r = \frac{N_r (C_n n_i + C_p p_i)}{C_n(n + n_i) + C_p(p + p_i)} \quad (7-273)$$

An analytical and exact solution of Equations 7-269 through 7-273 is not possible. However, Deuling¹⁰⁹ has obtained an exact numerical solution using a Rung-Kutta method with a computer for Au-doped Si, using the following physical parameters:

$$N_r \text{ (Au impurities in Si)} = 10^{16} \text{ cm}^{-3}.$$

$$E_F - E_v = 0.485 \text{ eV}, E_r - E_v = 0.620 \text{ eV}.$$

$$C_n = 1.65 \times 10^{-9} \text{ cm}^3 \text{ sec}^{-1}.$$

$$C_p = 1.15 \times 10^{-7} \text{ cm}^3 \text{ sec}^{-1}.$$

The voltage V is normalized to $V_o = d^2 (C_n N_r / \mu_n)$, and the current J is normalized to $J_o = \epsilon \mu_n d (C_n N_r / \mu_n)^2$.

Figure 7-25 shows Deuling's computed results. For the same problem, Lampert^{54,104,105} has used a regional approximation method based on the quasi-neutral approximation and obtained a solution. His result is also shown in Figure 7-25 for purposes of comparison. The quasi-neutrality approximation gives a general shape of J - V characteristics, although it overemphasizes the negative differential resistance region. Deuling¹⁰⁹ has also calculated the J - V characteristics as functions of the temperature, and his results are shown in Figure 7-26. The lower the temperature is, the longer the negative differ-

ential resistance region. This agrees with the experimental results of Bykovskii et al.¹¹⁰

Note that when the injection level becomes so high in the negative differential resistance region, a transition will occur to convert the negative differential resistance back to the positive differential resistance, but at higher current levels. After such a transition, $J \propto V^2$ in the semiconductor regime and $J \propto V^3$ in the insulator regime.

For details about the regional approximation method based on the quasi-neutrality approximation, see the review of Lampert and Schilling¹⁰⁵ and Lampert and Mark.⁵⁴ For details about other approximation methods to deal with double injection problems, see the review of Baron and Mayer¹⁰⁷ and Migliorato et al.¹⁰⁸

Before closing this section, it is worth mentioning that the threshold voltage for the onset and the length of the negative differential resistance (NDR) region both depend on the

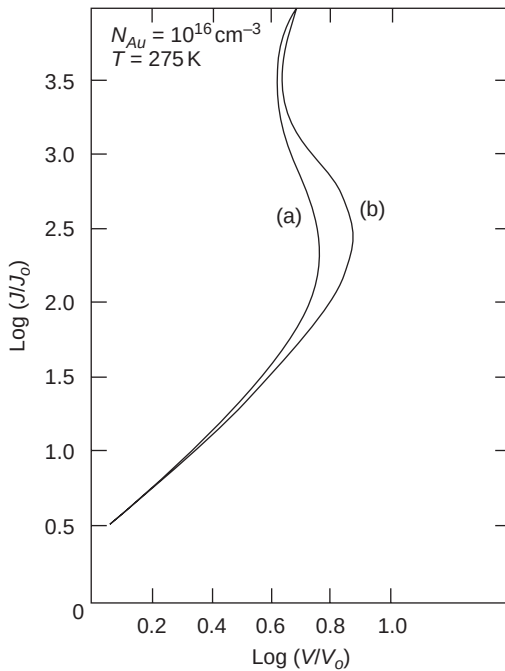


Figure 7-25 The J - V characteristics of gold-doped silicon computed by Deuling¹⁰⁹ based on (a) the exact numerical solution and (b) the quasi-neutrality regional approximation.

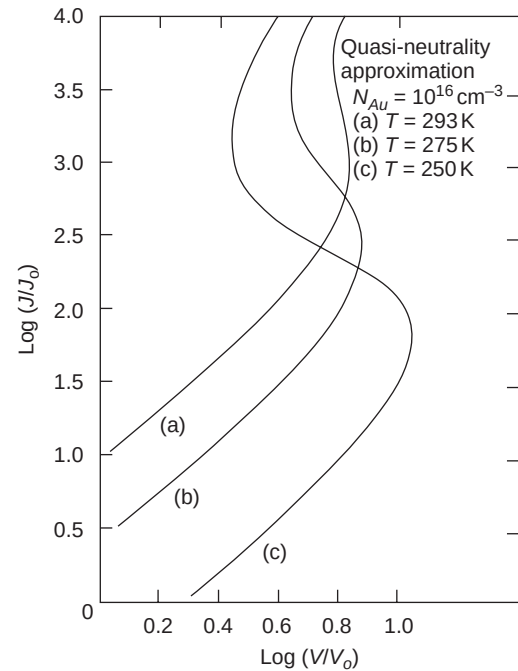


Figure 7-26 The negative differential resistance (NDR) regions of the J - V characteristics of gold-doped silicon as functions of temperature calculated on the basis of the quasi-neutrality approximation by Deuling.¹⁰⁹

externally applied magnetic field or intense light illumination. Generally, the threshold voltage and the length of the NDR region decrease with increasing magnetic field, which has been theoretically and experimentally studied.^{111–113} Several investigators have also reported that in $p^+ - i - n^+$ GaAs devices the threshold voltage and the length of the NDR region decrease with increasing intensity of the illuminating light.^{114,115} It can be imagined that if a current-controlled NDR device were connected to a constant DC bias voltage in series with a load resistance, it could be made to switch between a low-conductivity regime and a high-conductivity regime. This can be accomplished by the application of a pulse of electric field whose direction either aids or opposes the field produced by the DC bias voltage, depending on whether turn-on (to the high-conductivity regime) or turn-off (to the low-conductivity regime) is required. The switching may take place with a time between 10^{-6} and 10^{-9} sec, depending on load resistance. Such a switching function also may be achieved by the application of a pulse of magnetic field or intense light. However, two-terminal NDR switching is not convenient to use in practice, which may be why this switching feature has not yet been further developed for practical applications.

7.6 High-Field Effects

When the applied electric field becomes high, several effects may arise. The effects on the J – V characteristics are mainly caused either by a change in the distribution function or the mobility of charge carriers, or by a change in the injection or generation of charge carriers, or by both. High-field effects have been studied extensively both theoretically and experimentally by many investigators.^{2,116–121} For high-resistivity (or low mobility) solids such as inorganic and organic insulating materials, the high-field effects can be divided into three categories:

Electrode effects: The high-field effects are caused by the field-dependent rate of carrier injection or emission from electrodes

through either Schottky-type thermionic emission or tunneling through a potential barrier near the injecting contact or across a thin insulating film.

Bulk effects: The high-field effects are caused by the field-dependent carrier mobility due to various scatterings; by the field-dependent carrier density due to field-dependent trapping probability, the detrapping process (e.g., the Poole–Frenkel effect), and tunneling of trapped carriers; and by the field-dependent thermal effect arising from the joule heating of the specimen.

Combined electrode and bulk effects: The high-field effects may be changed from bulk-limited to electrode-limited or from electrode-limited to bulk-limited processes when the applied field is increased.

7.6.1 Filamentary Charge-Carrier Injection in Solids

It is well known that for either single or double injections, different materials used for electrodes or different techniques employed for preparation of the surfaces of semiconductors or insulators result in different values of critical voltages for transition from regime to regime. This indicates that there are no perfect ohmic contacts and that different electrode materials in contact with a crystal surface will form different potential barriers for carrier injection. It can also be imagined that the interface between an electrode and a crystal surface, which is not microscopically identical from domain to domain in asperity and surface conditions, is never homogeneous and uniform. Thus, there may be one or more microregions at which the potential barrier has a profile more favorable to carrier injection than at other regions of the interface. Furthermore, the crystal itself is never microscopically homogeneous or uniform.

Due to all these unavoidable imperfections, the current passing through a crystal specimen is filamentary, at least from a microscopic point of view and particularly under high fields. If an electric field is applied to the specimen longitudinally, the field will not be uniform

longitudinally (due to the effect of space charge) and the current density will not be uniform radially (due to the formation of filamentary paths). The current filaments formed in Si, GaAs, ZnTe, GeAs_xP_{1-x}, and polycrystalline Si have been observed by Barnett et al.,¹²²⁻¹²⁵ and in anthracene and in chalcogenide glasses observed by Kao et al.^{82,126,127}

The following sections present a theoretical model for filamentary injection and show that the model can be used to explain some experimental aspects of single and double injection.

Filamentary One-Carrier (Single) Injection

In our theoretical analysis, we will consider only electron injection and make the following assumptions:

- When applied voltages equal to or higher than the threshold voltage for the onset of carrier injection, there may be one or more filaments formed between electrodes. But, for mathematical simplicity, we will use cylindrical coordinates and consider only one filament formed along the z -axis, which coincides with the central line joining the two circular plane electrodes. The filament radius is r_0 . The whole system is symmetrical about the z -axis.
- In the filament, the longitudinal component of the diffusion current may be ignored because of the large longitudinal component of the electric field. The radial component of the drift current may be ignored because of the small radial component of the electrical field.
- Free electron density follows the Maxwell-Boltzmann statistics; trapped electron density follows the Fermi-Dirac statistics.
- The mobility of the free electrons μ_n is not affected by the presence of traps or by the high electric field.
- The treatment is two-dimensional with the plane at $z = 0$ as the electron-injecting contact and at $z = d$ as the electron-collecting contact, the specimen thickness being d .
- The thermally generated electron concentration is negligible.

The behavior of single injection in the steady state is governed by the current flow equation

$$\begin{aligned}\vec{J}(r) &= \vec{J}_z + \vec{J}_r \\ &= q\mu_n n F \vec{i}_z - qD_n \left(\frac{\partial n}{\partial r} \right) \vec{i}_r\end{aligned}\quad (7-274)$$

the continuity equation

$$\mu_n \left(\frac{\partial}{\partial z} \right) (nF) \approx 0 \quad (7-275)$$

$$\mu_n \frac{\partial (nF)}{\partial r} - \frac{D_n}{r} \frac{\partial}{\partial r} \left(r \frac{\partial n}{\partial r} \right) = 0 \quad (7-276)$$

and Poisson's equation

$$\nabla \cdot F = \frac{q}{\epsilon} (n + n_t) \quad (7-277)$$

where $\vec{i}_z$ and $\vec{i}_r$ are, respectively, the unit vectors in the z and r directions; the other symbols have the usual meanings.

The field F in the radial direction is assumed to be negligibly small, but dF/dr is not, and it should follow Poisson's equation. Thus, Equation 7-276 becomes

$$\frac{d^2 n}{dr^2} + \frac{1}{r} \frac{dn}{dr} = \frac{q\mu_n}{\epsilon D_n} n(n + n_t) \quad (7-278)$$

For trap-free solids, $n_t = 0$, and for solids with shallow traps we can write

$$\theta_n = \frac{n}{n + n_t} \quad (7-279)$$

An examination of Equation 7-278 shows that when r approaches zero, n approaches infinity. Physical reality requires n to be finite for all values of r , and this demands that $dn/dr \rightarrow 0$ when $r \rightarrow 0$. However, Equation 7-279 cannot be solved rigorously without the aid of numerical computation. To emphasize the physical picture of the problem, a simple analytical solution is always better than a computer solution. Therefore, we must make an approximation. It is postulated that by neglecting the term $\frac{1}{r} \frac{dn}{dr}$, the solution of Equation 7-279 is a good approximation of the exact solution of Equation 7-279.¹²³ On the basis of this approximation, and on the assumption that the traps are shallow traps, Equation 7-278 reduces to

$$\frac{d^2 n}{dr^2} = \frac{q\mu_n}{\varepsilon D_n} \theta_n^{-1} n^2 \quad (7-280)$$

Using the boundary conditions

$$\begin{aligned} r \rightarrow 0; \quad n \rightarrow n_o \\ r \rightarrow \infty \quad n \rightarrow 0 \quad \text{and} \quad \frac{dn}{dr} \rightarrow 0 \end{aligned} \quad (7-281)$$

the solution of Equation 7-280 gives

$$n = \frac{n_o}{\left[1 + \frac{n_o}{6} \frac{q\mu_n r}{D_n \varepsilon \theta_n}\right]^2} \quad (7-282)$$

where n_o is the electron density at the center of the filament, which is related to the current density at the center of the filament J_{zo} . Thus, we can write

$$n_o = J_{zo}/qu_n F \quad \text{or} \quad J_{zo} = qu_n n_o F \quad (7-283)$$

Using the boundary condition

$$F \rightarrow F_c \quad \text{when} \quad z \rightarrow 0 \quad (7-284)$$

and

$$V = \int_0^d F dz = F_{av} d \quad (7-285)$$

the solution of Equations 7-277 and 7-283 at $r = 0$ for $n_i \gg n$ yields¹²⁶

$$J_{zo} = \frac{9}{8} \varepsilon \mu_n \theta_n \frac{F_{av}^2}{d} M = \frac{9}{8} \varepsilon \mu_n \theta_n \frac{V^2}{d^3} M \quad (7-286)$$

where

$$\begin{aligned} M = & -\frac{4Y^2 - 3}{6} \\ & + \left[\left(-\frac{4Y^2 - 3}{6} \right)^2 - \frac{16Y^3(Y-1)}{27} \right]^{1/2} \end{aligned} \quad (7-287)$$

$$Y = F_c / F_{av} \quad (7-288)$$

V is the applied voltage, F_{av} is the average value of F along the z direction, and F_c is the electric field at the cathode ($z = 0$).

Equation 7-283 can also be expressed by approximation as

$$J_{zo} = q\mu_n \bar{n}_o F_{av} \quad (7-289)$$

where $\bar{n}_o$ is the average carrier density at the center of the filament (at $r = 0$). Comparing Equation 7-289 to 7-286, we obtain

$$\bar{n}_o = \frac{9}{8} \frac{\varepsilon \theta_n F_{av}}{qd} M \quad (7-290)$$

Using $\bar{n}_o$ for n in Equation 7-282, we obtain the average electron density as a function of r as

$$\bar{n} = \frac{9}{8} \frac{\varepsilon \theta_n F_{av}}{qd} M \left[1 + \left(\frac{3M\mu_n F_{av}}{16D_n d} \right)^{1/2} r \right]^{-2} \quad (7-291)$$

Thus, the average current density is

$$\begin{aligned} J_z(r) &= q\mu_n \bar{n} F_{av} \\ &= J_{zo} \left[1 + \left(\frac{3M}{16} \frac{\mu_n V}{D_n} \right)^{1/2} \frac{r}{d} \right]^{-2} \\ &= J_{zo} \left[1 + \left(\frac{J_{zo}}{6} \frac{1}{\theta_n D_n \varepsilon} \frac{d}{V} \right)^{1/2} r \right]^{-2} \end{aligned} \quad (7-292)$$

and the total current is

$$\begin{aligned} I &= \int_0^{2\pi} \int_0^{rd} J_z(r) r dr d\theta \\ &= 2\pi J_{zo} \int_0^{rd} \frac{r dr}{\left[1 + \left(\frac{J_{zo}}{6} \frac{d}{\theta_n D_n \varepsilon V} \right)^{1/2} r \right]^2} \\ &= 12\pi \theta_n D_n \varepsilon \frac{V}{d} \left\{ \ell n \left[1 + \left(\frac{J_{zo}}{6} \frac{d}{\theta_n D_n \varepsilon V} \right)^{1/2} r_d \right] \right. \\ &\quad \left. + \left[1 + \left(\frac{J_{zo}}{6} \frac{d}{\theta_n D_n \varepsilon V} \right)^{1/2} r_d \right]^{-1} - 1 \right\} \end{aligned} \quad (7-293)$$

Note that multiple current filaments may exist simultaneously between two parallel electrodes. In such a case, the total current between the plane electrodes may be expressed as

$$I_T = I_{\text{domain1}} + I_{\text{domain2}} + \dots = \sum_n I_n \approx HI \quad (7-294)$$

This means that the total current can be represented by the current in one filament I multiplied by a parameter H , which may be field dependent.

By assuming that the electron injection is due mainly to tunneling through a potential barrier at the cathode following Equation 6-97 (field emission) (see Chapter 6). We can write

$$J_c = aF_c^2 \exp(-b/F_c) \quad (7-295)$$

and making $J_c = J_{zo}$ since the current in the z direction is continuous, then from Equations 7-256 through 7-288 and 7-295, we obtain

$$Y = \frac{8}{27[Qe^{-s} + 2/3]^2 + 4} + \left\{ \left(\frac{8}{27[Qe^{-s} + 2/3]^2 + 4} \right)^2 + \frac{27Qe^{-s}}{27[Qe^{-s} + 2/3]^2 + 4} \right\}^{1/2} \quad (7-296)$$

where

$$Q = 8ad/q\epsilon\mu_n \quad \text{and} \quad S = b/YF_{av} \quad (7-297)$$

a and b are constants depending on the interfacial condition between the cathode and the material specimen, and J_c and F_c are, respectively, the current density and the field strength at the cathode. For a given material specimen and a given carrier-injecting contact (cathode electrode) Y decreases (or M increases) with increasing d and F_{av} , implying that the ratio F_c/F_{av} or the space charge effect depends on both specimen thickness and applied voltage.¹²⁶ This model has been used to explain semi-quantitatively high-field electric conduction and breakdown phenomena in dielectric liquids.¹²⁶ Of course, this model may also be used for high-field electric conduction in solids involving a field-emission contact.

Filamentary Two-Carrier (Double) Injection

Using a method similar to that in the previous section, we can easily derive the expressions for $J_z(r) - V$ and the $I-V$ characteristics for the case of filamentary two-carrier (double) injection. Because of the limited space, we shall not include the mathematical analysis here. The reader interested in that analysis is referred to references.¹²²⁻¹²⁵

The formation of the NDR region resulting from double injection in a solid with recombination centers, discussed in Section 7.5.3, has been attributed to the development of a current filament between two carrier-injecting elec-

trodes.⁹⁴ The $I-V$ characteristics observed in double injection are of the current-controlled form, as shown in Figures 7-25 through 7-27. Usually, a stable filament is formed after the transition from a low-current regime to a high-current regime, and sometimes after the occurrence of a current-controlled NDR region. Several investigators have reported that, prior to switching the current to the OFF state (low-current regime), the current is dependent on the electrode area, but after switching to the ON state (high-current regime), the current becomes practically independent of the electrode area. This indicates that electrical conduction takes place along a filament path between two electrodes.^{127,128} This is true for all materials with a switching phenomenon.

Using recombination radiation photography, Barnett and Milnes¹²⁵ have observed the development and growth of current filaments in indium-doped semi-insulating silicon $p^+ - \pi - n^+$ structures. Using a step-by-step sectioning

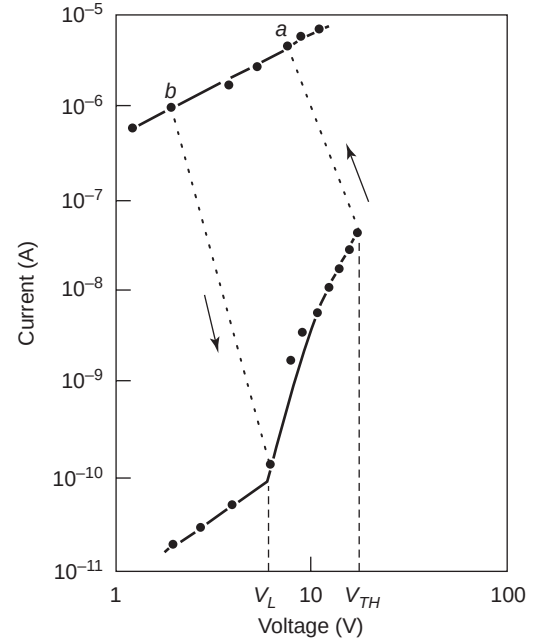


Figure 7-27 Typical DC current-voltage characteristics showing the switching and memory actions through the NDR region in anthracene thin films (6000 Å in thickness) with silver electrodes at 20°C.

method, Saji and Kao¹²⁷ have observed the distribution of the cross-sections of a typical current filament along its path between electrodes in chalcogenide glasses.

It is likely that the formation of a filament occurs when a critical injection level has been reached. We can assume that at that critical injection level, the carriers of one type (e.g., electrons) injected from the cathode have filled all the electron traps along a preferable path. Then, a negative space charge formed by the free and trapped electrons will enhance the field toward the anode, switching on hole injection and hence double injection. The voltage required for such a critical injection level to be reached is usually called the *threshold voltage* V_{TH} .

At this voltage, the conduction becomes current- or energy-controlled. Since the material is never homogeneous in structure and chemical composition, electrons and holes can always find an easy path with low resistance to drift along the applied field. This is why the shape of the current filament is irregular and its cross-section varies along its path. The current is continuous, but the current density in the filament is not uniform; it is highest in the region of the filament with a smallest cross-section. This region may produce the largest joule heating (or the highest temperature). Therefore, it becomes a hot spot in the filament.

It is likely that such a hot spot acts as a nucleus, initiating the growth of the filament. As the concentration of electrons and holes are increased by Joule heating, the increase in electrons and holes will, in turn, enhance the carrier injection from both electrodes, forming a mutual feedback carrier multiplication process. The increase of current in the filament implies a decrease in resistance of the material specimen. If a constant DC voltage is applied to a specimen connected in series with a resistor, the voltage across the specimen will decrease rapidly as soon as the filament starts to grow.

When the applied voltage across the specimen reaches V_{TH} , the current starts to increase rapidly in the filament, while the voltage across the specimen decreases accordingly, forming an NDR region. However, the I - V characteristics

in this region are very unstable until the current reaches a level at which the carriers lost from the filament by radial diffusion just balance those gained by the mutual feedback multiplication process. Under this condition, the voltage across the specimen becomes V_M , which can be considered the minimum voltage to maintain a stable microplasma in the filament. Thus, to make the specimen operate in a high-current regime, it is necessary to apply a voltage across the specimen higher than V_M after the formation of the filament.

7.6.2 The Poole–Frenkel Detrapping Model

The Poole–Frenkel effect is sometimes called the *internal Schottky effect*, since the mechanism of this effect is associated with field-enhanced thermal excitation (or detrapping) of trapped electrons or holes, which is very similar to the Schottky effect in the thermionic emission. The effect of an applied field in lowering the potential barrier for a trapped electron to escape in a one-dimensional model is shown in Figure 7-28, which is very similar to Figure 6-12. Both effects are due to coulombic interaction between the escaping electron and a positive charge, but they differ in that the positive charge is fixed for the Poole–Frenkel trapping barrier, while the positive charge is a mobile image charge for the Schottky barrier. The results in the barrier lowering due to the Poole–Frenkel effect are twice that due to the Schottky effect. The amount of the barrier lowering due to the Poole–Frenkel effect is

$$\Delta E_{pF} = \left(\frac{q^3 F}{\pi \epsilon} \right)^{1/2} = \beta_{pF} F^{1/2} \quad (7-298)$$

where β_{pF} is called the *Poole–Frenkel constant*.

$$\beta_{pF} = \left(\frac{q^3}{\pi \epsilon} \right)^{1/2} \quad (7-299)$$

Equation 7-298 differs from $\Delta \phi_B$ given by Equation 6-25 by a factor of two because the coulombic attractive force to the electron is $q^2/4\pi\epsilon(r_{pF})^2$ for the Poole–Frenkel effect, and $q^2/4\pi\epsilon(2x_m)^2$ for the Schottky effect. Thus, the Poole–Frenkel effect is effective only for traps

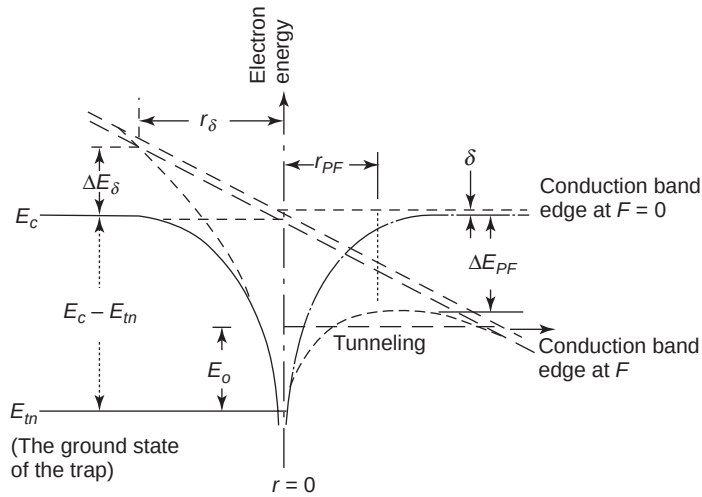


Figure 7-28 Schematic diagram illustrating the Poole-Frenkel effect.

that are neutral when filled and positively charged when empty. Traps that are neutral when empty and charged when filled will not manifest this effect for lack of coulombic interaction. The degree of field-dependent detrapping for various types of traps is summarized in Table 7-4.

Many investigators have used the Poole-Frenkel model to interpret high-field transport phenomena in insulators and semiconductors.¹²⁸⁻¹³⁴ It is obvious that the Poole-Frenkel effect is observed when electric conduction is bulk limited, and the Schottky effect is observed when electric conduction is electrode-limited. In both cases, $\ln \sigma \propto F^{1/2}$. To distinguish one case from the other, we must rely on the comparison of experimental and theoretical values of β_{sc} and β_{pF} , which can be calculated quite accurately if the value of ϵ for high frequencies, which should be approximately equal to η^2 (the refractive index of the material), is known. However, many investigators have reported that the value of β_{pF} determined from the slope of the $\ln \sigma - F^{1/2}$ plots does not agree with the theoretical one for the whole range of the experimental $J-V$ characteristics.¹²⁸⁻¹³⁴ This implies that there may be several slopes in the $\ln \sigma - F^{1/2}$ plots and that even from the best-fit curve, the value of ϵ is usually not close to what would be expected.

A trapped electron can be thermally released not only in the forward direction of the applied field, in which the potential barrier is lowered by the field; it can also be thermally released in other directions, although the probability is smaller. The relative probabilities for a trapped electron to be thermally released in the forward direction and in the reverse direction of the applied field are field dependent. The simple expression for the barrier lowering given in Equation 7-298, derived by Frenkel,¹²⁹ is based on a one-dimensional (planar) model. Several investigators have shown that a three-dimensional treatment gives field- and temperature-dependent conductivities that are different from the following original Poole-Frenkel relation

$$\sigma = \sigma_o \exp(\beta_{pF} F^{1/2} / 2kT) \quad (7-300)$$

where σ_o is the low-field conductivity.^{130,135} From Figure 7-28, it can be seen that the potential energy of an electron attracted to a positively charged trap located at $r = 0$ under the influence of a uniform field $\vec{F}$ may be written as

$$\phi = -q^2 / 4\pi\epsilon r - q(\vec{F} \cdot \vec{r}) \quad (7-301)$$

by setting $\frac{d\phi}{dr} = 0$ we obtain the barrier lowering

$$\Delta\phi_{pF} = \beta_{pF}(F \cos\theta)^{1/2} \quad (7-302)$$

which occurs at

$$r_{pF} = \left(\frac{q}{4\pi\epsilon F \cos\theta} \right)^{1/2} \quad (7-303)$$

where θ is an angle between $\vec{F}$ and $\vec{r}$. The potential barrier is lowered only in the forward direction for $0 \leq \theta \leq \pi/2$. In the reverse direction, Ieda et al.¹³⁰ have assumed that there is a state denoted by δ in which, by interaction with phonons, the transition probability that an electron to a distance r_δ will become a free carrier is much larger than that for the electron to the ground state. They have derived an expression for r_δ as

$$r_\delta = \frac{q^2}{4\pi r \epsilon \delta} \quad (7-304)$$

and for the increase of the potential barrier in the reverse direction as

$$\Delta E_\delta = \beta_{pF} F \cos\delta / 4\delta \quad (7-305)$$

Using the effective barrier lowering in the forward direction for $r_{pF} \leq r_\delta$

$$\Delta E_{pF} = \beta_{pF}(F \cos\theta)^{1/2} - \delta \quad (7-306)$$

and for $r_{pF} \geq r_\delta$

$$\Delta E_{pF} = \beta_{pF}^2 F \cos\theta / 4\delta \quad (7-307)$$

Ieda et al.¹³⁰ have also modified Equation 7-300, based on a one-dimensional treatment, to the following equation, based on a three-dimensional treatment:

$$\sigma = \sigma_o \left(\frac{4\gamma}{\alpha^2 F} \right) \sinh(\alpha^2 F / 4\gamma) \quad (7-308)$$

for $\alpha F^{1/2} \leq 2\gamma$

and

$$\sigma = \sigma_o \left(\frac{1}{\alpha^2 F} \right) [(\alpha F^{1/2} - 1) \exp(\alpha F^{1/2} - \gamma) - 2\gamma \exp(-\alpha^2 F / 4\gamma) + \exp(\gamma)] \quad (7-309)$$

for $\alpha F^{1/2} \geq 2\gamma$

where

$$\alpha = \beta_{pF} / 2kT \quad (7-310)$$

$$\gamma = \delta / 2kT \quad (7-311)$$

It can be seen that at low fields, $\alpha F^{1/2} / 4\gamma \ll 1$ the conductivity follows a simple Ohm's law and that at high fields Equation 7-309 approaches Equation 7-300 asymptotically. In the intermediate fields, the field-dependent σ follows Equations 7-308 and 7-309, which are quite different from Equation 7-300. Ieda et al.¹³⁰ have reported that Equations 7-308 and 7-309 are very consistent with the J - F characteristics as functions of temperature in polyacrylonitrile films and in SiO films.

There are many other modifications of the original Poole-Frenkel model based on various assumptions, such as those by Hartke,¹³⁵ Hill,¹³¹ Antula,¹³³ and Adamic and Calderwood.¹³⁶ The reader interested in their approaches is referred to their original papers.

7.6.3 The Onsager Detrapping Model

Unlike the Poole-Frenkel model, which is based on the assumption that the probability for the trapped carriers to be thermally released increases with increasing applied field because of the field-enhanced lowering of the traps' potential barriers, the Onsager model is based on a completely different assumption: that the probability $p(r, \theta, F)$ that an electron-hole pair (or electron-donorlike trap) thermalized with an initial separation r and orientation θ relative to the applied field F to escape initial recombination increases with increasing applied field. For the former model, the carriers are regenerated after they have been captured in the traps. For the latter model, the carriers are separated before they are trapped. On the basis of the postulate that the theory of geminate (or initial) recombination reduces to the problem of Brownian motion of one particle under the action of the coulombic attraction and the applied electric field, Onsager^{137,138} has approached this problem by solving the equation of Brownian motion

$$\frac{\partial f}{\partial t} = \frac{kT}{q} \mu_n \text{div} \left[\exp\left(-\frac{U}{kT}\right) \text{grad } f \exp\left(\frac{U}{kT}\right) \right] \quad (7-312)$$

where f is a probability function, μ_n is the electron mobility, and U is the coulombic potential modified by the applied field, given by

$$U = -(q^2/4\pi\epsilon r) - qFr \cos \theta \quad (7-313)$$

Using the boundary condition of zero initial separation between the electron and the positively charged center, the probability of ionization under the steady-state condition ($\partial f/\partial t = 0$) in the presence of the applied field is increased by the ratio^{139,140}

$$\begin{aligned} \frac{P(F)}{P(O)} &= \frac{J_1(j\alpha)}{j\alpha/2} \\ &= 1 + \frac{1}{2!} \left(\frac{\alpha^2}{4} \right) + \frac{1}{2!3!} \left(\frac{\alpha^2}{4} \right)^2 \\ &\quad + \frac{1}{3!4!} \left(\frac{\alpha^2}{4} \right)^3 + \dots \end{aligned} \quad (7-314)$$

where J_1 is the Bessel function of the first order and

$$\alpha = \left(\frac{q^3}{\pi\epsilon} \right)^{1/2} \frac{F^{1/2}}{kT} \quad (7-315)$$

If an initial separation between the electron and the positively charged center of r_o rather than zero is used as the boundary condition, Equation 7-314 is modified to¹³⁹

$$\begin{aligned} \frac{P(F)}{P(O)} &= 1 + \frac{1}{2!} \left(\frac{\alpha^2}{4} \right) + \frac{1}{2!3!} \left(\frac{\alpha^2}{4} \right)^2 \left(1 - \frac{2r_o}{r_c} \right) \\ &\quad + \frac{1}{3!4!} \left(\frac{\alpha^2}{4} \right)^3 \left(1 + 3! \frac{r_o^2}{r_c^2} - 3! \frac{r_o}{r_c} \right) + \dots \end{aligned} \quad (7-316)$$

where r_c is the cut-off separation distance to separate the bound and the free carriers and is defined by

$$r_c = \frac{q^2}{4\pi\epsilon kT} \quad (7-317)$$

Carriers within the separation distance $0 \leq r \leq r_c/2$ are bound charge carriers. When $r_o = 0$, Equation 7-316 reduces to Equation 7-314. Figure 7-29 shows the difference between

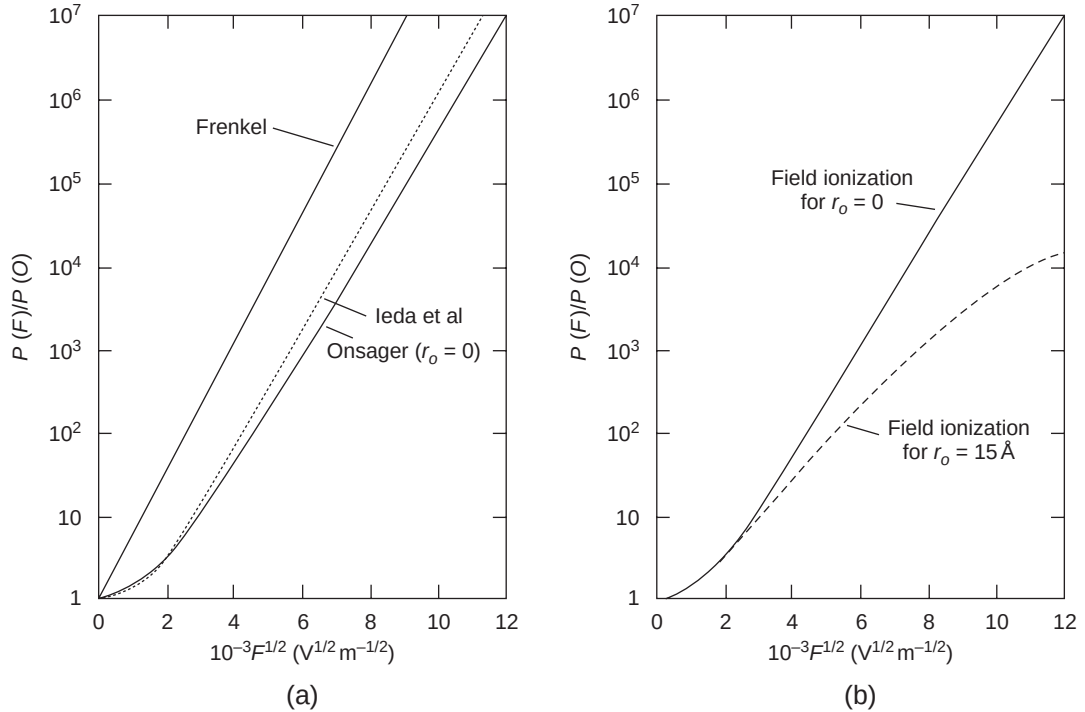


Figure 7-29 Computed results for the electric-field enhanced probability of ionization as a function of the square root of the electric field for a material with a dielectric constant of 3 at $T = 293$ K: (a) calculated from Equation 7-314 based on the Onsager model and compared to those calculated from Equations 7-300, 7-308, and 7-309 based on the Poole-Frenkel model and (b) calculated from Equation 7-314 for $r_o = 0$ and from Equation 7-316 for $r_o = 15$ Å based on the Onsager model.

Equation 7-316 and Equation 7-314. It also compares the Onsager model with the Poole–Frenkel model. It can be seen that the $P(F)/P(0)$ ratio is smaller for the Onsager model for a given field and that the larger the value of r_o is, the smaller this ratio.

Using the Onsager model, Pai¹³⁹ has derived an expression for the field-dependent electric conductivity, which is

$$\sigma(F) = K_1 \left[\frac{J_1(j\alpha)}{j\alpha/2} \right]^{1/2} \exp\left(-\frac{E_c - E_m}{2kT}\right) \quad (7-318)$$

and the field-dependent current density

$$J(F) = K_1 F \left[\frac{J_1(j\alpha)}{j\alpha/2} \right]^{1/m} \exp\left(-\frac{E_c - E_m}{mkT}\right) \quad (7-319)$$

where K_1 is a constant and the parameter m ranges between 1 and 2, depending on the complexity of the distribution of states in the forbidden band gap (donors, acceptors, traps, and the like). Equations 7-318 and 7-319 are based on uniform field distribution in a solid specimen. If the conduction involves carrier injection from the electrodes and space charge effects, these two equations must be modified to take into account the effect of field distribution.

Onsager has also derived the expression for $p(r, \theta, F)$, which is given by

$$p(r, \theta, F) = \exp(-A) \exp(-B) \times \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{A^m}{m!} \frac{B^{m+n}}{(m+n)!} \quad (7-320)$$

where

$$A = \frac{q^2}{4\pi\epsilon kTr} \quad (7-321)$$

$$B = (qFr/2kT)(1 + \cos\theta)$$

By defining ϕ_{po} as the ionization quantum yield (the efficiency of production of thermalized ion pairs per absorbed photon) and $g(r, \theta)$ as the initial spatial distribution of thermalized pair configuration (separation between ions of each ion pair or between electron and ionized donor), the carrier quantum yield (carrier generation efficiency) is given by

$$\phi_p(r_o, F) = \phi_{po} \int p(r, \theta, F) g(r, \theta) d^3r \quad (7-322)$$

By assuming that ϕ_{po} is independent of the applied electric field and $g(r, \theta)$ is an isotropic δ function and is expressed as

$$g(r, \theta) = (4\pi r_o^2)^{-1} \delta(r - r_o) \quad (7-323)$$

then substituting Equations 7-320 and 7-323 into Equation 7-322 and carrying out the integration, we obtain¹⁴⁰

$$\begin{aligned} \phi_p(r_o, F) &= \phi_{po} \left(\frac{kT}{qFr_o} \right) \exp(-A) \sum_{m=0}^{\infty} \frac{A^m}{m!} \\ &\times \sum_{n=0}^{\infty} [1 - \exp(-qFr_o/kT)] \\ &\times \sum_{\ell=0}^{m+n} \left(\frac{qFr_o}{kT} \right)^{\ell} \frac{1}{\ell!} \end{aligned} \quad (7-324)$$

The first few terms of Equation 7-324 can be written as

$$\begin{aligned} \phi_p(r_o, F) &= \phi_{po} \exp\left(-\frac{r_c}{r_o}\right) \left[1 + \left(\frac{q}{kT} \right) \frac{1}{2!} Fr_c \right. \\ &+ \left(\frac{q}{kT} \right)^2 \frac{1}{3!} F^2 r_c \left(\frac{1}{2} r_c - r_o \right) \\ &+ \left(\frac{q}{kT} \right)^3 \frac{1}{4!} F^3 r_c \left(r_o^2 - r_o r_c + \frac{1}{6} r_c^2 \right) \\ &\left. + \dots \right] \end{aligned} \quad (7-325)$$

Many investigators have reported that Equations 7-319 and 7-325 agree well with the experimental results on electric field and temperature dependence of electrical conductivity, and also with the electron and hole quantum yields in organic and inorganic solids.^{139–145} It should be noted that Onsager's theory, because of its inherent microscopic diffusive nature, takes into account the integrated ionization probability within the 4π solid angle. The physical soundness of the Onsager theory is to include the reverse ionization (or escape) probability in a most natural way, devoid of any approximation.¹³⁹

7.6.4 Field-Dependent Carrier Mobilities

It is well known that for crystals doped with either donor or acceptor impurities, charge

carriers (electrons and holes) will suffer two main types of scattering:

1. Scattering with phonons (lattice vibration)—for nonpolar crystals mainly with acoustic phonons, such as *Ge* and *Si*, and for polar crystals with optical phonons in addition, such as *GaAs*
2. Scattering with impurities (coulombic interaction)

Under a fixed electric field, the carrier mobilities controlled by these scatterings are strongly dependent on temperature.^{146–148} Apart from these two scattering mechanisms, other mechanisms also play important roles in determining the actual mobilities, such as intravalley and intervalley scattering.¹¹⁶ In general, at a fixed temperature and a fixed applied field, carrier mobility decreases with increasing concentration of impurities and with increasing effective mass of the carriers.

Under an applied field, an electron tends to gain energy from the field due to acceleration until it encounters a scattering either with lattice vibration waves or with other particles. If there is a net gain in energy, the energy of the electron increases. For an electron in an insulator or a nondegenerate semiconductor, an increase in energy by an amount kT is usually considered a large change in the mean energy. By assuming the electron energy distribution to be Maxwellian, the mean velocity of the electron may be written as

$$v = (2kT_e/m^*)^{1/2} \quad (7-326)$$

where T_e is defined as the electron temperature based on the expression of the electron energy in terms of kT_e . It can be seen from this simple equation that electron mobility would be independent of field only if $T_e = T$; if $T_e > T$, the electron becomes hot and electron mobility becomes field dependent.

The field dependence of the mean (or effective) mobility in crystalline semiconductors has been extensively reviewed by Conwell,¹¹⁶ and applied particularly to *Ge*, *Si*, and some III–V compounds. There exists a great deal of information about the energy band structures and lattice vibration modes of these semiconduc-

tors. Such information is not available for organic and inorganic dielectric materials, such as polymers or ceramics. However, the calculation of the field-dependent carrier mobility requires a knowledge of the energy distribution function of the carriers, and the mathematical solution for this is usually lengthy and complex even for simple semiconductors.^{2,116} This section discusses briefly the physical concept of field-dependent carrier mobilities.

In cases with a constant mean free path, Lampert^{149,54} has proposed the following formula for evaluating the mean field-dependent mobility

$$\mu(F) = \frac{1}{2} \mu_o (F_1/F) \{ [1 + (4F/F_1)]^{1/2} - 1 \} \quad (7-327)$$

where u_o is the mobility independent of electric field, which occurs only at very low fields, and F_1 is a critical field at which the drift velocity of the carriers varies as the square root of the applied field because the carriers become hot carriers. Thus, at low to moderate fields (i.e., $F < F_1$), Equation 7-327 can be approximated to

$$\mu(F) = \mu_o [1 - (F/F_1)] \quad (7-328)$$

At high fields (i.e., $F > F_1$), Equation 7-327 can be approximated to

$$\mu(F) = \mu_o (F_1/F)^{1/2} \left[1 - \frac{1}{2} (F_1/F)^{1/2} \right] \quad (7-329)$$

These equations predict well the variations of electron mobility with field in silicon, in which $u = u_o$ at very low fields, $\mu \propto F^{-1/2}$ at moderate fields, and $\mu \propto F^{-1}$ at high fields.¹⁵⁰

Dielectric materials consist of both shallow and deep traps. In the case of a single type of carrier (e.g., electrons) injected from an electrode, the electrons will encounter many traps during their journey to the other electrode. We can assume that at any moment there are n free electrons traveling across the specimen with the mobility u_n , and n_t trapped electrons always stationary in the traps. This assumption is reasonable only for cases in which the carrier transit time t_t is smaller than the time the trapped electrons remain in the traps, so that during the carrier transit time we can see only n electrons

contributing to electrical conduction. But if each injected electron ($n + n_t$) spends on average a total time τ_n as a free carrier and a total time τ_m in the trap, then the transit time can be expressed as

$$t_i = \tau_n + \tau_m \quad (7-330)$$

and the mean (or effective) mobility as

$$\mu_{\text{eff}} = \frac{\tau_n \mu}{\tau_n + \tau_m} \quad (7-331)$$

Equation 7-331 implies that during the carrier transit time, we can see all injected electrons contributing to electrical conduction and that there is no separation between free and trapped carriers. This assumption is reasonable for cases in which τ_n or τ_m is smaller than t_i . It should be noted that u can be field dependent due to scatterings with phonons and coulombic impurities, and τ_n and τ_m are field dependent due to the field-enhanced detrapping processes (see Sections 7.6.2 and 7.6.3).

Assuming that all injected electrons are available to drift across the specimen, we can write

$$(n + n_t)\mu_{\text{eff}} = nu \quad (7-332)$$

Assuming that there is only one trapping level in the forbidden gap located at E_t below the conduction band edge E_c , we can see the effect of trapping level on effective carrier mobility from the following relation

$$\begin{aligned} \frac{\mu_{\text{eff}}}{u} &= \frac{n}{n + n_t} = \frac{n}{n_t} \quad \text{If } n_t \gg n \\ &= (N_c/N_t) \exp(-E_t/kT) \end{aligned} \quad (7-333)$$

where N_t is the total concentration of traps and N_c is the effective density of states in the conduction band. For example, $N_c = 10^{19} \text{ cm}^{-3}$, $N_t = 10^{16} \text{ cm}^{-3}$ for silicon if $E_t = 0.2 \text{ eV}$ (or 7 kT) below E_c , $u_{\text{eff}}/u = 0.1$ for shallow traps. But if $E_t = 0.5 \text{ eV}$ (or 17.5 kT) below E_c , u_{eff}/u becomes 10^{-5} for deep traps. This example offers a numerical feeling about the effects of traps and applied field on effective carrier mobility.

It is worth describing briefly the physical concept of the intervalley scattering mechanism. In semiconductors with a two-valley conduction band, such as GaAs shown in Figure

7-30(a), the electrons in subband 1 (lower valley) have their mobility u_1 and the electrons in subband 2 (upper valley) have their mobility u_2 ; u_1 is much larger than u_2 because the effective mass of the electrons in the lower valley is much smaller than in the upper valley. Unlike the S-shaped J - F characteristics, which are current controlled, the N-shaped J - F characteristics are voltage controlled. The NDR region is associated with the transfer of electrons from a high mobility state to a low mobility state, as shown in Figure 7-30(b).

In thermal equilibrium, subband 1 contains essentially all the conduction electrons, while subband 2 contains very few electrons. However, when an applied field is larger than a certain critical value (or threshold field), which is about 3 kV cm^{-1} for GaAs, electrons in the lower valley (subband 1) can gain sufficient energy from the field to be scattered into the upper valley. The effective density of states in the upper valley is much larger than in the lower valley, so the probability that the electrons will be scattered into the upper valley is high. This electron transfer process leads to the formation of the NDR region.

Figure 7-30(b) shows that at fields below F_1 , almost all electrons are in the lower valley. At fields above F_2 , almost all electrons are in the upper valley. At fields between F_p and F_v , the intervalley electron transfer process produces an NDR region simply because a portion of the total number of electrons drifts at lower velocities. Since a positive resistance consumes power and a negative resistance supplies power, a negative differential resistance should generate AC power, that is, current oscillation in the microwave regime. The latter phenomenon was first observed by Gunn.¹⁵¹ Later, it was called the *Gunn effect*, and the microwave frequency oscillators based on this effect are called *Gunn diodes*.

There are several modes of operation for such transferred electron devices, and each mode is related to the process of the transfer of electrons from a high mobility state to a low mobility state. Obviously, a small fluctuation in the electron concentration in the NDR region can lead to the formation of a dipole layer,

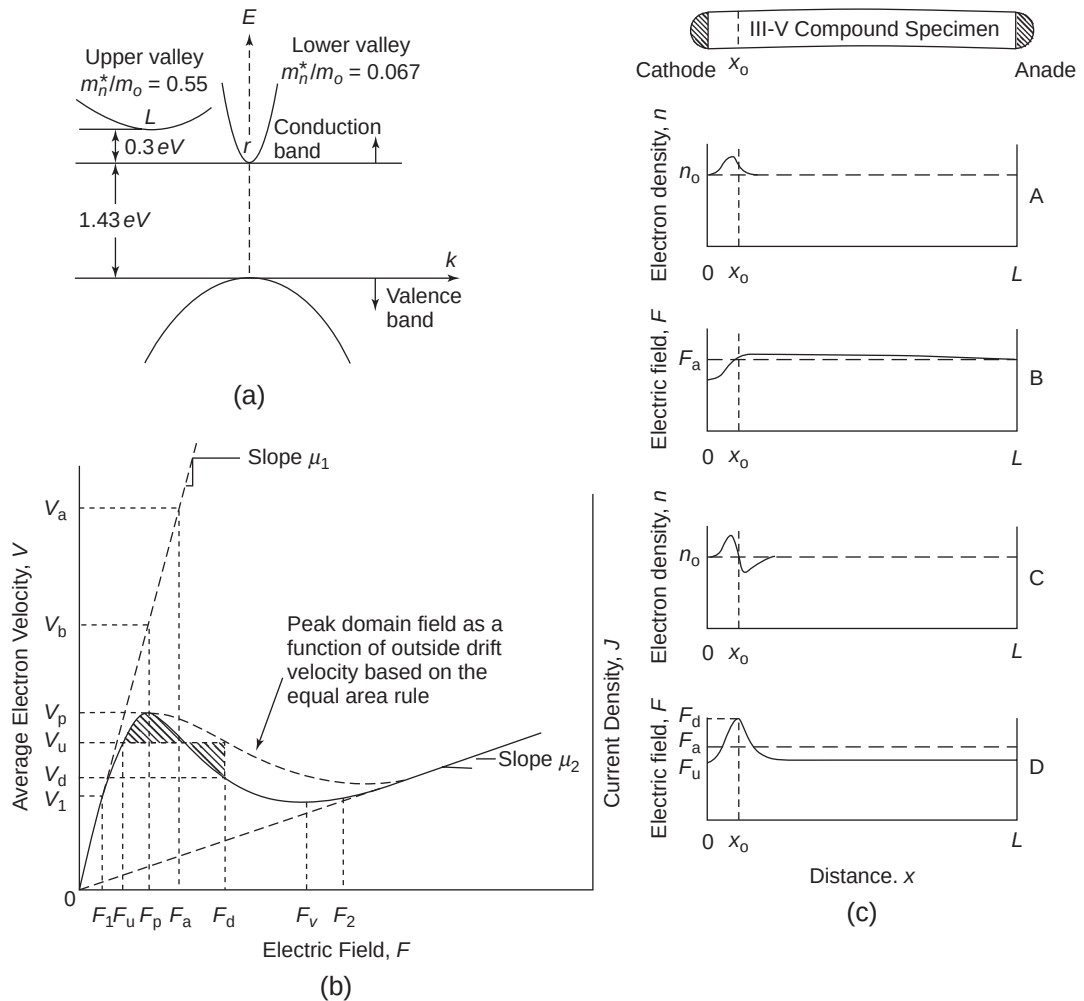


Figure 7-30 Schematic diagrams illustrating (a) the two-valley conduction band in GaAs, (b) the electron velocity-field characteristics for two-valley semiconductors, and (c) the formation of a high-field domain near the cathode.

which is generally referred as a *domain*. The movement of such domains results in the oscillation of the electric current.

Suppose that we have a III-V compound semiconductor specimen, as shown in Figure 7-30(c). Consider an average electric field $F_a (= V/L$ where V is the applied voltage and L is the specimen length), which is larger than F_p and is maintained as constant. When F_a is applied across the specimen, the electrons from the cathode and in the bulk will be accelerated under the field. Some of them may acquire enough energy to jump into subband 2.

However, the electrons near the cathode must travel a certain distance x_0 from the cathode before they have a chance to gain enough energy to transfer to subband 2. This distance x_0 , which is field dependent, should be of the order of $\Delta E/qF_a$ where ΔE is the separation in energy between the two subband edges and q is the electron charge. For n -GaAs, the value of x_0 is estimated to be about 10^{-4} cm.

This argument implies that the electrons in the region between $x = 0$ and $x = x_0$ are mainly in subband 1 with the average drift velocity v_a , while those in the region between $x = x_0$ and

$x = L$ are partly in subband 1 and partly in subband 2. Therefore, their average drift velocity is v_u . See Figures 7-30(b) and (c). Since v_a is much larger than v_u , an electron space charge will build up to form an accumulation layer near the cathode, tending to reduce the field at the cathode, as shown in A of Figure 7-30(c). There, n_o is the electron density in equilibrium and assumed to be the doping concentration, so the specimen is electrically neutral when $n = n_o$. This accumulation layer will cause the field distribution to change from the uniform field F_a to that shown by solid line in B of Figure 7-30(c). As the field at that instant increases from the lowest value at $x = 0$ to a field slightly higher than F_a at $x > x_o$, the velocity of the electrons, which are mainly in subband 1, increases spatially from the lowest value at $x = 0$ to the largest at x close to x_o . Hence, the electron density at the leading edge of the accumulation layer becomes depleted to form a dipole layer, as shown in C of Figure 7-30(c). This dipole layer then produces a high-field domain, as shown in D of Figure 7-30(c).

Taking into account the motion of electrons as a whole toward the anode, domain dynamics can be described simply by the following two one-dimensional equations¹⁵⁵:

$$\frac{dF}{dx} = \frac{q}{\epsilon}(n - n_o) \quad (7-334)$$

$$J = qnv(F) + \epsilon \frac{dF}{dt} \quad (7-335)$$

We shall not include the mathematical analysis here. It is clear that if the applied field is F_a , the domain consists of an accumulation layer where $n > n_o$, followed by a depletion layer where $n < n_o$. The carrier concentration $n = n_o$ at two values of the field, that is, $F = F_u$ outside the domain and $F = F_d$ being the peak domain field. By using the equal area rule,¹⁵³ the value of the peak domain field F_d can be determined if the value of the outside field F_u is known, as shown by the dashed curve in Figure 7-30(b). The formation and dynamics of domains have been extensively studied by many investigators.¹⁵²⁻¹⁵⁷ For details of this phenomenon, see these references.

7.7 Transitions between Electrical Conduction Processes

Most metallic contacts to semiconductors or insulators (for simplicity, the term *solids* will be used to refer to these materials) are generally nonohmic. The supply of charge carriers from a nonohmic contact is not unlimited but depends on the height and the width of the potential barrier near the contact. A nonohmic contact may act as a nearly ohmic contact at low electric fields, but at an appropriate range of fields it may act as a nearly blocking contact, with the carrier supply limited by Schottky-type thermionic emission. This nearly blocking behavior may be changed again to nearly ohmic behavior if the applied field can be increased, without causing breakdown of the solid, to a level that greatly reduces the width of the potential barrier to enable the carriers to inject by Nordheim–Fowler type tunneling.

For a given metal used for both anode and cathode, at low fields one contact may behave as a nearly ohmic contact for one type of carrier (e.g., electrons) and the other as a nearly blocking contact for the other type of carrier (e.g., holes), as shown in Figure 7-31(a) and (b). In cases in which different metals are used for the anode and the cathode, if both anode and cathode behave as nearly ohmic contacts, both contacts will supply carriers at low fields, forming a bulk-limited double-injection conduction, as shown in Figure 7-31(c). But if both anode and cathode behave as nearly blocking contacts, then the electrical conduction depends only on the carriers replenished by thermal excitation in the bulk, as shown in Figure 7-31(d), carrier injection from contacts being negligible.

7.7.1 Basic Transition Processes

We shall use the simplest metal–solid contacts to show the transitions between electrical conduction processes in a solid without defects (without traps).¹⁵⁸

A Solid between Similar Contacts

Using the same metal for both the anode and the cathode, we may make the contacts nearly

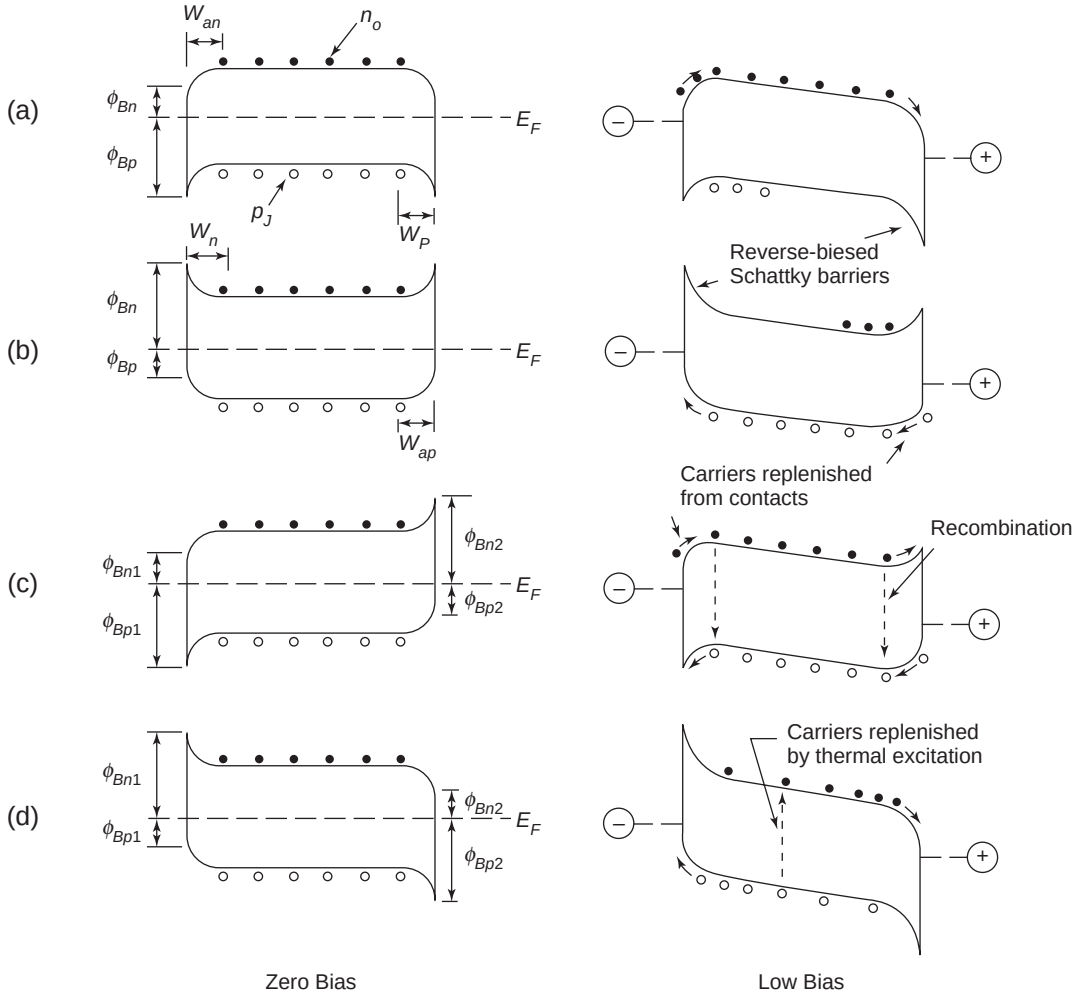


Figure 7-31 Nearly ohmic and nearly blocking contacts between a metal and a semiconductor or an insulator under thermal equilibrium conditions (zero applied field) and at low applied fields: (a) nearly electron-ohmic contacts and nearly hole-blocking contacts for cases using the same metal with $\phi_s > \phi_m$ for both the cathode and the anode, (b) nearly electron-blocking contacts and nearly hole-ohmic contacts for cases using the same metal with $\phi_s < \phi_m$ for both the cathode and the anode, (c) both electron- and hole-injecting contacts are nearly ohmic, but both electron- and hole-collecting contacts are nearly blocking, for cases using a metal with $\phi_s > \phi_{m1}$ for the cathode and a metal with $\phi_s < \phi_{m2}$ for the anode; (d) both electron- and hole-injecting contacts are nearly blocking, but both electron- and hole-collecting contacts are nearly ohmic, for cases using a metal with $\phi_s < \phi_{m1}$ for the cathode and a metal with $\phi_s > \phi_{m2}$ for the anode.

ohmic for electron injection and nearly blocking for hole injection, as shown in Figure 7-31(a), or vice versa, as shown in Figure 7-31(b). The two cases are similar, so we shall analyze only the case shown Figure 7-31(b). In this case, the barrier height of the nearly ohmic contact at the anode is

$$\phi_{BP} = E_g - (\phi_m - \chi)$$

which is much smaller than that of the nearly blocking contact at the cathode, which is

$$\phi_{Bn} = \phi_m - \chi$$

where χ is the electron affinity of the solid. This implies that at an applied field, the anode can replenish the holes in the solid extracted by the cathode, since $N_v \exp(-\phi_{BP}/kT) > n_i$, but

the cathode cannot replenish the electrons extracted by the anode, since $N_c \exp(-\phi_{Bn}/kT) < n_i$. As a result, electron density n after a transient event will reduce to a value

$$n_b = N_c \exp(-\phi_{Bn}/kT)$$

Supposing that $\phi_{BP} = E_g/3$ and $\phi_{Bn} = 2E_g/3$, where E_g is the energy band gap, then

$$n_b/n_i = \exp(-E_g/6kT)$$

which could be very small. For example, $n_b/n_i = 0.0015$ for $E_g = 1$ eV at $T = 300$ K. Thus, $p > n_i > n_b$. We shall derive the J - V characteristics for this case in three regimes.

Low-Field Bulk-Limited Regime

In this regime $p > n_i$ and $n_b < n_i$. Based on Equations 7-237 and 7-240, we can write

$$J = qn_b(\mu_n + \mu_p)F - \epsilon\mu_p F \frac{dF}{dx} \quad (7-336)$$

By setting $qn_b(\mu_n + \mu_p) = \sigma_0$ and $J/\sigma_0 = F_0$ and using the boundary condition

$$F_a = F(x = d - W_{op} \simeq d) = 0$$

Integration of Equation 7-336 yields

$$\frac{(d-x)\sigma_0}{\epsilon\mu_p F_0} = -\frac{F}{F_0} - \ln\left(1 - \frac{F}{F_0}\right) \quad (7-337)$$

Since F_0 is much larger than F , the expansion of the logarithmic term truncated to the cubic term is a good approximation. Thus, Equation 7-337 can be simplified to

$$\left(\frac{d-x}{\epsilon\mu_p F_0}\right) = \frac{1}{2}\left(\frac{F}{F_0}\right)^2 - \frac{1}{3}\left(\frac{F}{F_0}\right)^3 \quad (7-338)$$

From Equations 7-336 and 7-252 with $V = 0$ at $x = 0$ and $V = V$ at $x = d$, the voltage V_x at any point x is given by

$$V - V_x = \frac{\epsilon\mu_p F_0}{3} \left(\frac{F}{F_0}\right)^3 \quad (7-339)$$

Substituting Equation 7-339 into Equation 7-338 and letting $x = 0$, we obtain

$$\frac{(J - \sigma_0 V/d)^3}{J^2} = \frac{9}{8} \epsilon\mu_p \frac{V^2}{d^3} \quad (7-340)$$

The ohmic term (the term with a linear J - V relation) is contributed only by the carriers that

maintain neutrality in the solid. Since the right-hand side of Equation 7-340 is always larger than $\sigma_0 V/d$ for $V > 0$, there is no ohmic region (linear J - V behavior region). However, for large applied voltages, Equation 7-340 reduces to a single-injection Mott-Gurney equation⁵⁴:

$$J = \frac{9}{8} \epsilon\mu_p \frac{V^2}{d^3} \quad (7-341)$$

By expressing $J \propto V^m$, the current is contributed by two types of carriers at very low fields with $m > 2$, but it becomes mainly a hole current at higher fields with $m = 2$, as shown in Figure 7-32(a) for $V < V_c$.

Medium-Field Contact Limited Regime

Equation 7-341 is valid until the applied voltage reaches such a value that hole injection becomes saturated or electron injection becomes appreciable. If electron injection is still negligibly small after the onset of hole current saturation, then the current becomes mainly a field-enhanced thermionic emission hole current, which is given by

$$J = A_p \exp\left(-\frac{\phi_{BP}}{kT}\right) \exp(\beta F^{1/2}/kT) \quad (7-342)$$

$$\simeq J_{op}(1 + \beta F^{1/2}/kT)$$

(see Equations 6-25 and 6-34) where $A_p = A^*T^2$

$$J_{0p} = A_p \exp\left(-\frac{\phi_{BP}}{kT}\right) \quad (7-343)$$

The electrical conduction gradually becomes contact limited, as shown in Figure 7-32(a).

High-Field Bulk and Contact-Limited Regime

The function of the hole space charge is twofold: It reduces the field at the anode to limit the hole injection current and it enhances the field toward the cathode. When the applied voltage reaches the threshold voltage V_b , a transition from a dominant one-carrier transport to a dominant two-carrier transport occurs. By defining the threshold field at the cathode F_{cb} corresponding to the threshold voltage V_b for the onset of the field emission of electrons of the concentration $n = n_i$ from the cathode,

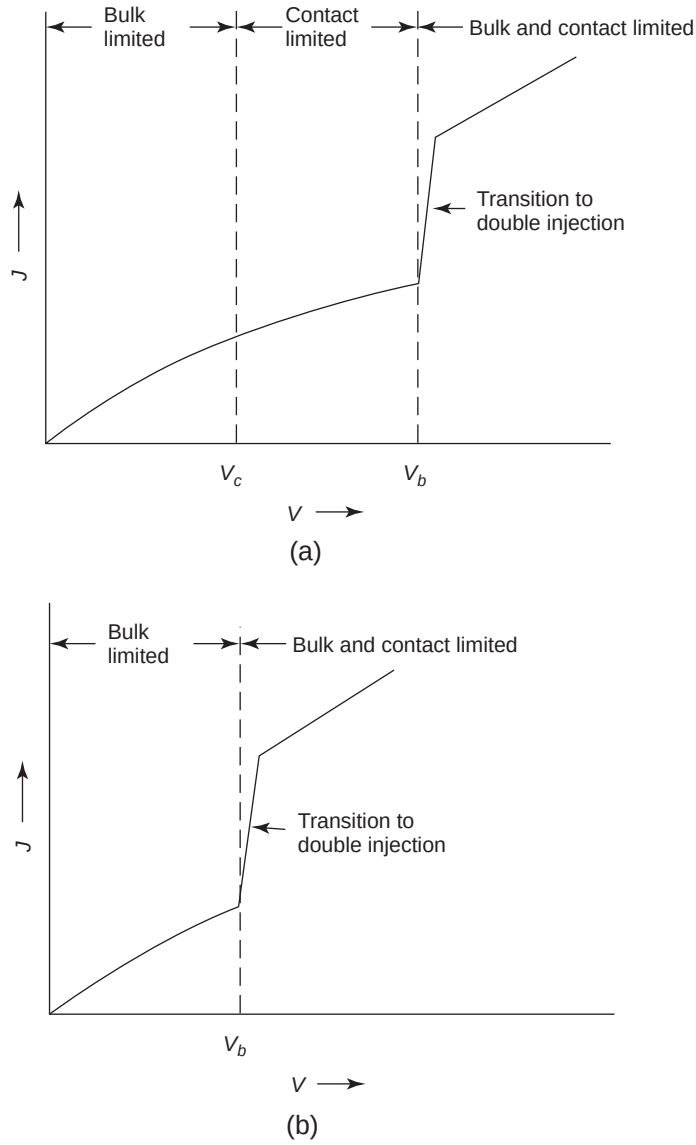


Figure 7-32 Schematic diagrams showing the J - V characteristics for an ideal intrinsic dielectric solid with a nearly ohmic contact for hole injection and a nearly blocking contact for electron injection (i.e., the same metal for both anode and cathode): (a) the occurrence of hole-injection saturation prior to the onset of double injection and (b) the onset of tunneling electron injection to switch on double injection prior to the onset of hole-injection saturation.

F_{cb} can be estimated by setting the field emission current equal to the drift current at the cathode

$$J_n = aF_{cb}^2 \exp\left(-\frac{b}{F_{cb}}\right) = qn_i\mu_n F_{cb} \quad (7-344)$$

where a and b are the constants of the Nordheim-Fowler equation. At the threshold

voltage V_b , J_p is still much larger than J_n . However, at $V > V_b$, the electron and hole densities are both larger than n_i ; recombination kinetics must be involved in deriving the J - V characteristics for this bulk- and contact-limited regime. Figure 7-32(a) shows the transition from contact-limited conduction to bulk- and contact-limited conduction.

It is also possible that under certain conditions, the threshold voltage for the onset of field emission at the cathode is lower than the threshold voltage V_c for the onset of current saturation at the anode. In this case, no contact-limited regime will be observed. The schematic illustration of the possible J - V characteristics is given in Figure 7-32(b). Note that at voltages slightly higher than V_b , the homo-space charges near the contacts have not built up to a level that limits carrier injection. Thus, the current increases very rapidly in the transition.

A Solid between Dissimilar Contacts

We shall consider one case for a solid between dissimilar contacts. For this case, $\phi_{Bn1} \approx \phi_{BP2}$, $\phi_{Bn1} > \phi_{Bn2}$ and $\phi_{BP1} < \phi_{BP2}$. Both the electron-injecting and the hole-injecting contacts are blocking, but neither the electron-collecting nor the hole-collecting contacts have barriers for

carrier extraction, as shown in Figure 7-33. Assuming that the number of carriers injected from the blocking contacts are negligibly small compared to those generated in the bulk by thermal excitation (or by photoexcitation), carriers can only be extracted from or flow out of the solid into contacts. This is a problem of double extraction rather than double injection. Before the application of a step voltage at $t = 0^-$, both the electron and hole densities are approximately uniformly distributed in the solid with $n_0 = p_0$, except near the contacts. After the application of a step voltage at $t = 0^+$, the excess holes near the cathode and the excess electrons near the anode disappear rapidly and the band becomes flat, as shown in Figure 7-33. It should be noted that if the solid contains traps, the trapped carriers form a space charge to cause band bending.¹⁵⁹

In perfectly intrinsic solids, the J - V characteristics follow Ohm's law at low fields because

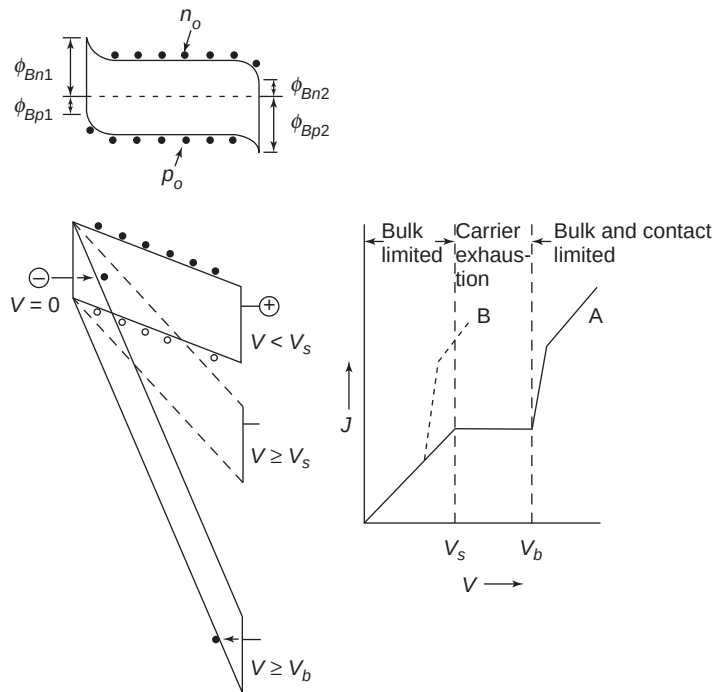


Figure 7-33 Energy band diagrams and J - V characteristics for an ideal intrinsic dielectric solid with nearly blocking contacts for both electron and hole injection (i.e., the metal for the cathode is dissimilar to the metal for the anode). Double extraction with nearly ohmic contacts for both electron and hole extraction. A (—): current saturation prior to the onset of double injection. B (----): the onset of double injection prior to current saturation.

n_0 and p_0 can be self-replenished by thermal excitation in the bulk. In this case, it is possible that current saturation may occur at $V = V_s$ and a current saturation region may exist at applied voltages between V_s and V_b . However, for $V > V_b$, carrier injection from both contacts will switch on the double-injection high-field bulk- and contact-limited regime, as shown in Figure 7-33 (curve A). It is also possible that the transition from the low-field bulk-limited regime to the high-field bulk- and contact-limited regime occurs at a voltage lower than the threshold voltage for the onset of current saturation, due to carrier exhaustion. In this case, there is no current saturation region, as shown in Figure 7-33 (curve B).

7.7.2 Transition from Bulk-Limited to Electrode-Limited Conduction Process

The current injection from a normal metallic contact into a dielectric solid follows the general Schottky thermionic emission equation given by

$$J = A^* T^2 \exp(-\phi_B/kT) \exp(\beta F^{1/2}/kT) \quad (7-345)$$

where $\phi_B = \phi_m - \chi$ and $\beta = (q^3/4\pi\epsilon)^{1/2}$ for electron injection. This equation is similar to Equation 7-342 for hole injection. The SCL current in a dielectric solid containing shallow traps is given by Equation 7-87, which, for traps distributed uniformly in space and for electron current only, is

$$J = \frac{9}{8} \epsilon \mu_n \theta_a V^2 / d^3 \quad (7-346)$$

For the cathode that is a nonohmic contact and at low fields, the contact may still supply more electrons than required to replenish those thermally excited and collected by the opposite electrode. Thus, a negative space charge exists in the vicinity of the cathode and a positive charge on the cathode surface, indicating that the electric conduction is bulk-limited (SCL).

When the applied field is increased to such a value that the number of electrons, which the cathode can inject into the specimen per

second, is exactly equal to what the anode can collect per second, there is no space charge in the specimen or any charge on the cathode surface. The electric conduction process begins to change from bulk limited to electrode limited. The critical voltage to bring on this transition can be obtained by setting Equation 7-345 equal to Equation 7-346)¹¹⁷:

$$V_c = T \left(\frac{8A^* d^3}{9\mu_n \epsilon \theta_a} \right)^{1/2} \exp\left(-\frac{\phi_B}{2kT}\right) \quad (7-347)$$

At low fields, Equation 7-345 can be simplified to

$$\begin{aligned} J &= A^* T^2 \exp(-(\phi_B/kT)(1 + F^{1/2})) \\ &\simeq A^* T^2 \exp(-\phi_B/kT) \text{ at very low fields} \end{aligned} \quad (7-348)$$

However, ϕ_B is lowered due to the combination of the applied electric field and the image force (Schottky effect), so the electrode-limited current for $V > V_c$ is governed by Equation 7-345. For $V > V_c$, there exists a negative charge on the cathode surface. The polarity of the charge on the cathode surface is generally used as a signal to distinguish between bulk-limited (SCL) and electrode-limited (injection-limited) processes. Since F at the cathode is not linearly dependent on V , the calculation of the J - V characteristics must be carried out by means of numerical methods. For further details about this topic, see references.^{117,160}

The transition from bulk-limited to electrode-limited conduction has been observed experimentally in inorganic insulators, such as Mylar, SiO, and Ta₂O₅ films,¹⁶¹ and in organic semiconductors, such as anthracene crystals.¹⁶² Figure 7-34 shows the transition from SCL currents at low fields to electrode-limited currents at high fields for anthracene crystals. The bulk $I \propto V^n$ curve with $n > 2$ indicates that the traps in the specimen may be distributed exponentially in energy, following Equation 7-110. The theoretical electrode-barrier J - V line is calculated by subtracting the extrapolated bulk voltage from the total voltage for each measured current; I_{sat} represents the saturation current due to thermionic emission in the absence of the Schottky effect.

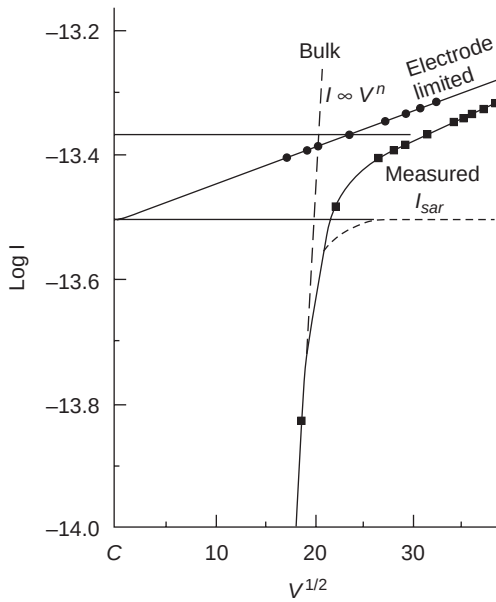


Figure 7-34 Plot of $\log I$ versus $V^{1/2}$ showing the total current (with ■ on the curve) for anthracene crystals (thickness: 2.5 mm; electrode area: 14 mm²) at T of 314.8 K, and the calculated electrode-limited current (with ● on the curve), and indicating the transition from bulk SCL current to electrode-limited current. The theoretical saturation current I_{sar} is due to Richardson thermionic emission in the absence of the Schottky effect.

7.7.3 Transition from Electrode-Limited to Bulk-Limited Conduction Process

When a dielectric solid has shallow donors, deep traps, and a blocking contact with a very narrow depletion region, as shown in Figure 7-35, all electrons from shallow donors will be captured by deep traps. Therefore, at low fields, the contact resistance is much higher than the resistance of the bulk, and the J - V characteristics are due mainly to field emission, following Equation 6-99, and will be practically independent of specimen thickness. Under such conditions, electric conduction is electrode limited. If the applied field is increased, bulk resistance and contact resistance tend to decrease, but the rate of the decrease for the former is much smaller than for the latter. At a certain critical voltage V_c , bulk resistance becomes higher than contact resistance. At this critical voltage, the transition from electrode-limited to bulk-limited conduction occurs. At $V > V_c$, the J - V characteristics become controlled by the space charge effect and the Poole-Frenkel effect, and the current becomes dependent on specimen thickness.¹¹⁷ Typical experimental results of Al-SiO-Al films demonstrating such a transition are shown in Figure 7-36.¹⁶³

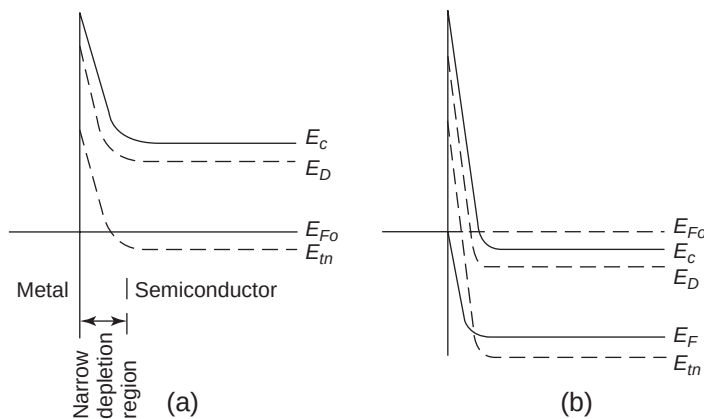


Figure 7-35 Schematic diagrams showing the energy levels for electron tunneling-field emission from a blocking contact into a dielectric solid with deep traps of trapping level E_{tn} (a) without bias in thermal equilibrium and (b) under a bias for tunneling injection. This illustrates the possible transition from electrode-limited to bulk-limited conduction processes by increasing applied fields.

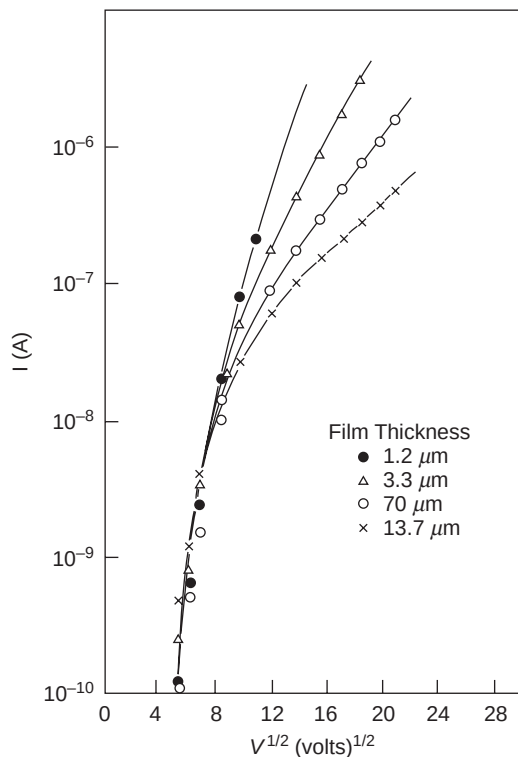


Figure 7-36 Transition from electrode-limited to bulk-limited electrical conduction in Al-SiO-Al films. Electrode area: 0.1 cm^2 .

7.7.4 Transition from Single-Injection to Double-Injection Conduction Process

If a dielectric solid has a pair of identical or different electrodes, it is possible that one of these electrodes is definitely electron injecting and the other is unable to inject holes efficiently, or vice versa, at low applied fields. But such single injection would become double injection at a critical applied field. This phenomenon has been reported by many investigators.¹⁶⁴⁻¹⁶⁶ Typical experimental results, for anthracene with silver electrodes, are shown in Figure 7-37.

For voltages below the threshold voltage V_{th} for the onset of electroluminescence, J is proportional to V^2 , following the relation for single injection into a solid containing hole-traps in a discrete energy level. The $\ln J - 1/T$ plot for $V = 2.0 \text{ kV}$ gives an activation energy E_{act} of 0.56 eV , which can be interpreted as this discrete

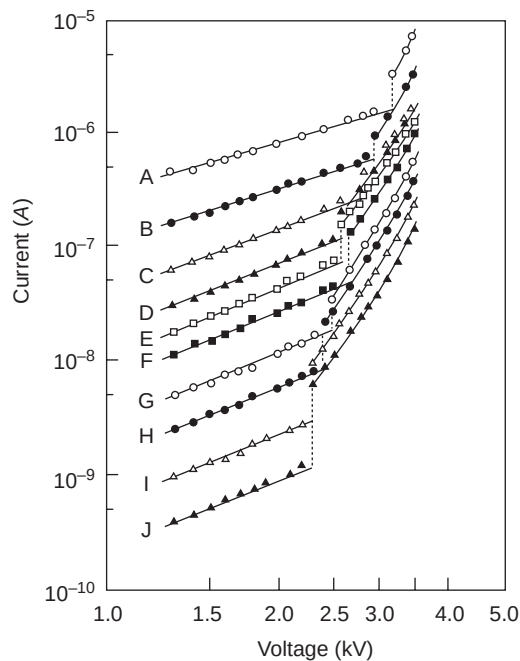


Figure 7-37 The DC current-voltage characteristics of anthracene crystals showing the transition from a single-injection to a double-injection process at high fields. Specimen thickness: 0.9 mm ; electrode diameter: 1.6 mm . Temperature: A, 87°C ; B, 63°C ; C, 51°C ; D, 41°C ; E, 34.5°C ; F, 20°C ; G, 10°C ; H, 0°C ; I, -10°C ; J, -20°C .

energy level above the valence band. For voltages above V_{th} , J is proportional to V^n with $n > 6$, implying that the large current may be associated with field-enhanced electron injection and the release of trapped carriers due to the reabsorption of electroluminescence. E_{act} for $V > V_{th}$ tends to increase with increasing V . This change in E_{act} may be attributed to the effect of the reabsorption of electroluminescence. The transition occurs when sufficient space charge builds to modify the potential barrier at the cathode to enable electron tunneling from it.

For insulating materials with a large band gap and a high ionic conductivity, the ionic space charge may be sufficient to turn on the transition from single-injection to double-injection conduction. The electrical contacts between a metal and an insulator can be assumed to be neutral contacts. The so-called *neutral contact* means that the regions adjacent to the contact on both sides are neutral electri-

cally. This implies that there is no space charge and no band bending present within the insulator in the absence of electric fields, so both the conduction band and the valence band edges are flat right up to the interface. A condition like this is sometimes referred to as the *flat-band condition*.

To satisfy this condition, the work function of the metal ϕ_m must be equal to the work function of the insulator ϕ . In wide-bandgap insulators, a small number of charge carriers always transfers between the metal and the insulator in order to bring the Fermi level of the metal to align with the Fermi level of the insulator. However, if the trapped space charge in the insulator due to this transfer of charge carriers is too small to cause significant band bending, then we can consider the contact neutral (see Types of Electrical Contacts in Chapter 6). For most insulating materials, such as polymers, the energy band gap is of the order of 9 eV (e.g., polyethylene) and most metals have work functions less than 4 eV. Thus, the neutral contact means that the potential barrier height ϕ_{Bn} for electron injection from the cathode is lower than that ϕ_{Bp} for hole injection from the anode.

When a steady DC voltage is applied across a metal-insulator-metal (MIM) system, electrons will be injected from the cathode to the insulator by three possible processes: Schottky emission, thermally assisted tunneling, and tunneling via traps. If the applied voltage is sufficiently high, direct tunneling from the Fermi level of the metal to the conduction band, based on the Fowler–Nordheim type tunneling injection, is possible. The probability of tunneling depends on both the population of electrons available for tunneling and the barrier height. The population of available electrons is highest at the Fermi level of the metal and decreases exponentially at levels higher than the Fermi level, but the barrier height for tunneling decreases at levels higher than the Fermi level. So, there is a tradeoff between these two factors governing tunneling processes. Since insulating polymers contain a large quantity of various traps and ionic impurities, carriers injected from the electrical contacts will quickly be trapped to form trapped space charges, as shown in Figure 7-38(a).

The positively charged cations moving toward the cathode and the negatively charged anions moving toward the anode under an applied electric field will create hetero-space charges near the electrodes. If the ions' charges are not neutralized at the electrodes, they will accumulate near the electrodes, producing an internal potential and hence an internal field opposite to the applied voltage, as shown in Figure 7-38(b). The positive ions tend to neutralize the injected trapped electrons near the cathode, while the negative ions tend to enhance the field toward the anode, thus modifying the potential barrier near the interface and creating a chance for the holes to tunnel from the metal to the valence band of the polymer, as shown in Figure 7-38(c). Obviously, the higher the applied field, the more holes will tunnel, thus switching on the double-injection process.¹⁶⁷

The ionic conduction process is very slow because the movement of an ion involves the transport of a mass and the activation energy both for the ion to surmount a potential barrier and for the creation of a neighboring vacancy for the ion to move into.¹¹ It can be imagined that to create hole tunneling, the applied field must be sufficiently high and the stressing time must be sufficiently long for the formation of a high concentration of hetero-space charges. Several investigators have observed hole injection from a metallic electrode into polyethylene after the material has been subjected to electrical stressing for a prolonged period of time, and hence, double injection.^{168,169}

7.8 Current Transient Phenomena

So far, we have considered only steady-state injection currents. Prior to the attainment of the steady state, a variety of current transient phenomena may occur immediately after the application of an electric field. Such phenomena may provide useful information about transport, trapping, recombination, and photogeneration processes in solids. On the basis of the total charge Q injected into the solid, the current transient phenomena can be classified into three types:

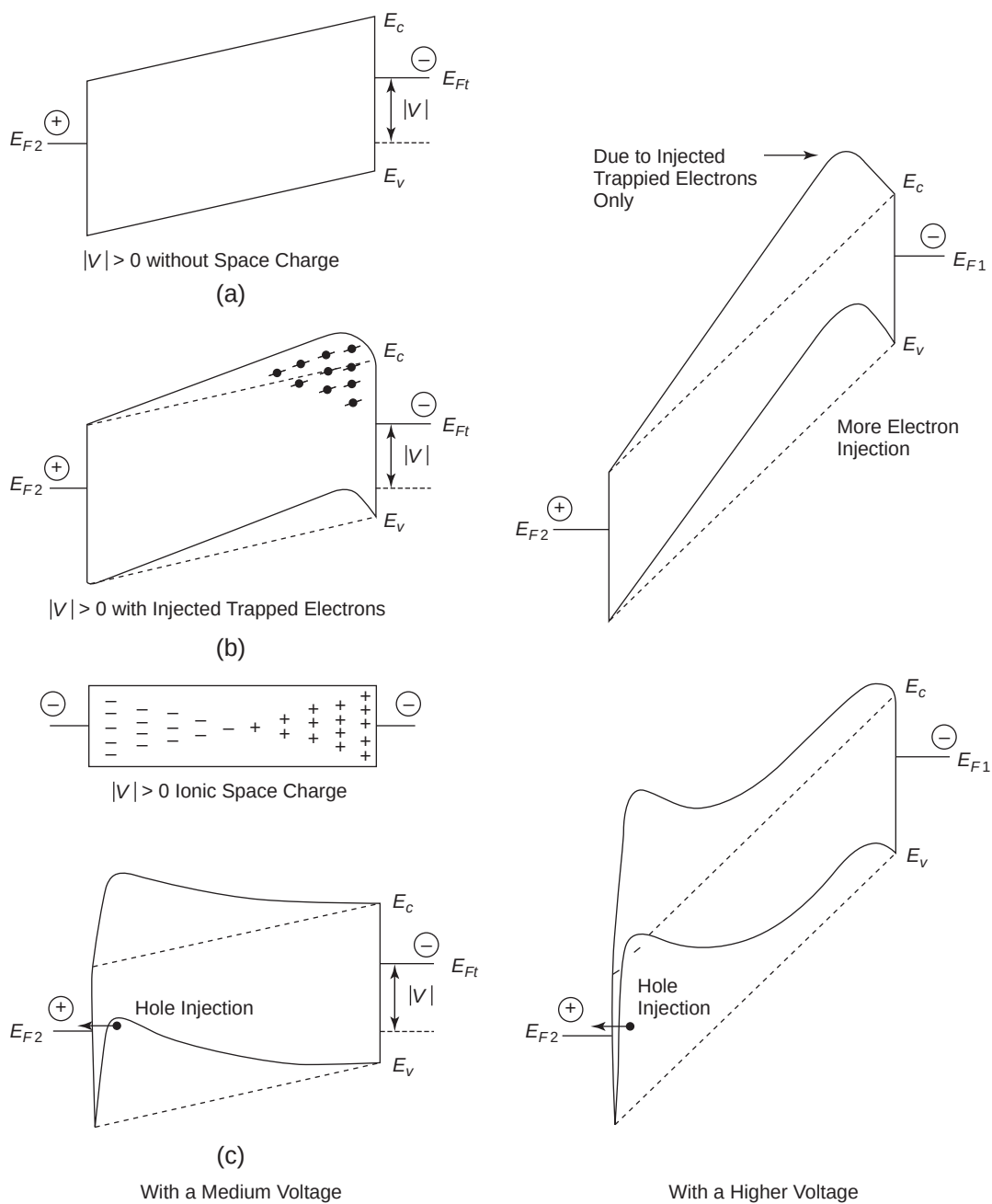


Figure 7-38 Schematic diagrams illustrating the effects of the combined space charges formed (a) by injected trapped electrons and (b) by positive and negative ions on the band bending and hence (c) the potential distribution resulting in hole tunneling from the anode to the insulating polymer.

The space-charge free (SCF) transient, for which Q is much smaller than required to influence significantly the electric field near the injecting electrode, $Q \ll 2\epsilon F$

The space-charge limited (SCL) transient, for which Q is large enough to create a charge reservoir near the injecting electrode, $Q \gg 2\epsilon F$

The space-charge perturbed (SCP) transient, for which the values of Q are intermediate between SCF and SCL

Current transients can be induced by the application of a voltage pulse, electron beam bombardment, corona discharge, light pulse, etc. This section briefly reviews the theory and applications of current transient phenomena resulting from one-carrier (single) planar injection.

Suppose that a step-function voltage of the form

$$\begin{aligned} V(t) &= 0 \quad \text{when } t < 0 \\ V(t) &= V \quad \text{when } t \geq 0 \end{aligned}$$

is applied to a solid specimen provided with an electron-injecting contact at $x = 0$ as cathode. The current transient begins at $t = 0$, and its behavior is governed by the current flow equation

$$\begin{aligned} J(x,t) &= J_c(x,t) + \epsilon \frac{\partial F(x,t)}{\partial t} \\ &= q\mu_n n(x,t) F(x,t) \\ &\quad - qD_n \frac{\partial n(x,t)}{\partial x} + \frac{\partial F(x,t)}{\partial t} \end{aligned} \quad (7-349)$$

Poisson's equation

$$\frac{\partial F(x,t)}{\partial x} = \frac{q}{\epsilon} [n(x,t) - n_o + n_t(x,t) - n_{to}] \quad (7-350)$$

the rate equation (see Equation 7-206)

$$\begin{aligned} \frac{\partial n_t(x,t)}{\partial t} &= C_n \{n(x,t)[N_n - n_t(x,t)] \\ &\quad - N_c \exp[-(E_c - E_t)/kT] \bullet n_t(x,t)\} \\ &= C_n \{n(x,t)[N_n - n_t(x,t)] - n_t n_t(x,t)\} \\ &= \tau_n^{-1} [n(x,t) - \theta_o n_t(x,t)] \end{aligned} \quad (7-351)$$

and the continuity equation

$$\frac{\partial J(x,t)}{\partial x} = \frac{\partial J_c(x,t)}{\partial x} + \epsilon \frac{\partial}{\partial x} \left[\frac{\partial F(x,t)}{\partial t} \right] = 0$$

or

$$-\frac{1}{q} \frac{\partial J_c(x,t)}{\partial x} = \frac{\partial n(x,t)}{\partial t} + \frac{\partial n_t(x,t)}{\partial t} \quad (7-352)$$

where n_o and n_{to} are, respectively, the free-electron and trapped-electron densities under thermal equilibrium (assumed to be uniformly distributed throughout the specimen), C_n is the electron capture coefficient, and N_n is the density of electron traps in a discrete set of localized states located at E_t in the band gap. τ_n is the trapping time, defined as the mean free time of a mobile free carrier before it is trapped, which is

$$\tau_n = [C_n(N_n - n_{to})]^{-1} \quad (7-353)$$

and n_1 and θ_o are defined by

$$n_1 = N_c \exp[-(E_c - E_t)/kT] \quad (7-354)$$

$$\theta_o = n_o/n_{to} = n_1/(N_n - n_{to}) \quad (7-355)$$

and $(N_n - n_{to}) \gg n_t(x,t) - n_{to}$. Equation 7-351 may be extended for any form of trap distribution, provided that the total trapping rate is larger than the total detrapping rate.

It is obvious that an analytical solution of Equations 7-349 through 7-352 cannot be obtained without making some simplifying assumptions. Assuming that $n \gg n_o$ and $n_t \gg n_{to}$, integration of Equation 7-349 yields

$$\begin{aligned} J(t) &= \frac{\epsilon \mu_n}{2d} [F_a^2(t) - F_c^2(t)] \\ &\quad - \frac{q\mu_n}{d} \int_0^d n_t(x,t) F(x,t) dx \\ &\quad - \frac{\epsilon D_n}{d} \left\{ \left[\frac{\partial F(x,t)}{\partial x} \right]_{x=d} - \left[\frac{\partial F(x,t)}{\partial x} \right]_{x=0} \right\} \\ &\quad + \frac{qD_n}{d} \{ [n_t(x,t)]_{x=d} - [n_t(x,t)]_{x=0} \} \end{aligned} \quad (7-356)$$

since $(\epsilon/d) \frac{\partial}{\partial t} \int_0^d F(x,t) dx = (\epsilon/d) \frac{\partial V}{\partial t} = 0$, by assumption, where $F_a(t) = F(d,t)$ and $F_c(t) = F(0,t)$ are, respectively, the fields at the anode and the cathode. Integration of Equation 7-350 yields the relation between $F_c(t)$ and $F_a(t)$ in

terms of the total charge $Q(t)$ per unit area in the specimen at time t .

$$F_a(t) = F_c(t) + Q(t)/\varepsilon \quad (7-357)$$

In the following sections, we shall consider three major current transient phenomena.

7.8.1 Space-Charge Free (SCF) Transient

In this section we shall describe two cases, one is the trap-free case and the other the effect of traps.

Case 1: In the Absence of Traps and Diffusion

For this case, Equation 7-356 becomes

$$\begin{aligned} J(t) &= \frac{\varepsilon \mu_n}{2d} [F_a^2(t) - F_c^2(t)] \\ &= \frac{\mu_n Q(t)}{2d} [2F_c(t) - Q(t)/\varepsilon] \\ &= \frac{\mu_n Q(t)}{2d} [2F_c(t) + Q(t)/\varepsilon] \end{aligned} \quad (7-358)$$

and $Q \ll 2\varepsilon F_c$ by assumption. This implies that Q is smaller than the maximum charge that the specimen can store (the maximum charge = $V\varepsilon/d$) and that this case corresponds to the case of a blocking-contact cathode and a weak light pulse to produce the injected charge. From Equation 7-357, we have

$$F_a \approx F_c \approx F_{av} = \frac{V}{d} \quad (7-359)$$

Thus, Equation 7-358 becomes

$$\begin{aligned} J &= \frac{\mu_n Q F_c}{d} = \frac{\mu_n Q V}{d^2} = \text{constant} \\ &= \frac{Q}{t_{io}} \end{aligned} \quad (7-360)$$

where t_{io} is the SCF electron transit time.

The simplest example is the time-of-flight technique for measuring carrier mobilities. This technique is similar to that employed for the determination of the mobility of minority carriers in semiconductors.^{170,171} The basic experimental arrangement for the time-of-flight technique is shown in Figure 7-39(a). The specimen is subjected to a constant applied voltage V between two noninjecting electrodes in the dark. At $t = 0$, a light pulse (or electron beam)

is applied to the cathode to produce a sheet of charge Q , which drifts from $x = 0$ at $t = 0$ to $x = d$ at $t = t_{io}$. The so-called *sheet of charge* has a finite thickness due to the duration of the light pulse (or electron beam). As soon as a sheet of electrons is introduced in the narrow region near the cathode, a current will start to flow and can be recorded by a measuring circuit. The slow rise of the current transient is caused by the time constant of the measuring circuit; the long tail is due to the spreading of the charge carrier sheet by diffusion, as shown in Figure 7-39(b). The times $t = 0, t_1, t_2, t_3$ correspond to the position of the charge carrier sheet along the x direction in the specimen, as shown in Figure 7-39(c). This technique has been used to measure carrier mobilities in organic solids¹⁷²⁻¹⁷⁴ and in inorganic solids.¹⁷⁵⁻¹⁷⁹

In general, the concentration of imperfection centers is higher near a surface. These, coupled with other possible surface states, will considerably shorten carrier lifetime in the generation region. It is not easy to find a light source that can produce an intense flash of much shorter duration than the carrier transit time and yet generate sufficient electron-hole pairs less than $1\mu\text{m}$ from the illuminated surface. This problem becomes more difficult for wide-bandgap materials. Spear^{176,177} has suggested using electron-beam excitation instead of optical excitation for the following advantages:

- The intensity of the beam is high and its duration can be made sufficiently short for the generation of sufficient carriers.
- The depth of the generation region below the top surface can be varied to any desired value simply by adjusting the accelerating potential.
- It does not involve absorption in the electrode.
- For wide-bandgap materials for which optical excitation is not feasible, electron beam or α particle excitation is the only means of generating electron-hole pairs.

Case 2: In the Absence of Diffusion but with Traps

The presence of traps may limit the applicability of the transient method for drift mobility

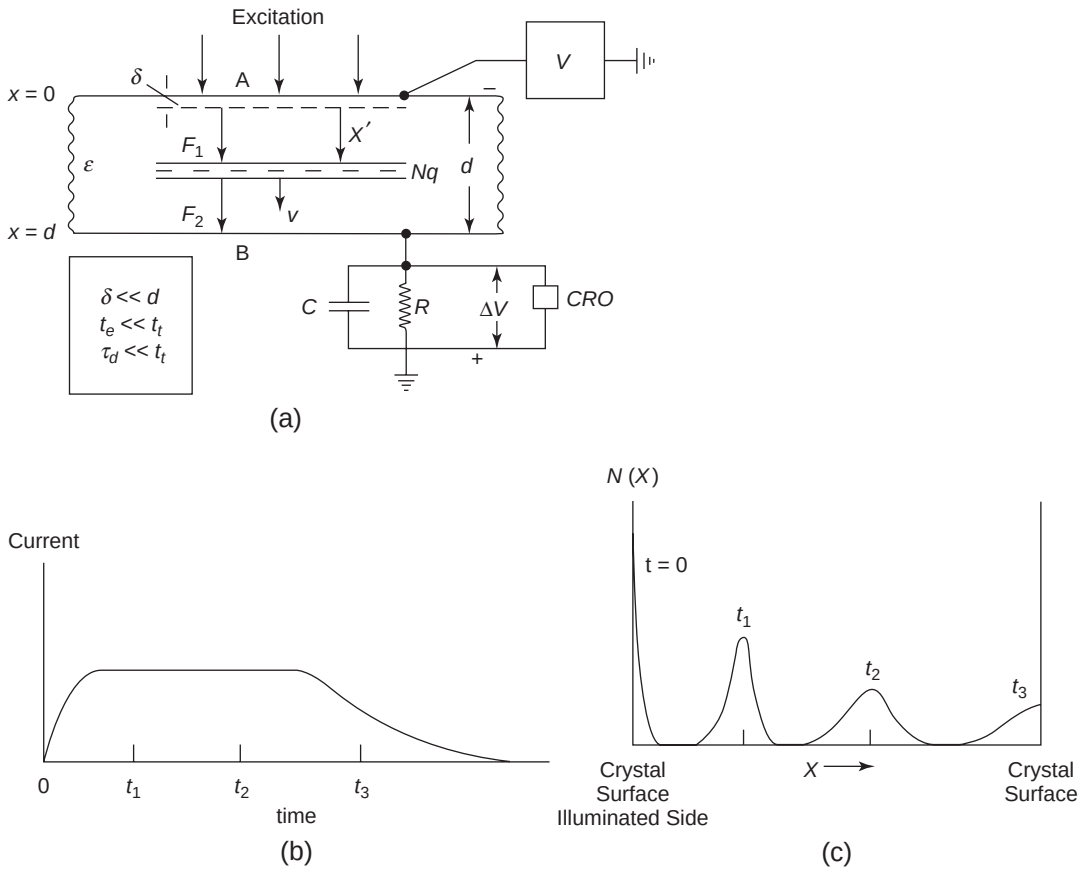


Figure 7-39 Schematic diagrams showing (a) the experimental arrangement for the time-of-flight technique, in which A and B are the metallic electrodes with electrode A semitransparent, V is the applied voltage with negative polarity at A with respect to ground, t_e is the duration of the excitation pulse, t_t is the transit time, τ_d is the dielectric relaxation time, Nq is the narrow charge sheet due to N carriers generated near the cathode surface, F_1 and F_2 are, respectively, the electric field behind and in front of the charge sheet; (b) the current transient as a function of time; and (c) the position of the charge sheet along the axis in the specimen corresponding to $t = 0$, t_1 , t_2 , and t_3 .

measurements. Suppose that μ_n is the mobility of the free electron carriers whose concentration is n . But the trapped carrier concentration is n_t , so the measured drift mobility μ' is given by

$$n\mu_n = (n + n_t)\mu'_n \quad (7-361)$$

If the traps are shallow electron traps located at a single discrete level E_t , then in thermal equilibrium we have

$$\frac{n_t}{n} = \frac{N_t}{N_c} \exp[(E_c - E_t)/kT] \quad (7-362)$$

From Equations 7-361 and 7-362, we obtain

$$\mu'_n = \mu_n \left\{ 1 + \frac{N_t}{N_c} \exp[(E_c - E_t)/kT] \right\}^{-1} \quad (7-363)$$

At very high temperatures, $\mu'_n \approx \mu_n$. At low temperatures

$$\mu'_n \approx \frac{N_c \mu_n}{N_t} \exp[-(E_c - E_t)/kT] \quad (7-364)$$

Similar expressions apply to hole carriers, except that n , n_t , μ_n , μ'_n , and $(E_c - E_t)$ are replaced with p , p_t , μ_p , μ'_p , and $(E_t - E_v)$, respectively.

For shallow traps, the average time for a trapped carrier to stay in the trap before thermal release (the trap release time) $\tau_{tr} \ll t_t$ (transit time), and the lifetime $\tau \ll t_t$ (depending on the specimen thickness and applied field). However, for deep traps $\tau_{tr} \gg t_t$. Thus, the effect of traps on the measurements of drift mobilities depends on whether $\tau < t_t$ or $\tau > t_t$. It is obvious that the presence of traps has little effect if $\tau > t_t$ and makes the measurements impossible if $\tau < t_t$ for deep traps. With a short lifetime $\tau \approx t_t$ for deep traps, the concentration of free carriers will gradually decrease during the transit from one electrode to the other. Typical transient responses for this case are shown in Figure 7-40(c) and (d).

7.8.2 Space-Charge Limited (SCL) Transient

Similar to the previous section, this section describes two cases. One deals with the trap-free case and the other explores the effects of traps.

Case 1: In the Absence of Traps and Diffusion

In this case, the injected charge Q is large enough to create an electron charge carrier

reservoir at the cathode (or at the anode for holes), due either to intense optical excitation at the cathode or to strong carrier injection from the ohmic-contact cathode. Since $Q \gg 2\epsilon F_c$ in this case, we have

$$F_c(t) = 0 \quad (7-365)$$

and Equation 7-356 becomes

$$J(t) = \frac{\epsilon \mu_n}{2d} F_a^2(t) \quad (7-366)$$

Suppose that t_1 is the time required for the leading front of the injected charge to arrive at the anode (the transit time of the leading front). Then, during the period $0 \leq t \leq t_1$, the total current is simply the displacement current at the anode because the conduction current J_c , which depends on the charge collected at the anode, is practically equal to zero. Thus,

$$J(t) = \epsilon \frac{\partial F_a(t)}{\partial t} \quad (7-367)$$

The combination of Equations 7-366 and 7-367 yields

$$\frac{dF_a}{dt} = \frac{\mu_n}{2d} F_a^2 \quad (7-368)$$

Using the boundary condition $F(x, 0) = V/d$, since there is no injected charge at $t = 0$, and the following relations

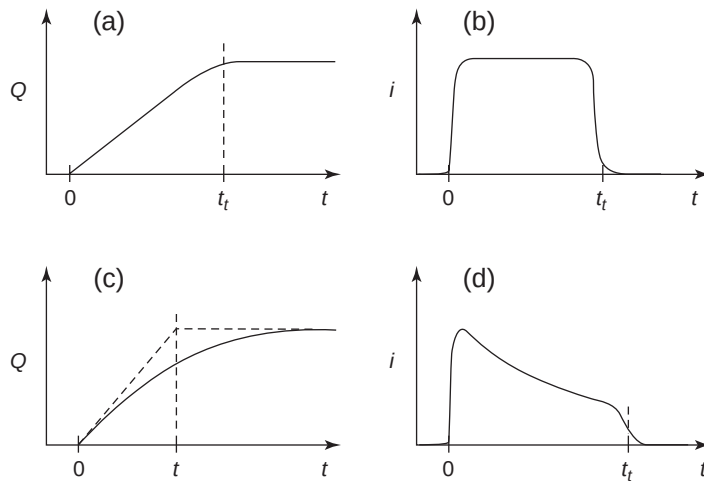


Figure 7-40 Schematic diagrams illustrating some typical pulse shapes in the time-of-flight measurements and the effects of traps for carriers generated close to the cathode: (a) integrated signal (i.e., Q) for $CR \gg t_t$ [see Figure 7-39(a)], (b) current pulse for $CR \ll t_t$. In both (a) and (b) deep traps are absent; (c) and (d) show the corresponding pulse shapes for deep traps.

$$\left. \begin{aligned} t_{io} &= d^2/\mu_n V \\ Q_c &= \frac{\epsilon V}{d} \end{aligned} \right\} \quad (7-369)$$

the solution of Equation 7-368 yields

$$\epsilon F_a(t) = Q(t) = \frac{Q_c}{1 - (t/2t_{io})} \quad (7-370)$$

Hence, we obtain

$$J(t) = J_o/[1 - (t/2t_{io})]^2 \quad (7-371)$$

where

$$J_o = J(o) = \frac{\epsilon \mu_n V^2}{2d^3} \quad (7-372)$$

Since the field at the leading front of the charge at the time $t < t_1$ can be considered the same as the field at the anode at time t , the value of t_1 can be determined by the following relation

$$\begin{aligned} d &= \int_0^{t_1} \mu_n F_a(t) dt \\ &= (2t_{io} \mu_n V/d) \ln[1 - (t_1/2t_{io})]^{-1} \end{aligned} \quad (7-373)$$

with the aid of Equation 7-370. From Equation 7-369, this gives

$$t_1 = 2(1 - e^{-1/2})t_{io} \approx 0.786t_{io} \quad (7-374)$$

Letting $J_1 = J(t_1)$ and $J_\infty = J(\infty) = 9\epsilon\mu_n V^2/8d^3$ = the trap-free steady-state SCL current, Equations 7-371 and 7-374 yield

$$\left. \begin{aligned} J_1/J_o &= e = 2.72 \\ J_1/J_\infty &= 4e/9 = 1.21 \end{aligned} \right\} \quad (7-375)$$

The curve for $R = \tau_n/t_{io} = \infty$ shown in Figure 7-41 corresponds to the case in which trapping is absent. It can be seen that $J_o/J_\infty = 0.44$ occurs at $t/t_{io} = 0$ and $J_1/J_\infty = 1.21$ occurs at $t/t_{io} = 0.786$. The time rate of change of the current at $t = t_1^-$ just before it reaches the peak is

$$\left. \frac{dJ}{dt} \right|_{t=t_1^-} = \frac{e^{1/2} J_1}{t_{io}} = 1.65 J_1/t_{io} \quad (7-376)$$

and at $t = t_1^+$ just after it passes the peak is

$$\left. \frac{dJ}{dt} \right|_{t=t_1^+} = \frac{1 - 2e^{-1/2}}{1 - e^{-1/2}} \left(\left. \frac{dJ}{dt} \right|_{t=t_1^-} \right) = 0.90 J_1/t_{io} \quad (7-377)$$

At $t = t_1$ when the leading front reaches the anode ($x = d$), the specimen contains the maximum amount of space charge and $J(t)$

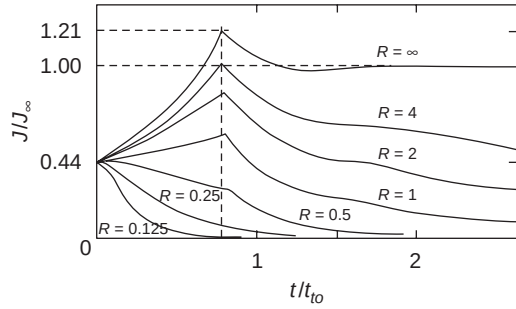


Figure 7-41 Theoretical SCL current transients plotted in current density as a function of time for various trapping times τ_n for insulating crystals in which $\theta_o = n_o/n_{io} = 0$, $R = \tau_n/t_{io}$, where t_{io} is the space charge free transit time, J_∞ is the steady-state current without trapping, and $R = \infty$ corresponds to the absence of trapping.

attains its peak value. Therefore, $J(t)$ decays toward J_∞ because for $t > t_1$ more space charge leaves the specimen than is injected into it, and the space charge distribution at $t = t_1$ relaxes toward its steady-state configuration with the relaxation time inversely proportional to V and of the order t_1 . Since at $t = t_1$, the specimen contains more space charge than it can hold under steady-state conditions. An undershoot occurs in the decay due to the inhibition of charge injection in the vicinity of t_1 , because of the overshoot of charge already present at the time. This may cause the oscillations of the decaying current.¹⁸⁰⁻¹⁸²

Suppose that a light flash of much shorter duration than the transit time produces a very thin sheet (much thinner than the specimen thickness) of charge carriers at $x = 0$ and $t = 0$. Then the field is V/d in front of the sheet and zero behind it. Thus, the width of the sheet as a function of time is

$$S(t) = \frac{V\mu_n t}{d} \quad (7-378)$$

and the current density of such sheet current is given by

$$J(t) = J_o \exp(t/t_{io})/2 \quad (7-379)$$

where J_o is the same as J_o given by Equation 7-372.⁵³ This equation is valid as long as the leading front has not yet reached the anode. The time t_2 required for the leading front to reach the anode is

$$t_2 = 0.792t_{i0} \quad (7-380)$$

The current transient has a cusp at $t = t_2$. The time rate of change of the current at $t = t_2^-$ just before it reaches the peak is

$$\left. \frac{dJ}{dt} \right|_{t=t_2^-} = \frac{J(t_2)}{t_{i0}} \quad (7-381)$$

and that at $t = t_2^+$ just after it passes the peak is

$$\left. \frac{dJ}{dt} \right|_{t=t_2^+} = -2.3J(t_2)/t_{i0} \quad (7-382)$$

Theoretically, another transit time is required for all the space charge to leave the specimen. However, the initial charge distribution has a negligible effect on the general shape of the transient, provided that the change is confined in a sheet very thin compared to the specimen thickness, and that the width of the charge sheet increases with time, becoming almost equal to the specimen thickness when the leading front arrives at the collecting electrode (the anode in this case). Therefore, the transients of SCL current and SCL sheet current are similar.⁵³

For details about the derivation of the above equations, see references.^{53,180,184}

Case 2: In the Absence of Diffusion but with Traps

The SCL transient in a solid with traps has been studied by several investigators.^{180,185-187} Many and Rakavy have numerically computed $J(t)$ as a function of time for a solid containing traps, and their results are shown in Figure 7-41. The smaller the value of $R = \tau_n/t_{i0}$, the faster the current decays and the higher the value of t_1 . This implies that the concentration of trapped carriers near $x = 0$ increases with decreasing value of R , giving rise to a correspondingly lower field under which the leading front moves. A current cusp always exists in the presence of trapping, provided that the trapping is not too fast ($R = \tau_n/t_{i0} > 0.5$). Also, the peak of the current occurs at about the same time t_1 , and the initial current $J_0 = J(t = 0)$ is independent of R (i.e., independent of the trapping time). For details about this analysis, see references.^{186,187}

7.8.3 Space-Charge Perturbed (SCP) Transient

In most practical situations, the ohmic contact is not perfect and the carrier reservoir is not large enough to maintain zero field at the injecting electrode, whether part of the time or throughout the duration of the current. In this section, we shall consider two cases in which the current transient which is partly space-charge controlled and partly electrode limited.

Carrier Generation by a Strong Light Pulse

The intensity of the light pulse is assumed to be sufficiently great that initially zero field is established at the electrode illuminated by the light pulse, and the current initially follows an SCL form.^{184,188} After the termination of the light pulse, however, the carrier reservoir is gradually diminished by recombination and $F(0,t) = 0$ for times up to t_A , where t_A is the duration of the light pulse and $t_A \leq t_1$.

Assuming that the carrier lifetime $\tau < t_A$ so the carrier reservoir at the cathode collapses completely at $t = t_A$, and that $F(0,t) = 0$ for $t \leq t_A$ and $F(0,t) \neq 0$ for $t > t_A$, then for the time interval $t < t_A$, the current as a function of t is given by Equations 7-371 and 7-372. For the time interval $t > t_A$ Weisz et al.¹⁸⁸ have derived the expression for $J(t)$ in a case with shallow traps but excluding diffusion, which is

$$J(t) = \frac{\epsilon\mu_n V^2}{2d^3} [1 - (t_A/2t_{i0})]^2 \times \exp[(t - t_A)/(t_{i0} - t_1/2)] \quad (7-383)$$

for $t_A \leq t \leq t_1$

and

$$J = -(M/t)\{\beta(t) + \beta^2(t)[\beta(t)/y^2 - 1/y^2]\} \quad (7-384)$$

for $t_1 < t < t_2$

where

$$\left. \begin{aligned} \beta(t) &= y[J_1(y) + cH_1^{(1)}(y)]/[J_0(y) + cH_0^{(1)}(y)] \\ y &= \frac{2\mu_n Mt}{\epsilon d} \\ M &= -\frac{\epsilon V}{d} \quad \text{for } t_A = 0 \end{aligned} \right\} \quad (7-385)$$

and $J_0(y)$ and $J_1(y)$ are the zero and first-order Bessel functions, respectively; $H_0^{(1)}(y)$ and $H_1^{(1)}(y)$ are the zero and first-order Hankel functions of the first kind, respectively; and c is the constant of integration; t_1 and t_2 are the transit times for the leading front and the tail of the injected charge to arrive at the anode, respectively. For $t_A \geq 0$, M and c can be determined from the following equations

$$\left. \begin{aligned} t \left(\frac{dQ}{dt} \right) + Q &= \frac{\mu_n}{2\epsilon d Q^2} + M \\ Q(t_A) &= \frac{\epsilon V}{d} [1 - (t_A/2t_{io})]^{-1} \\ t \frac{dQ}{dt} &= -\epsilon F(d, t) \\ F(d, t) &= (V/2d) [1 - (t_A/2t_{io})]^{-1} \\ &\quad \times [1 + \exp[(t - t_A)/(t_{io} - t_A/2)]] \\ tQ &= \frac{\epsilon dy}{\mu_n} \{ [J_1(y) + cH_1^{(1)}(y)] / [J_0(y) + cH_0^{(1)}(y)] \} \end{aligned} \right\} \quad (7-386)$$

Weisz et al.¹⁸⁸ have computed $J(t)$ as a function of t for $t_A = 0, 0.5t_1, t_1$ and ∞ , and the results are shown in Figure 7-42. In the top four curves, the current is initially SCL, but for $t > t_A$, the

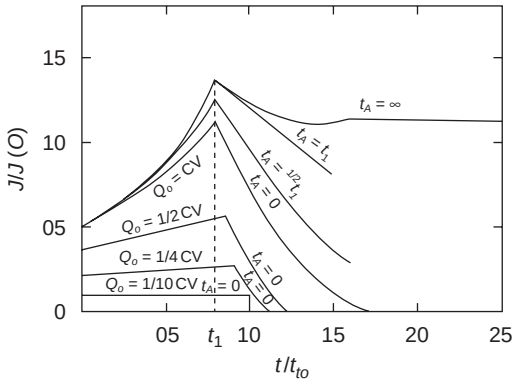


Figure 7-42 Transient current as a function of time due to injected charge Q_0 in a dielectric specimen. The top four curves are associated with strong light excitation and the bottom three curves with weak light excitation. The physical constants used for computing these curves are $\mu_n = 1 \text{ cm}^2/\text{V-sec}$, absorption coefficient: (light penetration length) $^{-1} = 10^3 \text{ cm}^{-1}$, $\epsilon/\epsilon_0 = 2$, and $d = 0.1 \text{ cm}$.

current changes more rapidly with time than for the SCL transient. This prediction agrees well with the experimental results on anthracene^{180,183} and on iodine crystals.¹⁸⁹

Carrier Generation by a Weak Light Pulse

The light intensity is assumed to be so weak or the pulse duration so short compared to the transit time that the charge Q_0 injected by the light pulse is smaller than $CV = \epsilon V/d$. In this case, $F(0, t) \neq 0$ for $t < t_A$, but $F(0, t) = 0$ for $t > t_A$. Weisz et al.¹⁸⁸ have analyzed this case, and their analysis yields

$$\begin{aligned} J(t) &= (\mu_n Q_0 / d) (V/d - Q_0 / 2\epsilon) \\ &\quad \times \exp(\mu_n Q_0 t / \epsilon d) \\ &\text{for } t \leq t_1 \end{aligned} \quad (7-387)$$

If $Q_0 \ll CV$, then Equation 7-387 becomes

$$J(t) = \mu_n V Q_0 / d^2 \quad \text{for } t \leq t_1 \quad (7-388)$$

For $t_1 < t < t_2$, the value of $J(t)$ can be calculated from Equation 7-384. They have also computed $J(t)$ for $Q_0 < CV$, and their results are shown in Figure 7-42 (the bottom three curves). From Equation 7-387, the initial current at $t = 0$ for $Q_0 < CV$ is

$$J(0) = (\mu_n Q_0 V / d^2) - (\mu_n Q_0^2 / 2\epsilon d) \quad (7-389)$$

Theoretically, $J(0)$ is directly proportional to V , but experimentally, it is proportional to V^2 over a considerable range of V , indicating that Q_0 depends on V even when the intensity and duration of the light pulse are constant. Of course, the higher the applied voltage, the more carriers of the appropriate type will be swept into the bulk of the specimen. In the illuminated electrode region, carriers of both types are present and encounter carrier recombination as well as carrier trapping. In the bulk, only carrier trapping is possible. Thus, the lifetime of a carrier in the electrode region can be much smaller than in the bulk. This implies that Q_0 increases with increasing V . Weisz et al.¹⁸⁸ have estimated the voltage dependence of Q_0 , which is expressed as

$$Q_0(V) \approx [2K/(1+K)]CV \quad (7-390)$$

and $J(0)$ becomes

$$J(o) \approx [4K/(1+K)^2](\epsilon\mu_n V^2/2d^3) \quad (7-391)$$

where

$$K = \rho(o)\mu_n\tau_n \exp(-2)/2\epsilon \quad (7-392)$$

$\rho(o)$ is the charge density produced by a light flash at $t = o$ and at $x = o$. Equations 7-390 and 7-391 are valid only for $K \leq 1$. The intensity of light at which the initial current undergoes a transition from the electrode-limited to SCL regime is obtained by the condition $K = 1$, which is

$$\rho(o)\exp(-2) = 2\epsilon/\mu_n\tau_n \quad (7-393)$$

For details of SCP transients, see references.^{184,188,190-192}

Tabak and Scharfe¹⁹² have observed a transition from emission- (or electrode-) limited to SCL current transient in amorphous selenium films. Their results are shown in Figure 7-43. If the light is absorbed close to the surface, only the carriers with the same polarity as the illuminated surface will move across the specimen. If $Q_o < CV$, a fraction of the total charge during carrier transport will be trapped. This fraction is determined by the ratio of the average distance traveled before trapping to specimen thickness. For a constant applied voltage, the accumulation of space charge will cause the electric field at the illuminated electrode to approach zero and the current to become space-charge limited. Eventually, at $t = \infty$, F varies with $x^{1/2}$. The dividing line between emission-limited and SCL regimes is labeled $T_{1/2}$ in Figure 7-43. $T_{1/2}$ is defined as the time required for the current to decay to one-half of its initial value; it is also the time taken to trap one CV of charge.^{75,192}

Finally, note that, for a rigorous treatment of current transients, the effects of image force and charge exchange in traps on permittivity and carrier mobility should be considered.

7.9 Experimental Methodology and Characterization

To characterize the electrical properties of materials, experimental work usually involves several standard techniques and instruments,

such as x-ray diffraction, Auger electron spectroscopy, infrared absorption spectroscopy, Raman spectroscopy, electron paramagnetic resonance (EPR) spectroscopy, ellipsometry, I - V and C - V characteristics, etc. Information about these standard techniques and instruments is available elsewhere and therefore not included in this section. The commonly used thermally stimulated discharge current (TSDC) techniques were discussed in Relaxation Times of Dipoles and the Thermally Stimulated Discharge Current (TSDC) Technique in Chapter 5. Here, we shall describe only a few simple and powerful methods used to determine the type of carrier species and the trap parameters that control the electrical properties of dielectric materials.

7.9.1 Simultaneous Measurements of Charging Current and Photocurrent Transients

By measuring simultaneously the charging current transient and the photocurrent transient superimposed on it, we can determine the carrier species responsible for both the dark conduction and the photoconduction currents. The experimental arrangement for this measurement is shown in Figure 7-5, and Section 7.3 discussed the experimental results in some detail.

7.9.2 Alternating Measurements of the I - V and C - V Characteristics

By alternately measuring the I - V and C - V characteristics using linear ramp voltages, we can determine the carrier species and trap parameters. The experimental arrangement for these measurements is shown in Figure 7-44(a) and the form of the linear ramp voltage in Figure 7-44(b). A metal-insulating material-semiconductor (MIS) system and standard instruments for monitoring the current, voltage, and capacitance are used for these measurements. The capacitance is usually measured at a high frequency (1 MHz).

We will now describe briefly how this method can determine the following parameters:

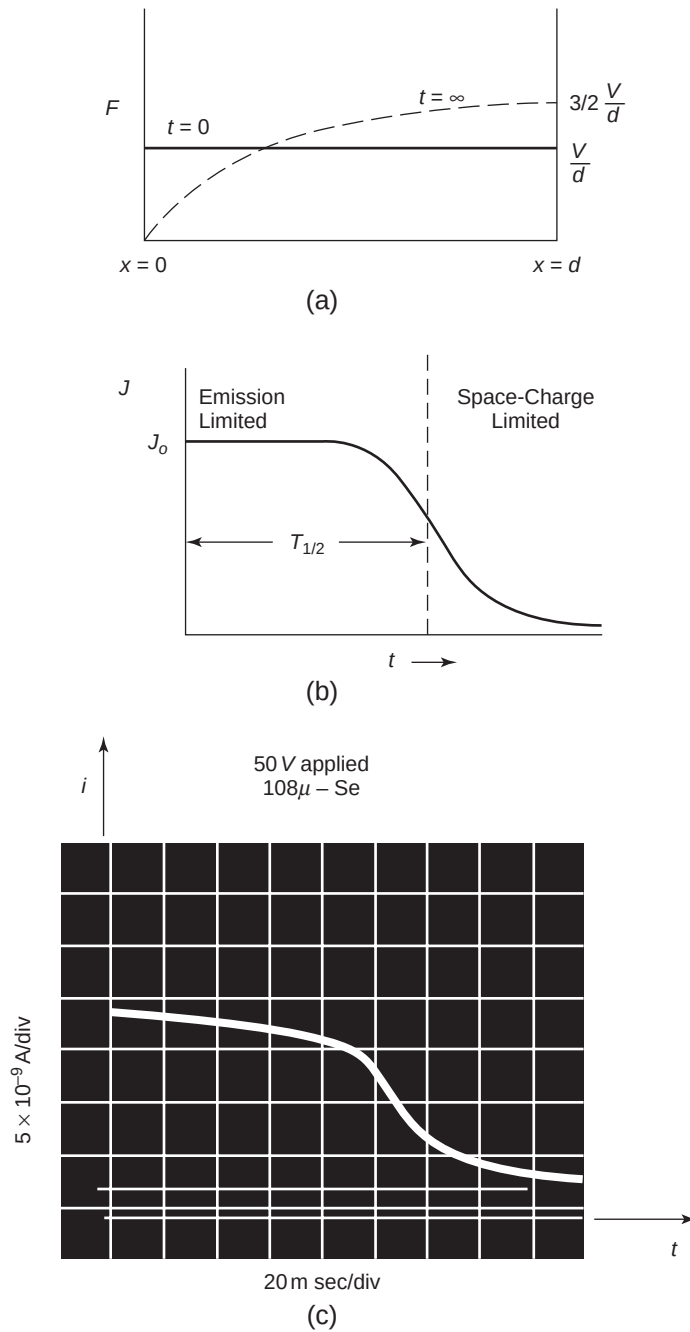


Figure 7-43 Transition from emission-limited to SCL current transients: (a) theoretical electric field distribution at $t = 0$ and at $t = \infty$, (b) theoretical J - t curve, and (c) experimental J - t curve for an amorphous selenium film with $d = 108 \mu\text{m}$ and the illuminated (4500 Å light) electrode positive so that holes are the injected carriers and $V = 50 \text{ V}$.

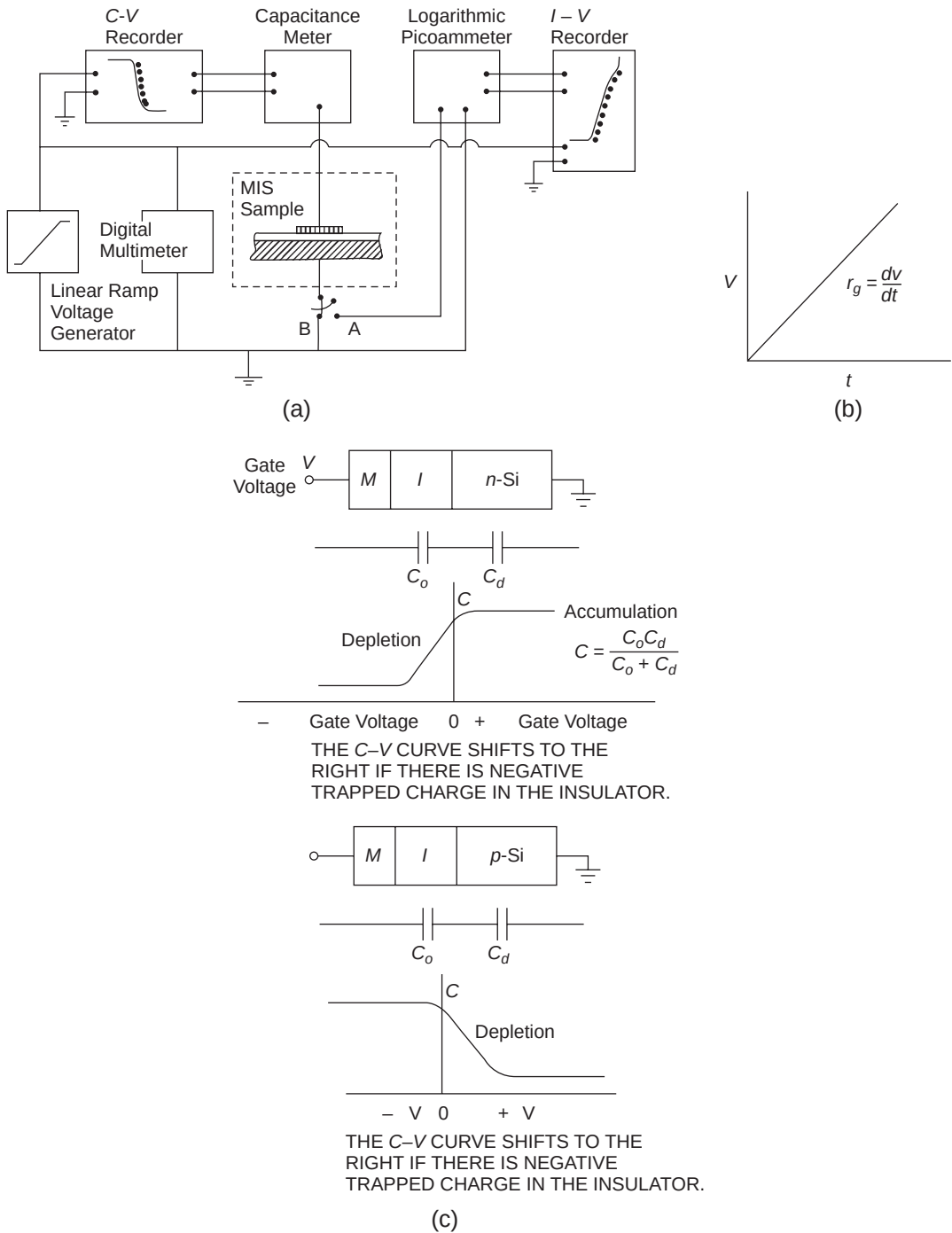


Figure 7-44 (a) Experimental arrangement for measuring the ramp current–voltage (I – V) and the capacitance–voltage (C – V) characteristics; switch at A for I – V and switch at B for C – V measurements; (b) the form of the ramp voltage, and (c) the shift of the C – V curve for MIS systems.

- By measuring the I - V characteristics at different ramp rates, the electron trapping rate can be determined on the basis of the shift of the I - V characteristics as a function of the ramp rate.
- To determine the type of carrier species responsible for high-field electrical conduction, we measure the C - V characteristics right after the MIS specimen has been electrically stressed. Figure 7-44(c) shows that if an n-type semiconductor (e.g., n -Si) is used for the MIS system, the C - V curve shifts to the right if the space charge in the insulating material is negative (due to trapped electrons) and to the left if the space charge is positive (due to trapped holes). When a p-type semiconductor (e.g., p -Si) is used for the MIS system, the direction of the shift of the C - V curve is the same as for an n-type semiconductor. For the C - V measurements, the voltage sweep must be limited to a swing range lower than the threshold voltage for carrier injection so that during the sweeping, there is no carrier injection. In general, we first measure the C - V characteristics of a virgin MIS specimen (called the *virgin C-V curve*). After that, the specimen is stressed with a first cycle of linear ramp voltage up to a predetermined peak voltage. After the first cycle of stressing, the specimen is immediately connected to the circuit for the C - V measurement. By repeating this procedure to measure I - V and C - V curves alternately for the second cycle, third cycle, and so on, we can determine the type of carrier species and the amount of trapped charge based on shifts of the I - V and C - V curves.
- After several stressing cycles, the trap-filling will reach its saturation level, and the shift will also reach its maximum value. Based on the maximum shift of the C - V curve from the virgin C - V curve and the corresponding maximum shift in the I - V curves, we can determine the total trapped charge and hence the trap concentration, as well as the centroid of the trapped charge in the specimen.
- We can also use computer simulation to estimate the trap concentration and the cen-

troid by comparing the experimental I - V characteristics with the computed I - V characteristics obtained from Equations 6-98 through 6-108.

Using polyethylene thin film specimens and an MIS system with a p -Si semiconductor, we demonstrate this technique for characterization purposes. One advantage of using thin film specimens for experiments is that very high fields can easily be achieved at relatively low applied voltages, so surface leakage or discharge associated with high voltages can be avoided. In this case, when the gate electrode is negatively biased, holes are injected from the p -Si through the Si-polymer contact to the polyethylene films. Typical I - F and C - F characteristics are shown in Figure 7-45.

When a ramp voltage is applied to the MIS system, the current is constant up to the threshold field for hole injection. This constant current is the displacement current due to $C(dV/dt)$, where C is the total capacitance of the MIS system. For the $F > F_{th}$, the I - F characteristics are controlled by the effective field at the injecting contact, which in turn is governed by the rate of hole trapping. The shifts of the I - F curves to the right and of the corresponding C - F curves to the left after the first cycle of electrical stressing, as shown in Figure 7-45(a), indicate clearly that injected holes have been trapped, forming a homo-space charge near the injecting contact. The shifts gradually diminish after the second and further cycles of electrical stressing, indicating that the filling of the bulk hole traps has reached a dynamic equilibrium, unless the stressing field or stressing time is increased. When a sweeping cycle with a swing range from 0.5 MV cm^{-1} to -1.5 MV cm^{-1} is applied to the gate electrode for C - F measurements, the C - F curves (branches 1 and 2) form a hysteresis loop which always runs counter-clockwise regardless of the sweeping direction, that is, the sweeping direction starting either from the positive gate voltage or from the negative gate voltage, as shown in Figure 7-45(b). This provides good evidence about hole injection and trapped-hole space charge in polyethylene.¹⁹³

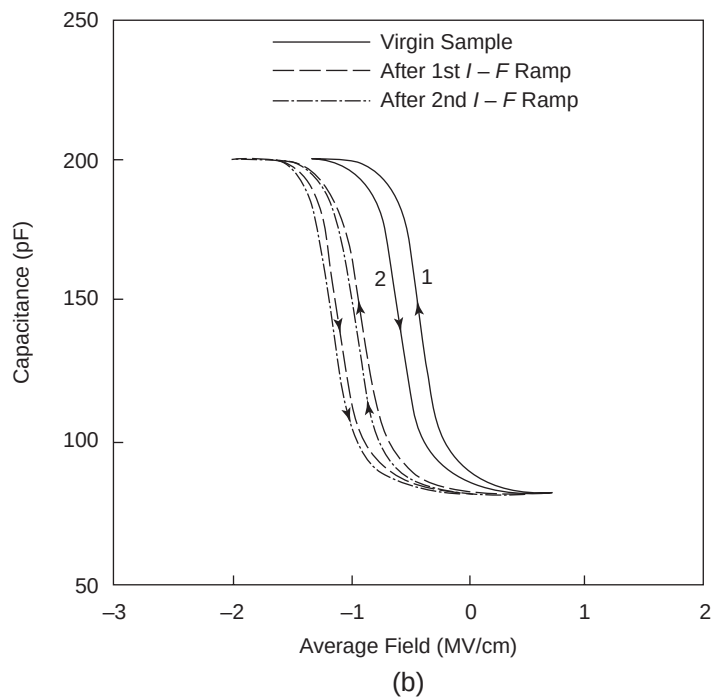
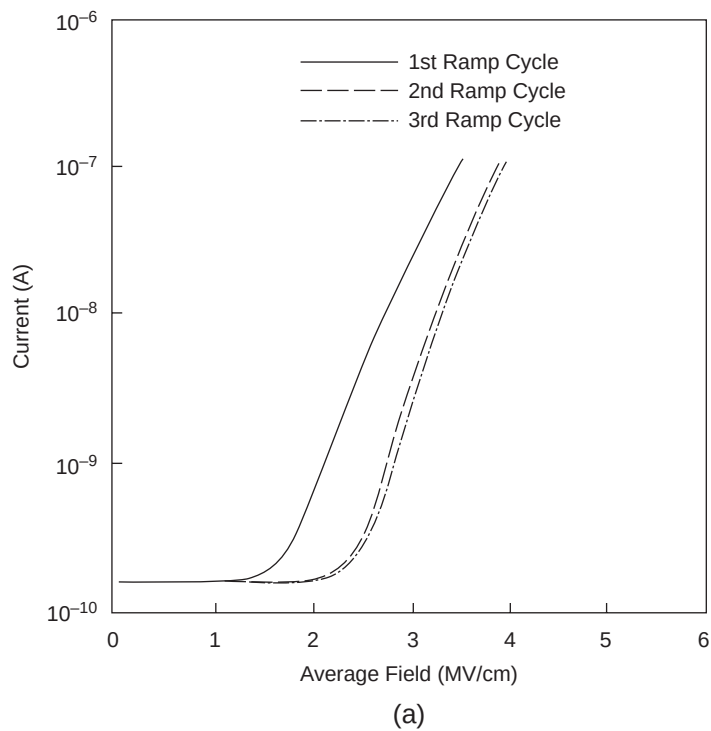


Figure 7-45 Effects of ramp voltage stressing on (a) the current–average field (I – F) characteristics and (b) the corresponding capacitance–average field (C – F) characteristics of polyethylene films for a virgin specimen after first, second, and third stressing cycles for an Al gate electrode negatively biased and a ramp rate of $0.06 \text{ MV cm}^{-1} \text{ s}^{-1}$.

7.9.3 Measurements of Surface Potentials

It is likely that the trapped carriers are concentrated in a narrow region of δ in thickness from the carrier-injecting contact, as shown in Figure 7-46, because of the small mean free path and the high concentration of traps in most dielectric materials, such as polymers. This trapped charge will produce a surface potential. If the injected carriers are electrons and the trapped charge is formed by trapped electrons of concentration n_t , then the surface potential V_s can be written as

$$V_s = \frac{\delta dq n_t}{\epsilon} \quad (7-394)$$

All parameters in Equation 7-394 have been previously defined. The electron trapping rate can be written as¹⁹⁴

$$\frac{dn_t}{dt} = \frac{\sigma J}{q} (N_t + N'_t - n_t) \quad (7-395)$$

Equation 7-395 includes the newly field-created trap concentration N'_t . If the stressing field is not very high or the period of electrical stressing is not very long, then $N_t > N'_t$. In this case, N'_t can be ignored.

To measure trap concentration by measuring surface potentials, a step-function voltage is used as a test voltage, with a magnitude sufficiently large to cause injection of electrons to

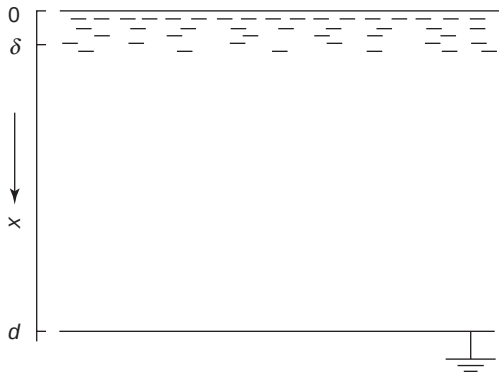


Figure 7-46 Schematic diagram illustrating the possible distribution of trapped electrons injected from the electron-injecting contact in a dielectric material under high electric fields.

gradually fill the traps, but not large enough to create a significant number of new traps. Then N'_t , created by the high-field electrical stressing or due to electrical aging, would not be affected by this test voltage. The dark conduction current decay with time after the application of a step-function electric field has been experimentally proved to be associated with the trap-filling process.¹⁰ This phenomenon was discussed in Section 7.3. Several investigators have studied this phenomenon.¹⁹⁵⁻¹⁹⁹ Their analysis and experimental evidence indicates that J is proportional to t^{-m} with m close to unity. Therefore, J can be reasonably expressed as

$$J = J_0(1 + Kt^{-m}), \quad t > 0, 0 < m < 1 \quad (7-396)$$

where K and m are constants depending on the material structure and the potential barrier profile of the injecting contact, and J_0 is the quasi-steady state current after the current decay transient period, which is field dependent.

Substituting Equation 7-396 into Equation 7-395 and solving it with the boundary condition, $n_t = 0$ at $t = 0$, we obtain¹⁹⁴

$$n_t = (N_t + N'_t) \left\{ 1 - \exp \left[-\frac{\sigma J_0 t}{q} \left(1 + \frac{K}{1-m} t^{-m} \right) \right] \right\} \quad (7-397)$$

Substitution of Equation 7-397 into Equation 7-394 gives

$$V_s(t) = V_s(\infty) \left\{ 1 - \exp \left[-\frac{\sigma J_0 t}{q} \left(1 + \frac{K}{1-m} t^{-m} \right) \right] \right\} \quad (7-398)$$

where

$$V_s(\infty) = \frac{\delta q d}{\epsilon} (N_t + N'_t) \quad (7-399)$$

When a step-function test voltage is applied to the specimen for a long time (e.g., 40 to 60 minutes), trap filling would reach a quasi-equilibrium state, (i.e., most traps have been filled by electrons). We refer this time as $t \rightarrow \infty$. Thus, the measurement of $V_s(\infty)$ would reveal the total trap concentration $N_t + N'_t$. Before the specimen is subjected to high-field or prolonged electrical stressing, the surface potential denoted by $V_s(\infty)$ measures only the originally existing trap concentration N_t , which is

$$V_{so}(\infty) = \frac{\delta q d}{\epsilon} N_t \quad (7-400)$$

Therefore, the newly field-created trap concentration N_t' can be determined by

$$\Delta V_s(\infty) = V_s(\infty) - V_{so}(\infty) = \frac{\delta q d}{\epsilon} N_t' \quad (7-401)$$

7.9.4 Capacitance Transient Spectroscopy

Deep-level transient spectroscopy (DLTS) has been used to study trap parameters in the bulk and at the interface mainly of MIS systems.²⁰⁰ In practice, it is more convenient to use a simple MIM system because it is sometimes difficult to make an intimate contact between the insulating material and the semiconductor. The method of employing an MIM system to measure the capacitance transient is referred to as *capacitance transient spectroscopy*.²⁰¹

Consider an MIM system with an electron-injecting contact injecting electrons into a dielectric specimen at fields higher than F_{th} . The injected electrons will soon be trapped, forming a trapped charge which will affect the total capacitance of the system. In general, the trapped electrons are distributed between the two metal plates. To simplify the analysis, we assume that all trapped electrons are concentrated on one sheet with the total trapped charge q_s located at the centroid x_o from the injecting contact, as shown in Figure 7-47(a). Obviously, with the charge sheet located at x_o , the potential distributions in region I and in region II are different, depending on the location of x_o . For low-mobility insulating materials, x_o should be very close to the injecting contact (plate A). This sheet charge q_s will induce charges q_a and q_b on the plates A and B, respectively. The induced charges are given by

$$q_a = q_s(1 - x_o/d) \quad (7-402)$$

$$q_b = q_s(x_o/d) \quad (7-403)$$

$$q_s = q_a + q_b \quad (7-404)$$

Thus, based on the principle of linear superposition, the total capacitance C of the MIM

system consists of three components and can be written as

$$C = C_o - C_I + C_{II} \quad (7-405)$$

as shown in Figure 7-47(b). These components are given by²⁰¹

$$C_o = \epsilon \frac{S}{d} \quad (7-406)$$

$$C_I = \frac{q_s(1 - x_o/d)}{V(x_o/d) + \frac{q_s x_o}{\epsilon S}(1 - x_o/d)} \quad (7-407)$$

$$C_{II} = \frac{q_s(x_o/d)}{V(1 - x_o/d) - \frac{q_s x_o}{\epsilon S}(1 - x_o/d)} \quad (7-408)$$

where S is the area of the plates.

Physically, the sheet charge q_s produces depolarization in region I and polarization in region II. The former tends to decrease the capacitance, while the latter to increase it. It can be seen that C is a function of three variables: q_s , x_o , and V . If the applied stressing voltage is a step-function voltage, then as soon as the voltage V is applied to the MIM system, electron injection will start, electrons will be trapped, and the trapped electron concentration n_t will increase with time. Assuming that the trap filling is extended to $2x_o$, then the total trapped charge may be approximately expressed as

$$q_s = qS(2x_o)n_t \quad (7-409)$$

Thus, by measuring the capacitance transient spectrum, we can obtain information about the trapped charge and x_o . By measuring the spectrum as a function of temperature, we can also obtain information about the distribution of the trap energy levels.

We will use a polyimide (PI) film specimen and the *A1-PI-A1* MIM system to illustrate the use of this technique. A specimen containing either only the originally existing traps (N_t) or the combination of N_t and the newly created traps N_t' is neutral if the traps are not filled with charge carriers (i.e., $q_s = 0$ and therefore, $C = C_o$). To enable the use of this technique to determine trap parameters, we must find a means of filling the traps with electrons. The easy way is to use a test voltage similar in concept to the

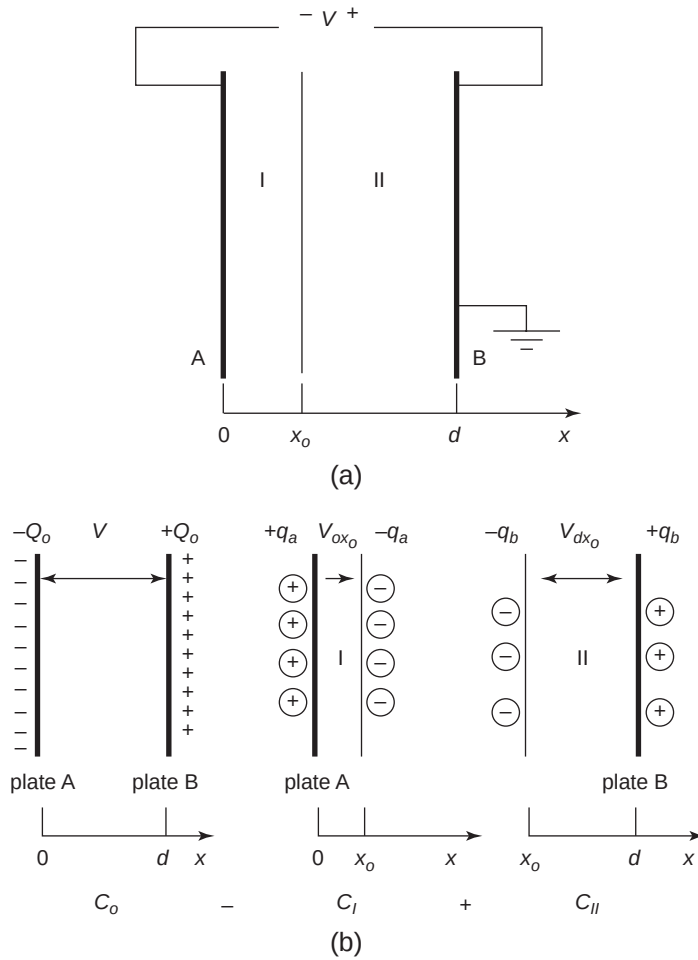


Figure 7-47 (a) An MIM system with a sheet charge q_s located at x_0 and an applied DC voltage between metal plates A and B. (b) The three components of the total capacitance C . C_0 : the capacitance of the MIM system without the sheet charge q_s ; C_I : the capacitance due to the depolarization by q_s ; C_{II} : the capacitance due to the polarization by q_s .

test voltage used for surface potential measurements. A step-function DC voltage can be used as the test voltage V with a magnitude sufficiently large to cause injection of electrons from the injecting contact to fill the traps gradually, but not large enough to create a significant number of new traps. This test voltage can also be considered a bias for the capacitance measurements based on Equations 7-405 through 7-408.

Substitution of Equation 7-397 into Equation 7-409 gives

$$q_s = qS(2x_0)(N_t + N_t') \times \left\{ 1 - \exp \left[-\frac{\sigma J_0 t}{q} \left(1 + \frac{K}{1-m} t^{-m} \right) \right] \right\} \quad (7-410)$$

As C is a function of q_s based on Equations 7-405 through 7-408, C varies with time following the variation of q_s with time during the trap-filling process. It can be seen that for large t (i.e., $t \rightarrow \infty$), $q_s \approx qS(2x_0)(N_t + N_t')$. By assuming that x_0 is close to the electron-injecting contact, we can write for $x_0/d < 1$

$$C(t \rightarrow \infty) - C_0 = \frac{qSd}{V} (N_t + N'_t) \quad (7-411)$$

Figure 7-48 shows $C - C_0$ as a function of t for polyimide films at various values of V (or $F = V/d$). After each measurement, the specimen was short-circuited for about one hour or longer to get rid of all the electrons trapped inside the specimen before starting another measurement with a different applied voltage. When the applied field is lower than 0.1 MV cm^{-1} , the capacitance is practically constant (i.e., $C = C_0$), indicating $q_s = 0$ although there are traps in the specimen. When $F \geq 1.5 \text{ MV cm}^{-1}$, C decreases with time, implying that electron injection has started and a negative space charge is being formed due to the trap-filling process. Before the application of any field, the specimen is a virgin specimen containing only the originally existing traps N_t . For $F = 1.5 \text{ MV cm}^{-1}$, the field may not create many new traps. In this case, we may assume $N_t > N'_t$ and estimate N_t . Using $S = 2 \times 10^{-2} \text{ cm}^2$, $d = 940 \text{ \AA}$ and $\epsilon = 3.4\epsilon_0$, N_t has been estimated to be about $4 \times 10^{17} \text{ cm}^{-3}$ from Equation 7-411. For $F = 3.2 \text{ MV cm}^{-1}$, the field is high enough to create a large quantity of new traps, making $N'_t > N_t$. In this case, we can ignore N_t , and N'_t is approximately $6 \times 10^{18} \text{ cm}^{-3}$. The incessant creation of new traps by carrier trapping under a stressing

field is the major mechanism leading to electrical aging.

PART II: PHOTOCONDUCTION

The photoconduction phenomenon was first discovered in selenium by Willoughby Smith in 1873.²⁰² As this phenomenon is a quantum process, it was not physically interpreted until 1911, by R. Pohl,²⁰³ who, with his school, systematically investigated it in the 1920s.²⁰⁴ Photoconduction in a material is induced by the absorption of photons of energy of electromagnetic waves, such as gamma rays, x-rays, and ultraviolet, visible, or infrared light.

Photoconduction is one of the important phenomena in solid-state materials. It has been used as a powerful tool to study defects and other physical parameters in solids. A variety of practical applications exists.²⁰⁵⁻²⁰⁷ Photoconduction has been observed in materials with resistivities ranging from less than 1 ohm-cm to more than 10^{18} ohm-cm . The lifetimes of the photogenerated carriers (a measure of photosensitivities) range from seconds to 10^{-13} second. The response times of photoconductors are strongly dependent on the densities of various defect states.

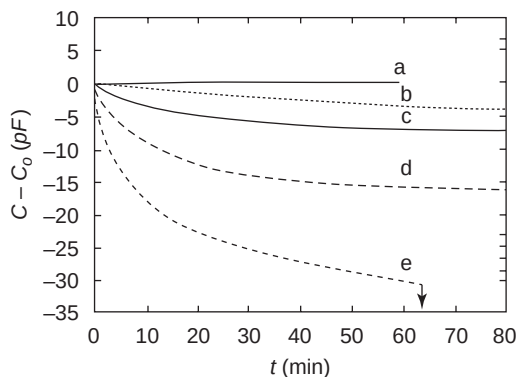


Figure 7-48 Typical capacitance transient spectra of an MIM system with polyimide film as the insulator at applied average fields (Vd^{-1}): (a) 0.1; (b) 1.5; (c) 2.1; (d) 2.7; and (e) 3.2 MV cm^{-1} . $S = 2.0 \times 10^{-2} \text{ cm}^2$; $d = 940 \text{ \AA}$. The arrow in (e) indicates the breakdown of the specimen.

7.10 Quantum Yield and Quantum Efficiency for Photoconduction

Quantum yield η is generally referred to as

- The number of free electron-hole pairs generated per light quantum absorbed (for intrinsic photoconduction)
- The number of free electrons or holes per light quantum absorbed (for extrinsic photoconduction)

Quantum efficiency g_{ph} is defined as the ratio of the number of charge carriers generated by photoexcitation passing through a solid between two electrodes to the number of light quanta absorbed by this solid during the same period of time.⁷³ Suppose that in the steady state free electrons of density Δn and free holes of

density Δp are generated during continuous photoexcitation. Then, we can write

$$\Delta n = G_n \tau_n \quad (7-412)$$

$$\Delta p = G_p \tau_p \quad (7-413)$$

where G_n and G_p are, respectively, the generation rates of electrons and holes per unit volume, and τ_n and τ_p are, respectively, the lifetimes of electrons and holes. For perfect crystals, both excitation and recombination are of a direct band-to-band process. In this case $\Delta n = \Delta p$, $G_n = G_p$ and hence $\tau_n = \tau_p$ (intrinsic photoconduction). For real crystals in which there are traps and recombination centers introduced by imperfections (chemical or structural), then $\Delta n \neq \Delta p$, $G_n \neq G_p$ and hence $\tau_n \neq \tau_p$ (extrinsic photoconduction).

In general, photocurrent density can be written as

$$J_{ph} = q \left[\frac{G_n \tau_n d}{t_n} + \frac{G_p \tau_p d}{t_p} \right] \quad (7-414)$$

or

$$J_{ph} = q[G_n X_n + G_p X_p] \quad (7-415)$$

where t_n and t_p are, respectively, the transit time for an electron and a hole to travel across the specimen of thickness d , and X_n and X_p are the distances that an electron and a hole have traveled in the direction of the electric field F during their lifetimes before they are trapped. Since

$$\begin{aligned} X_n &= \tau_n \mu_n F, & t_n &= d / \mu_n F \\ X_p &= \tau_p \mu_p F, & t_p &= d / \mu_p F \end{aligned} \quad (7-416)$$

it can be seen that at low fields, $J_{ph} = q(\mu_n n + \mu_p p)F$, which is proportional to F . This is because in this case, $X_n, X_p < d$ and $\tau_n, \tau_p < t_n, t_p$. At fields in which $X_n, X_p > d$ and $\tau_n, \tau_p > t_n, t_p$, the photocurrent tends to become saturated (independent of applied field). This is because in this case, the migrations of a large number of photogenerated carriers in the opposite direction will form a space charge if both the cathode and the anode are of blocking contacts. This means that a contact collecting holes cannot inject electrons or vice versa. (See Types of Electrical Contacts in Chapter 6).

The quantum efficiency g_{ph} (also called the *photoconductive gain*) can be expressed as

$$\begin{aligned} g_{ph} &= \frac{J_{ph}/q}{(G_n/\eta_n + G_p/\eta_p)d} \\ &= \frac{G_n X_n + G_p X_p}{[G_n/\eta_n + G_p/\eta_p]d} \\ &= \frac{G_n(\tau_n/t_n) + G_p(\tau_p/t_p)}{G_n/\eta_n + G_p/\eta_p} \end{aligned} \quad (7-417)$$

where η_n and η_p are, respectively, the quantum yields for electrons and holes. It is obvious that, with blocking contacts, the maximum value of $X_n + X_p$ is d . If it is assumed that $G_n = G_p$ and $\eta_n = \eta_p = \eta$, the maximum gain is $g_{ph} = \eta$.

If both electrodes are ohmic contacts (this implies that the cathode will inject electrons and the anode will inject holes to the photoconductor), then the situation is quite different. In this case, the gain can be higher than unity (if η is assumed to be unity). The charge carriers injected into the photoconductor from the electrodes are continuously neutralized by dielectric relaxation if the carrier transit time is larger than the dielectric relaxation time τ_d . Under this condition, there is negligible space charge near the electrodes and the photocurrent is linearly proportional to the applied voltage. If the applied voltage is increased to such values that the carrier transit time becomes equal to τ_d , and both decrease together with increasing applied voltage, then the photocurrent becomes space-charge limited and behaves in a manner similar to that described in Sections 7.4 and 7.5.

If the photoconductor is a perfect crystal, free of traps, the condition for the onset of the SCL photocurrent is carrier transit time equal to τ_d . If the photoconductor is a real crystal containing traps, the threshold voltage for the onset of the SCL current increases, because the presence of traps reduces the average carrier mobility and makes the experimentally observed decay time of the photocurrent (after the removal of excitation) longer than the carrier lifetime. (In trap-free perfect crystals, the observed decay time is equal to the carrier lifetime). This is due to two reasons:

1. The carriers released thermally from traps prolong the observed decay time of the

photocurrent if n_t and p_t (trapped electron and hole concentrations) are larger than the corresponding Δn and Δp .

2. The traps reduce effective drift mobility and carrier lifetime.

Supposing that $\Delta p \gg \Delta n$ and the minority carriers (electrons) are trapped immediately by traps after excitation, then Equation 7-417 reduces to

$$g_{ph} = \eta_p (\tau_p / t_{ip}) \quad (7-418)$$

Assuming that $\eta_p = 1$, $g_{ph} = \tau_p / t_{ip}$. This indicates the importance of carrier lifetime to photosensitivity of a solid. It is obvious that the value of g_{ph} is field dependent, temperature dependent, specimen-thickness dependent, and purity and structure dependent even for a given material. In Section 7.5, we discussed carrier lifetimes for perfect crystals free of traps, $\tau_p = (\langle v \sigma_p \rangle \Delta n)^{-1}$; while for real crystals with $\tau_p = (\langle v \sigma_p \rangle n_r)^{-1}$, where n_r is the electron-occupied trap concentration ready to capture holes (occupied recombination center concentration). If $n_r > \Delta n$, then τ_p (with recombination centers) is about $\Delta n / n_r$ times τ_p (without recombination centers). If the photoconductor also contains shallow hole traps, then the effective hole mobility is decreased and its transit time is increased. In this case, the transit time must be reduced to about $t_{ip} = (\Delta p / p) \tau_d$ or the applied voltage raised to $p_t / \Delta p$ times the case without traps for the onset of SCL photocurrent.^{73,99}

For photoconductors with ohmic contacts, the maximum gain for photoconduction is set when the SCL current starts to be comparable to the photocurrent. Thus,

$$(g_{ph})_{\max} = \frac{\tau_p}{t_{ip}} = \frac{\tau_p}{(\Delta p / p_t) \tau_d} \quad (7-419)$$

$$= \frac{\tau_{rp}}{\tau_d}$$

where τ_{rp} is the observed decay time or the response time, which is

$$\tau_{rp} = (p_t / \Delta p) \tau_p \quad (7-420)$$

In most cases, the observed decay time τ_{rp} does not exceed τ_d , so $(g_{ph})_{\max}$ is limited to unity. But $(g_{ph})_{\max} > 1$ can be expected if the energy levels

of traps are such that the effect of p_t on t_{ip} is greater than the effect of p_t on τ_{rp} . We can rewrite Equation 7-419 as

$$(g_{ph})_{\max} = \frac{\tau_{rp}}{\tau_d} \frac{(P_t)_{\text{SCL injection}}}{(P_t)_{\text{light excitation}}} \quad (7-421)$$

$$= \frac{\tau_{rp}}{\tau_d} M$$

where $(p_t)_{\text{SCL injection}}$ is the filled-trap density due to SCL injection and tends to increase the transit time, and $(p_t)_{\text{light excitation}}$ is the filled-trap density due to light excitation and tends to increase the observed decay time.⁷³ It is obvious that M can be made larger than unity.

7.11 Generation of Nonequilibrium Charge Carriers

In thermal equilibrium, the concentrations of electrons n_o and holes p_o thermally generated follow the mass-action law

$$n_o p_o = n_i^2 \quad (7-422)$$

where n_i is the intrinsic carrier concentration. When such an equilibrium condition is changed to a nonequilibrium one by an external force such as photoexcitation, the concentrations of excess electrons Δn and holes Δp generated by the photoexcitation are

$$\Delta n = n - n_o \quad (7-423)$$

$$\Delta p = p - p_o$$

and the total concentrations of electrons n and holes p are no longer governed by the mass action law

$$np \neq n_i^2 \quad (7-424)$$

Several processes may occur simultaneously in competing for the absorption of the energy of the photons penetrating into the photoconductor.

Process 1: If the photon energy for the photoexcitation is equal to or slightly higher than E_g , it will produce Δn and Δp through a band-to-band transition process. Δn and Δp can be written as

$$\Delta n = \eta_n \alpha I \tau_n \quad (7-425)$$

$$\Delta p = \eta_p \alpha I \tau_p \quad (7-426)$$

where η_n and η_p are, respectively, the quantum yields for electrons and holes, α is the absorption coefficient, I is the light intensity, and τ_n and τ_p are, respectively, the lifetimes of the excess electrons and the excess holes.

Process 2: The photon energy may be absorbed by the material in causing a transition within the allowed bands. For example, an electron may be raised from a lower level to a higher

level in the conduction band, as shown in Figure 7-49(a).

Process 3: Similar to Process 2, an electron in the valence band may be raised from a lower level to occupy a hole available near the valence band edge. This is equivalent to saying that a hole is raised from a lower level to a higher level, as shown in Figure 7-49(a).

Process 4: The photon energy may be absorbed by the material in causing the transition of a

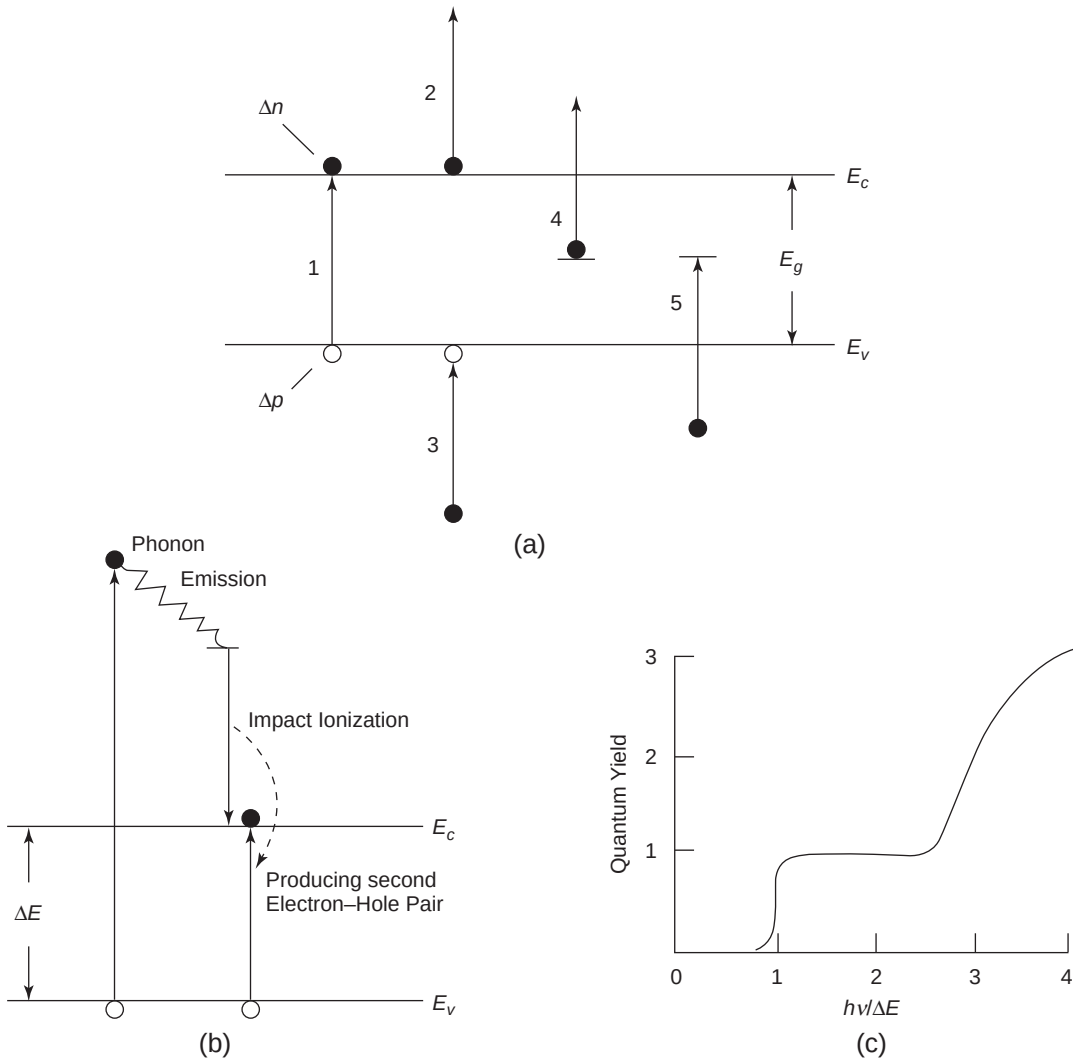


Figure 7-49 Schematic diagrams illustrating (a) the five competing processes for the absorption of the exciting photon energy, (b) the process of absorbing a high photon energy with $h\nu > 2E_g$ leading to the generation of a second electron-hole pair, and (c) quantum yield as a function of photon energy.

trapped electron in the recombination center to the conduction band, producing an excess electron. This process is, in fact, a photo-detrapping process.

Process 5: The photon energy may be absorbed by the material in causing the transition of an electron in the valence band to occupy a state in the recombination center, thus producing an excess hole. This process is also a photo-detrapping process. Processes 2, 3, 4 and 5 tend to reduce the quantum yield η_n and η_p to a value less than unity.

In general, if ΔE is the energy required for making a band-to-band transition, ΔE is greater for optical transition than for thermal transition, due mainly to the restriction of the Franck–Condon principle (see The Franck–Condon Principle in Chapter 3). Thus, we can write

$$(\Delta E)_{\text{optical}} > (\Delta E)_{\text{thermal}} \approx E_g \quad (7-427)$$

When the photon energy $h\nu > 2\Delta E$, the absorption of a single high-energy photon may lead to the generation of two electron–hole pairs, as shown in Figure 7-49(b). Theoretically, when $h\nu = 2\Delta E$, impact ionization may lead to the generation of a second electron–hole pair. Experimentally, however, impact ionization becomes significant only when $h\nu > 2\Delta E$ because of the limitation of selection rules, which govern the probability of photoionization and impact ionization. Quantum yield as a function of photon energy is shown in Figure 7-49(c). It should be noted that a high-energy electron will lose its excess energy after about 10^3 collisions with phonons or defects. If the electron has a mean free path of 10^{-6} cm and a thermal velocity of 10^7 cm/sec, it will lose its excess energy in less than 10^{-10} sec.

7.11.1 Energy Distribution of Nonequilibrium Charge Carriers

The energy distribution of the nonequilibrium carriers in a band does not differ from that of the equilibrium carriers for the majority of their lifetimes. In thermal equilibrium, we use Fermi level E_F to describe electron and hole concentrations:

$$\begin{aligned} n_o &= N_c \exp[-(E_c - E_F)/kT] \\ &= n_i \exp[(E_F - E_i)/kT] \\ p_o &= N_v \exp[-(E_F - E_v)/kT] \\ &= n_i \exp[(E_i - E_F)/kT] \end{aligned} \quad (7-428)$$

In nonequilibrium steady state, we can use the so-called quasi-Fermi levels to describe electron and hole concentrations

$$\begin{aligned} n &= n_o + \Delta n = N_c \exp[-(E_c - E_{Fn})/kT] \\ &= n_i \exp[(E_{Fn} - E_i)/kT] \\ p &= p_o + \Delta p = N_v \exp[-(E_{Fp} - E_v)/kT] \\ &= n_i \exp[(E_i - E_{Fp})/kT] \end{aligned} \quad (7-429)$$

where E_F and E_i are, respectively, the Fermi level in thermal equilibrium and the intrinsic Fermi level; E_{Fn} and E_{Fp} are, respectively, the quasi-Fermi levels for the total electron and hole concentrations, including excess electrons and holes. The energy distributions of the nonequilibrium and equilibrium carriers are identical. The generation of nonequilibrium carriers simply alters the concentration of free carriers, leaving unaffected the energy distribution of these carriers in the bands and the average kinetic energy per free carrier. Consequently, the mobilities of the nonequilibrium carriers do not differ from those of the equilibrium ones. Equilibrium and the nonequilibrium carriers have the same average probability of recombination.

7.11.2 Spatial Distribution of Nonequilibrium Charge Carriers

The concentrations of photogenerated charge carriers Δn and Δp are not uniformly distributed in space inside the specimen because the light intensity I is spatially nonuniform, due to the spatial dependence of the absorption following the absorption law

$$I(y) = I_o \exp(-\alpha y) \quad (7-430)$$

where I_o is the light intensity at the illuminated surface (see Absorption and Dispersion in Chapter 3). Obviously, this will make $\Delta n(y)$ change with y . Two electrode-photoconductor-electrode configurations used most commonly

for photoconduction measurements are transverse photoconduction and longitudinal photoconduction. For transverse photoconduction, the light beam is normal to the applied electric field. For longitudinal photoconduction, the light beam is parallel to the applied field and penetrates into the specimen through a semi-transparent electrode, as shown in Figure 7-50.

Consider a photoconductor specimen like the one shown in Figure 7-50(a). If we take the photogenerated electrons as example, Δn will vary with y , following the form similar to Equation 7-430. Thus, we can write

$$\Delta n(y) = (\Delta n)_o \exp(-\alpha y) \quad (7-431)$$

$$= \eta_n \alpha I_o \exp(-\alpha y) \tau_n$$

Assuming that the diffusion of the carriers along the y direction may be neglected for photoconductors with a high conductivity, by dividing the specimen into layers, each of

thickness Δy , the photoconductance within one layer Δy can be written as

$$q\mu_u \Delta n(y) \frac{w\Delta y}{d} \quad (7-432)$$

Since all layers are connected in parallel, the total photoconductance can be expressed as

$$q\mu_u (w/d) \int_0^h \Delta n(y) dy \quad (7-433)$$

$$= q\mu_u (w/d) \eta_n \tau_n I_o [1 - \exp(-\alpha h)]$$

Thus, the average transverse photoconductivity for photogenerated electrons is

$$\sigma_{ph} = q\mu_n h^{-1} \eta_n \tau_n I_o [1 - \exp(-\alpha h)] \quad (7-434)$$

Similarly, we can obtain the average transverse photoconductivity for photogenerated holes. However, it can be seen from Equation 7-434 that for a thick specimen with a large value of h , photoconductivity is independent of the

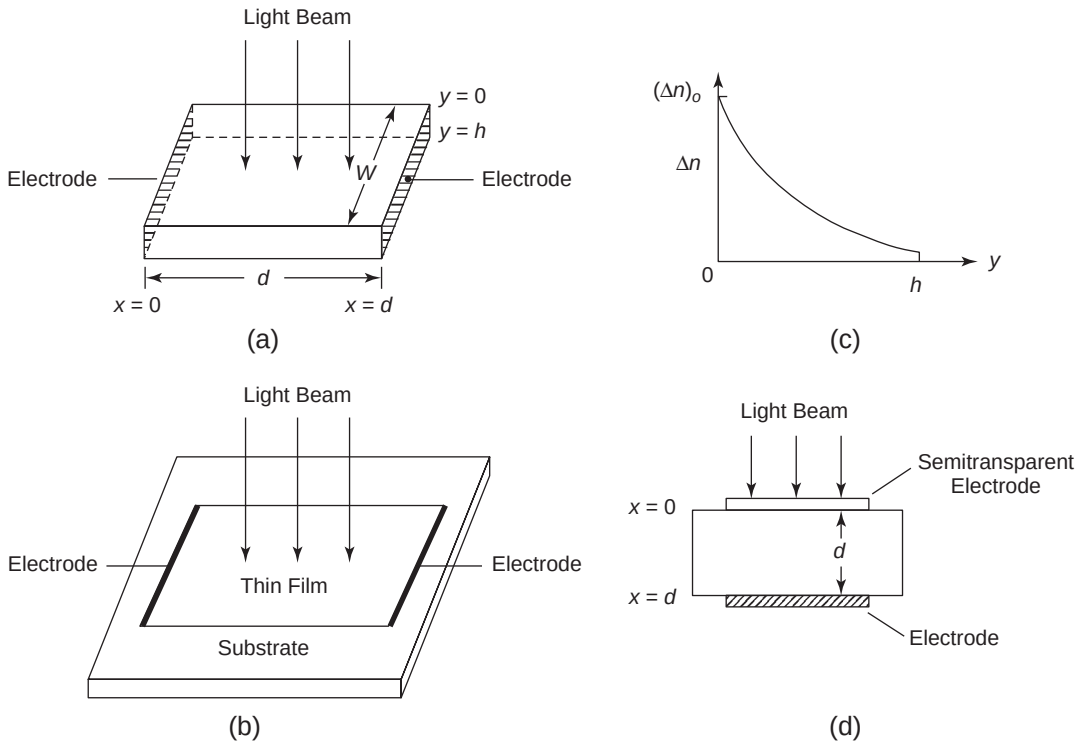


Figure 7-50 Schematic diagram showing electrode-photoconductor-electrode configurations: (a) for measuring transverse photoconductivity in plate form (b) for measuring transverse photoconductivity in thin-film form, (c) dependence of the excess carrier density on the distance from the illuminated surface for transverse photoconduction, and (d) for measuring longitudinal photoconductivity.

absorption coefficient and is governed only by the total amount of light energy I_o penetrating into the specimen.

In longitudinal photoconduction, the relationship between the photoconductivity and the phenomenological parameters (η , α , and τ) is quite complex. However, a simple case with $\Delta n \ll n_o$ has been analyzed.²⁰⁸

7.11.3 Lifetimes of Nonequilibrium Charge Carriers

If no other processes affect the photogeneration of nonequilibrium carriers, the excess carrier concentration will increase with time without limit, as shown in Figure 7-51(a). As the number of photogenerated carriers increases, however, the rate of recombination also increases. Finally, the carrier generation rate becomes equal to the carrier recombination rate. In thermal equilibrium, the carrier generation rate G_{th} is equal to the recombination rate R_{th} . Thus, we can write

$$G_{th} = R_{th} = C_r n_o p_o = C_r n_i^2 \quad (7-435)$$

where C_r is the recombination coefficient. In nonequilibrium, we can write $G = R$, similar to the thermal equilibrium one as

$$G = R = C_r np = C_r (n_o + \Delta n)(p_o + \Delta p) \quad (7-436)$$

So the recombination rate for only the nonequilibrium excess carriers Δn and Δp can be written as

$$\begin{aligned} \Delta R &= R - R_{th} = G - G_{th} \\ &= R_{th} \left[\frac{n_o \Delta p + p_o \Delta n + \Delta n \Delta p}{n_i^2} \right] \end{aligned} \quad (7-437)$$

Each nonequilibrium carrier, (e.g., an electron) will undergo thermal motion. While moving in the conduction band, it will recombine with a free hole or a trapped hole in the recombination center. Based on Equations 7-188 and 7-189, the lifetime of nonequilibrium carriers can be written as

$$\tau_n = \frac{\Delta n}{\Delta R} = \frac{1}{R_{th}} \left\{ \frac{n_i^2 \Delta n}{n_o \Delta p + p_o \Delta n + \Delta n \Delta p} \right\} \quad (7-438)$$

$$\tau_p = \frac{\Delta p}{\Delta R} = \frac{1}{R_{th}} \left\{ \frac{n_i^2 \Delta p}{n_o \Delta p + p_o \Delta n + \Delta n \Delta p} \right\} \quad (7-439)$$

We shall discuss several cases.

Linear Recombination

When the photoexcitation is low, then $n_o + p_o \gg \Delta n$ and $\Delta n = \Delta p$. In this case (linear recombination), recombination velocity is linearly related to Δn , so we have

$$\Delta R = R_{th} \frac{(n_o + p_o) \Delta n}{n_i^2} \quad (7-440)$$

and

$$\tau_n = \tau_p = \frac{1}{R_{th}} \left(\frac{n_i^2}{n_o + p_o} \right) \quad (7-441)$$

There are three subcases:

Intrinsic materials: $n_o = p_o = n_i$. The lifetime of the excess electrons is

$$\tau_i = \tau_n = \tau_p = \frac{n_i}{2R_{th}} \quad (7-442)$$

Extrinsic n-type materials: $n_o \gg p_o$. The lifetime of the excess electrons is

$$\tau_n = \frac{n_i^2}{n_o} \left(\frac{1}{R_{th}} \right) \quad (7-443)$$

Since $n_o \gg n_i$, τ_n for extrinsic materials is always smaller than for intrinsic materials.

Extrinsic p-type materials: $n_o \ll p_o$. The lifetime of the excess electrons is

$$\tau_n = \frac{n_i^2}{p_o} \left(\frac{1}{R_{th}} \right) \quad (7-444)$$

As in the case of extrinsic n-type materials, since $p_o \gg n_i$, the lifetime of excess carriers in extrinsic materials is smaller than for intrinsic materials.

The simple analysis above indicates that the recombination of nonequilibrium excess carriers via impurity centers (localized states in the band gap) is much faster than through band-to-band recombination.

For the linear recombination, the recombination rate is proportional to the nonequilibrium

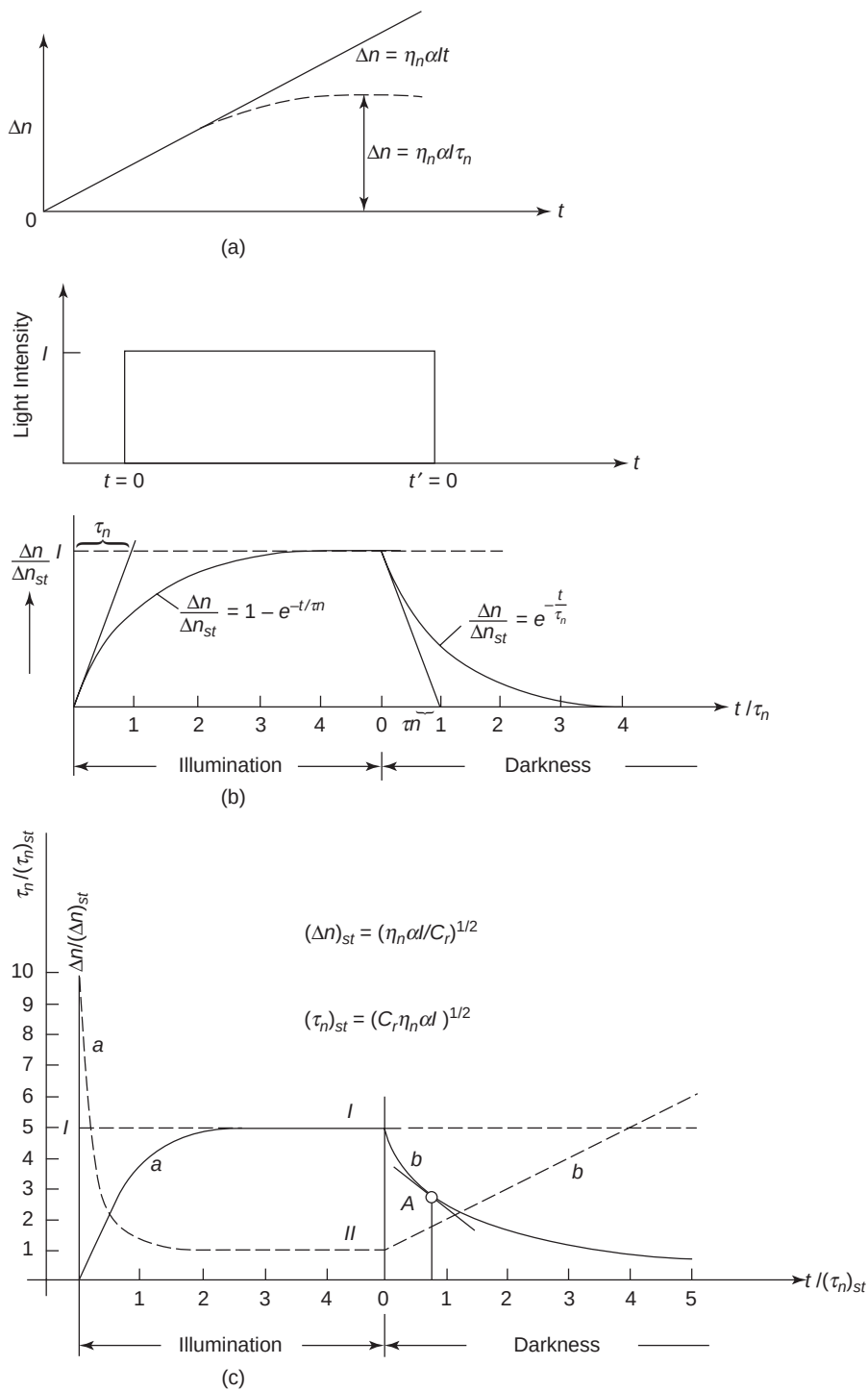


Figure 7-51 Schematic diagrams showing the variation of the concentration of photogenerated excess electrons with optical excitation time during the rise period and also during the decay period after the removal of optical excitation: (a) Δn as a function of time without limiting processes, (b) Δn as a function of time in the case of linear recombination, (c) Δn as a function of time in the case of quadratic recombination. In curve I, a is the rise period and b is the decay period of Δn . Curve II shows the instantaneous lifetime during rise period a and decay period b .

carrier concentration. Thus, the change in the nonequilibrium carrier concentration (for electrons as an example) may be written as

$$\begin{aligned}\frac{d(\Delta n)}{dt} &= \text{generation rate} - \text{recombination rate} \\ &= \eta_n \alpha I - \frac{\Delta n}{\tau_n}\end{aligned}\quad (7-445)$$

Now, if we use a rectangular light beam pulse to excite a dielectric specimen, we can find the variation of Δn with time during the rise period, when the light pulse is switched on, and during the decay period, after the light pulse is switched off, by solving Equation 7-445.

During the rise period, we use the initial condition: when $t = 0$, $\Delta n = 0$. Then, the solution of Equation 7-445 yields

$$\begin{aligned}\Delta n(t) &= \eta_n \alpha I \tau_n (1 - e^{-t/\tau_n}) \\ &= (\Delta n)_{st} (1 - e^{-t/\tau_n})\end{aligned}\quad (7-446)$$

where $(\Delta n)_{st}$ is the steady-state value of Δn . Later, we will use Δn instead of $(\Delta n)_{st}$ for the steady-state value of Δn and $\Delta n(t)$ for the transient value of Δn .

During the decay period, after $\Delta n(t)$ has reached a steady-state value $(\Delta n)_{st}$, the light beam is switched off and (Δn) starts to decay with time. Thus, using the boundary condition when $t' = 0$, $I = 0$ and $\Delta n(t) = (\Delta n)_{st}$, the solution of Equation 7-445 gives

$$\begin{aligned}\Delta n(t') &= \eta_n \alpha I \tau_n e^{-t'/\tau_n} \\ &= (\Delta n)_{st} e^{-t'/\tau_n}\end{aligned}\quad (7-447)$$

The variation of $\Delta n(t)$ with time is shown in Figure 7-51(b). The rise and decay relaxation curves for this case are similar. The time constant of these exponential curves is the lifetime of the nonequilibrium excess carriers. This linear recombination case prevails when holes or empty states in the impurity centers are available to recombine with the nonequilibrium electrons. Similarly, this case prevails when electrons or occupied states in the impurity centers are available to recombine with the nonequilibrium holes. This is a simple way to determine τ_n and τ_p from the relaxation curves. Equations 7-446 and 7-447 indicate that τ_n and τ_p are constant, and that $(\Delta n)_{st}$ (and hence the

photoconductivity τ_{ph}) are linearly proportional to light intensity I .

Quadratic Recombination

When the photoexcitation is high, then $\Delta n \gg n_o + p_o$ and $\Delta n = \Delta p$. In this case, the recombination rate R and ΔR can be written as

$$\begin{aligned}R &= C_r np = C_r (n_o + \Delta n)(p_o + \Delta p) \\ \Delta R &= R - R_0 = C_r \Delta n \Delta p = C_r (\Delta n)^2\end{aligned}\quad (7-448)$$

The recombination rate is quadratically related to Δn (quadratic recombination). So the change of the nonequilibrium carrier concentration (for electrons as an example) may be written as

$$\frac{d(\Delta n)}{dt} = \eta_n \alpha I - C_r (\Delta n)^2 \quad (7-449)$$

Using the same procedure as for linear recombination, we can obtain Δn as a function of time for quadratic recombination.

During the rise period, we have

$$\Delta n(t) = (\eta_n \alpha I / C_r)^{1/2} \tanh[(C_r \eta_n \alpha I)^{1/2} t] \quad (7-450)$$

During the decay period, we have

$$\Delta n(t) = (\eta_n \alpha I / C_r)^{1/2} [(C_r \eta_n \alpha I)^{1/2} t + 1]^{-1} \quad (7-451)$$

In this case, the rise and the decay relaxation curves are no longer symmetrical. The decay curve changes much more slowly than the rise curve, implying that the lifetime is not a constant value. It varies with light intensity as well as with time. During the rise period, the lifetime is

$$\begin{aligned}\tau_n(t) &= \frac{1}{C_r \Delta p} = \frac{1}{C_r \Delta n} \\ &= \frac{1}{(C_r \eta_n \alpha I)^{1/2}} \coth[(C_r \eta_n \alpha I)^{1/2} t]\end{aligned}\quad (7-452)$$

During the decay period, the lifetime is

$$\tau_n(t) = \frac{1}{(C_r \eta_n \alpha I)^{1/2}} [(C_r \eta_n \alpha I)^{1/2} t + 1] \quad (7-453)$$

The variation of Δn and τ_n with time is shown in Figure 7-51(c). This case prevails when

thermal equilibrium carrier concentrations are very small compared to photogenerated carrier concentrations. From Equations 7-452 and 7-453, it can be seen that when the light is just switched on, $\Delta n \rightarrow 0$ and $\Delta p \rightarrow 0$, the lifetime $\tau_n \rightarrow \infty$. After the light has been switched off for some time, $\Delta n \rightarrow 0$ and $\Delta p \rightarrow 0$, so τ_n is also approaching infinity. For quadratic recombination, the steady-state values of Δn and τ_n become

$$\begin{aligned} (\Delta n)_{st} &= \left(\frac{\eta_n \alpha I}{C_r} \right)^{1/2} \\ (\tau_n)_{st} &= (C_r \eta_n \alpha I)^{-1/2} \end{aligned} \quad (7-454)$$

Since $(\Delta n)_{st}$ is proportional to $I^{1/2}$, the photoconductivity σ_{ph} is also proportional to $I^{1/2}$.

Instantaneous Lifetimes

The lifetime of nonequilibrium carriers is generally not constant but varies with light intensity and time. So, the lifetime should be expressed as

$$\tau = f(I, t) \quad (7-455)$$

Only in the special case of linear recombination may the lifetime be considered independent of light intensity and time. It is clear that for quadratic recombination, the lifetime depends on both light intensity and time, as shown in Figure 7-51(c).

If there are several types of capture centers acting as free holes, or trapped holes which would capture electrons, and if each type has its own capture cross-section, density, and average relative velocity of motion, then the lifetime of nonequilibrium carriers due to the j th-type capture centers, based on Equation 7-188, can be written as

$$\tau_{nj} = \frac{1}{\langle v_j \sigma_{nj} \rangle (N_{rj} - n_{rj})} \quad (7-456)$$

where N_{rj} and n_{rj} are, respectively, the concentrations of j th-type centers (including empty and occupied centers) and occupied centers. The effective average lifetime is the summation of the lifetimes due to each type of center. Thus, we have

$$\begin{aligned} \frac{1}{\tau_n} &= \sum_i \frac{1}{\tau_{ni}} \\ \frac{1}{\tau_p} &= \sum_i \frac{1}{\tau_{pi}} \end{aligned} \quad (7-457)$$

Average recombination rates for electrons and holes can be expressed as

$$\begin{aligned} \overline{\langle v_n \sigma_n \rangle (N_r - n_r)} &= \frac{\Delta n}{\tau_n} \\ \overline{\langle v_p \sigma_p \rangle (n_r)} &= \frac{\Delta p}{\tau_p} \end{aligned} \quad (7-458)$$

Obviously, since τ_n and τ_p depend on $(N_r - n_r)$ and n_r , respectively, they are not constant for a given material consisting of multiple capture centers. They vary with time under non-steady state conditions, with light intensity and temperature.

Carrier trapping affects the lifetime of non-equilibrium carriers. So-called carrier trapping means that nonequilibrium electrons or holes may be captured by traps, and later, the trapped electrons or trapped holes will be thermally reexcited back to the bands. For example, if a dielectric specimen has only electron traps of concentration N_t located at ΔE_t below E_c , then with photoexcitation, the total carrier concentrations are

$$\begin{aligned} n &= n_o + \Delta n \\ n_t &= n_{to} + \Delta n_t \\ p &= p_o + \Delta p \end{aligned} \quad (7-459)$$

where n_t , n_{to} and Δn_t are, respectively, the concentrations of total trapped electrons, trapped thermal equilibrium electrons, and trapped photogenerated electrons.

Following the same procedure used in Sections 7.11.3, the steady-state lifetimes τ_n and τ_p can be readily derived, and they are

$$\tau_n = \frac{1}{C_r [(1+a)n_o + p_o]} \quad (7-460)$$

$$\tau_p = \frac{1+a}{C_r [(1+a)n_o + p_o]} \quad (7-461)$$

where

$$a = \frac{\Delta n_t}{\Delta n} = \frac{n_t - n_{to}}{n - n_o} \quad (7-462)$$

Since $\Delta n_i/\Delta n$ is always positive, Equations 7-460 and 7-461 indicate that with electron traps only, the electron trapping tends to decrease the lifetime of the nonequilibrium electrons and to increase the lifetime of nonequilibrium holes, and vice versa. For more details about the effect of carrier trapping, see references.^{73,208}

7.12 Photoconduction Processes

The carrier mobilities u_n and u_p are generally assumed to be unaffected by light excitation. Some changes may occur under certain conditions, but these are small and insignificant. If the applied electric field is small and if the photoconductor is homogeneously excited (so the distribution of photogenerated carriers remains uniform), the photoconduction current (simply called the *photocurrent*) depends only on how the free carriers are photogenerated. The photocurrent depends on the wavelength and the intensity of the exciting light, applied electric or magnetic fields, temperature, surface condition, and ambient atmosphere, because photo-generated Δn and Δp are dependent on these parameters. Photoconduction can be intrinsic or extrinsic, depending on the dominant photocarrier generation process.

7.12.1 Intrinsic Photoconduction

Photoconduction means the excitation of an electron into the conduction band by the absorption of energy from an incident photon in the photoconductor. Photoconduction is intrinsic if this electron originates in a full valence band. This is, in fact, a band-to-band transition. In this case, both electrons and holes contribute to the generation of a photocurrent. When the photons incident on the photoconductor have energies above the fundamental absorption threshold ($>E_g$), intrinsic photocurrent density can be expressed as

$$\begin{aligned} J_{ph} &= \sigma_{ph} F = q(u_n \Delta n + u_p \Delta p) F \\ &= q(u_n + u_p) \Delta n F \end{aligned} \quad (7-463)$$

If the field F is small, J_{ph} is proportional to F . At high fields, the spatial distribution of

carriers becomes nonuniform because of the space charge effect. In such cases, Ohm's law is not obeyed, and the photocurrent becomes saturated.

Direct band-to-band transition generally occurs in inorganic crystals, and, to a lesser extent, in organic crystals. Single-photon intrinsic photoconduction can occur only when photon energies exceed the optical absorption threshold. However, in organic crystals, carriers are generally produced via intermediate steps involving excitons. The formation and the interaction of excitons was discussed in Formation and Behavior of Excitons in Chapter 3.

Two-photon carrier generation processes are directly associated with the generation of two photoexcited states (i.e., excitons), which interact, resulting in a direct transition of an electron from the valence band to the conduction band or to an autoionization state above E_c .^{209,210} There are several possible processes leading to the generation of intrinsic electron-hole pairs, such as singlet exciton-singlet exciton collision ionization and singlet exciton-triplet exciton collision ionization. For details about photoconduction resulting from exciton interaction processes, see references.^{52,144,209-213}

7.12.2 Extrinsic Photoconduction

Extrinsic photoconduction is generally unipolar, that is, it involves mainly one type of carrier (either electrons or holes are dominant). This section discusses features of extrinsic photoconduction involving trapping and recombination centers.

Previously, we discussed the rise time and the decay time of photogenerated carriers Δn and Δp . The trapping of free carriers causes rise time and decay time to be much greater than carrier lifetime. Rise time is the time required for the traps to capture the photogenerated free carriers and for the steady state to be established between the new density of free carriers and the new occupancy of traps after an exciting light is switched on. Decay time is the time required for the trapped carriers to be released thermally after the excitation is terminated.

The observed decay time (or response time) can be related to the carrier lifetime by the following approximate equation⁷³:

$$\tau_{rn} = \tau_n(1 + n_i/n) \quad (7-464)$$

$$\tau_{rp} = \tau_p(1 + p_i/p) \quad (7-465)$$

depending on whether electrons or holes are the dominant carriers. For perfect crystals free of traps, the decay time is equal to the lifetime. In general, the lifetimes τ_n and τ_p are insensitive to light intensity. If $p_i \gg p$ for hole-dominant photoconductors, Equation 7-465 reduces to

$$\tau_{rp} = \tau_p(p_i/p) \quad (7-466)$$

By assuming $\Delta p = p - p_o \approx p$, from Equation 7-412 we obtain

$$p_i = G_p \tau_{rp} \quad (7-467)$$

Thus, we can estimate the concentration of trapped carriers. Since p_i is very sensitive to both temperature and photon energy, τ_{rp} is expected to decrease with increasing temperature and increasing photon energy. Measurements of photocurrent decay as a function of time at various temperatures, wavelengths, and intensities of the exciting light allow the determination of trap parameters.^{73,214}

Photocurrent–Voltage Characteristics

Case 1: With Carrier-Injecting Ohmic Contacts

If the concentration of photogenerated carriers is higher than injected carriers from electrodes, the J – V characteristics closely follow Ohm's law. However, depending on the applied voltage V , there is a critical voltage V_Ω at which a transition from the linear ohmic region to the superlinear SCL region takes place. The value of V_Ω shifts toward a higher voltage as the exciting light intensity is increased. For traps confined in a single discrete energy level in the band gap, the SCL current in the dark, for the case of dominant hole carriers, is given by

$$J = \frac{9}{8} \epsilon \mu_p \theta \frac{V^2}{d^3} \quad (7-468)$$

where $\theta = p/(p + p_i)$ (see Equation 7-87). Therefore, $\theta = 1$ if $p_i = 0$ (without traps).

If the exciting light intensity is low, thermal detrapping dominates the detrapping processes. Under this condition, the SCL current will be the same as in the dark. When the exciting light intensity is increased to such a level that optical detrapping becomes dominant, then the photocurrent depends on light intensity, and θ is no longer a function of temperature but a function of light intensity.

With carrier-injecting contacts, the photocurrent can become saturated if the concentration of majority carriers injected from the ohmic contact is suppressed at high fields by minority carriers injected from the opposite blocking contact through a recombination process, particularly in relaxation semiconductors^{50,51,215,216} or if the bulk-limited regime is changed to an electrode-limited regime at high fields.¹¹⁷

Case 2: With Noninjecting Blocking Contacts

It is obvious that with noninjecting blocking contacts, the photocurrent would become saturated when all photogenerated carriers were extracted from the photoconductor. Photoconductive gain is maximal when the applied field reaches a value at which the photocurrent is saturated. Suppose that the electrical contacts are noninjecting for electrons and holes, and that the photogenerated electron–hole pairs are uniformly distributed throughout the specimen. Then, all photogenerated carriers are extracted if the mean electron drift length $X_n (= \tau_n \mu_n F)$ and the mean hole drift length $X_p (= \tau_p \mu_p F)$ are equal to or larger than the specimen thickness d . If either $X_n < d$ or $X_p < d$ or both are smaller than d , then space charge will form due to accumulated trapped carriers, making the applied field nonuniform.

To illustrate the space charge effects, we will consider a specific case in which $\mu_p \tau_p > \mu_n \tau_n$ and $X_n < d$. This implies that trapped electrons tend to increase the field near the anode and to decrease the field near the cathode, as shown in Figure 7-52. A high field zone is formed near the anode; its width can be assumed to be equal to the electron drift length in the zone, which is given by

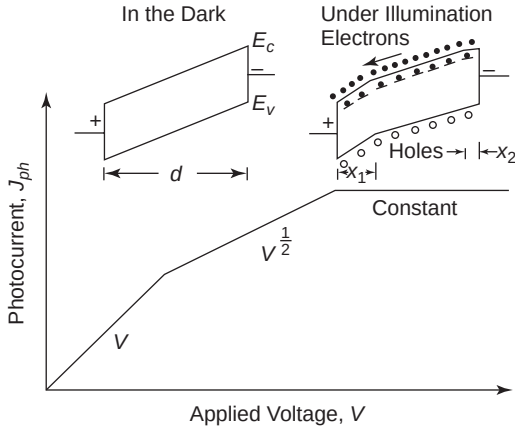


Figure 7-52 The variation of the photocurrent with applied voltage in the case of noninjecting blocking contacts and $\mu_p \tau_p \gg \mu_n \tau_n$.

$$X_1 = \tau_n \mu_n F_1 \quad (7-469)$$

There is also a low field zone. Its width is

$$X_2 = \tau_n \mu_n F_2 \quad (7-470)$$

where F_1 and F_2 are, respectively, the fields in zone 1 and zone 2. At low applied voltages, X_1 is small and F_1 is not much different from V/d . Then J_{ph} follows Ohm's law

$$J_{ph} = qG_p \mu_p \tau_p V/d \quad (7-471)$$

The contribution of electrons is neglected here because $\mu_p \tau_p$ is assumed to be much larger than $\mu_n \tau_n$. At higher voltages, X_1 increases. Assuming that the major portion of the applied voltage is across X_1 , then

$$F_1 \approx V/X_1 \quad (7-472)$$

From Equation 7-415 with $G_p X_p \gg G_n X_n$, it can easily be shown that

$$J_{ph} = qG_p (\mu_p \tau_p)^{1/2} V^{1/2} \quad (7-473)$$

At this particular voltage range, J_{ph} is proportional to $V^{1/2}$. At still higher voltages, X_1 will extend to the cathode and $X_1 \approx d$, and the limiting photocurrent becomes

$$J_{ph} = qG_p d \quad (7-474)$$

The J_{ph} - V characteristics for these three voltage ranges are shown schematically in Figure 7-52. For more details, see reference.¹⁵⁹

To culminate a clear understanding of photoconduction processes and space charge effects, we will consider a simple case: a photoconductor having a set of recombination centers located at E_r below E_c . If an exciting light with energy about $E_c - E_r$ is used to illuminate the photoconductor specimen, we have only free electrons as the dominated carriers. If these photogenerated electrons are replenished at the cathode, then the photocurrent can be written as

$$\begin{aligned} J_{ph} &= qG_n \tau_n \mu_n \frac{V}{d} \\ &= qG_n d \frac{\tau_n}{t_m} \\ &= qG_n d g_{ph} \end{aligned} \quad (7-475)$$

where t_m is the transit time of the electrons and $g_{ph} = \tau_n/t_m$ is the photoconductive gain, which is the quantum efficiency defined in Section 7.10 (i.e., the number of photogenerated electrons passing through the photoconductor per absorbed photon). So g_{ph} can be expressed as

$$g_{ph} = \frac{\tau_n}{t_m} = \mu_n \tau_n \frac{V}{d^2} \quad (7-476)$$

Equation 7-476 indicates that g_{ph} can be made very large either by increasing V or decreasing d if μ_n and τ_n are assumed to be constant. If there is no replenishment of electrons from the cathode, the maximum gain is $g_{ph} = 1$ when $\tau_n = t_m$ since t_m decreases with increasing V .

When a photoconductor is used as a photodetector, its performance is determined by two important parameters: photoconductive gain g_{ph} and the speed of response (or response time) τ_{rn} . Thus, for good photodetectors, the $g_{ph} \Delta B$ product must be large. This criterion is, in fact, similar to that for good amplifiers. ΔB is the bandwidth equivalent to the amplifier passband width, which is directly related to the response time τ_{rn} and can be expressed as

$$\Delta B = \frac{1}{\tau_{rn}} \quad (7-477)$$

It can be seen from Equations 7-476 and 7-477 that g_{ph} can be made larger than 1 if the photogenerated electrons can be replenished at the cathode. But when g_{ph} increases, ΔB will

decrease because the response time increases, implying that the $g_{ph}\Delta B$ product has a tradeoff nature: it is necessary to make g_{ph} increase more than ΔB decreases. We shall discuss briefly two simple cases.

A Photoconductor without Traps

In the ohmic region, $g_{ph} = \mu_n \tau_n V/d^2$ and $\Delta B = 1/\tau_{rn} = 1/\tau_n$. Thus, we have

$$\begin{aligned} g_{ph}\Delta B &= \mu_n V/d^2 \\ J_{ph} &= qG_n \tau_n \mu_n V/d = qG_n d g_{ph} \end{aligned} \quad (7-478)$$

At the onset of the space charge limited current (SCLC) region (the upper limit of the performance), $t_{in} = \tau_d$ and $\Delta B = 1/\tau_{rn} = 1/\tau_n$. Thus, we have

$$g_{ph}\Delta B = \frac{\tau_n}{t_{in}} \frac{1}{\tau_n} = \frac{1}{\tau_d} \quad (7-479)$$

Since $\tau_d = \epsilon/\sigma$ photoconductors with a high resistivity have a high value of dielectric relaxation time τ_d . For $V > V_\Omega$, the value of τ_d decreases with increasing applied voltage V , as shown in Figure 7-53(a). This implies that high-gain photoconductors are usually very sluggish in response. For $V > V_\Omega$, the SCL current sets in and the injected carriers become dominant; the photosensitivity will decrease.

A Photoconductor with Shallow Traps

In the ohmic region, $g_{ph} = \tau_n/t_{in}$ and $\tau_{rn} = \tau_n(1 + n_t/n)$. Thus, we have

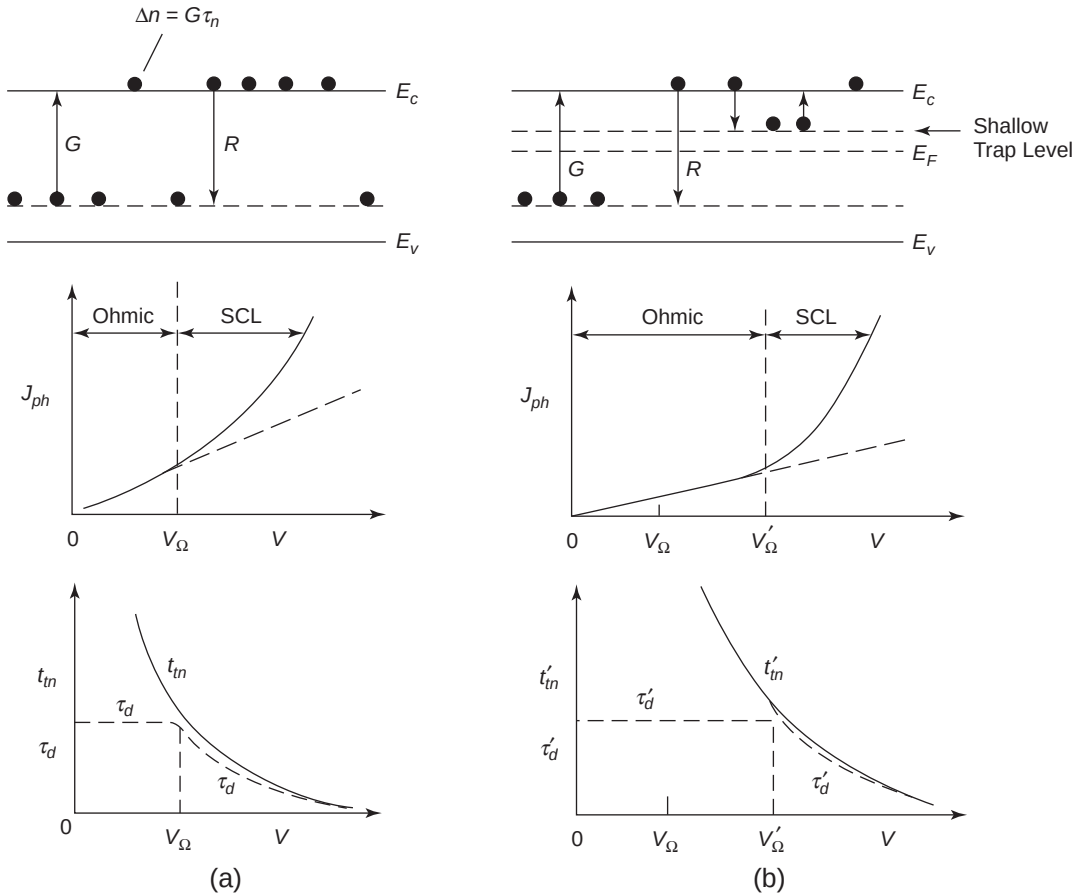


Figure 7-53 Schematic diagrams illustrating the variation of J_{ph} , t_{in} , and τ_d with applied voltage V for (a) a photoconductor without traps and (b) a photoconductor with shallow traps.

$$g_{ph}\Delta B = \frac{\tau_n}{t_{in}} \frac{1}{\tau_n(1+n_t/n)} = \frac{1}{t_{in}(1+n_t/n)} \quad (7-480)$$

With shallow traps, the response time and transit time increase; hence, the $g_{ph}\Delta B$ product decreases. At the onset of the SCL region, $t_{in} = \tau_d$. Since t_{in} increases, τ_d also increases. Therefore, the threshold voltage V_Ω for the onset of the SCL region increases, as well, as shown in Figure 7-53(b).

Light Intensity Dependence

The relationship between photoconductivity σ_{ph} and the light intensity I can be expressed in the form

$$\sigma_{ph} = AI^m \quad (7-481)$$

where A is a constant and m is the light intensity exponent, whose values are generally in the range of $0.5 \leq m \leq 1.0$. The photoconductivities of many materials, such as Sb_2S_3 , Sb_2SeS_2 , ZnTe , MgI , ZnSe , Zn -doped silicon, and amorphous silicon, have been observed to follow this relationship. In the steady state, the carrier generation rate is always equal to the sum of the recombination rates of all recombination channels. According to Rose,⁷³ an exponent m lying between 0.5 and 1.0 (but not equal to 0.5 or 1.0) requires a distribution of trapping states in energy in the band gap. Rose was the first to derive an expression to explain the relationship of Equation 7-481, based on the following assumptions:

The density of empty recombination centers in the dark is negligibly small.

The trapping states are exponentially distributed in energy following

$$N_t(E_t) = B \exp\left(-\frac{E_c - E_t}{kT_c}\right) \quad (7-482)$$

where B is the preexponential factor, E_t is the trapping state energy level, and T_c is the characteristic constant for exponential trap distribution (see Equation 7-106). It is assumed that the photon energy $h\nu \geq E_g$ and that the dominant carriers for photoconduction are electrons.

Rose's analysis leads to the relationship between free electron density n and the photo-generation rate G_n as⁷³

$$n = (G_n N_c^{T/T_c} / BkT_c \langle v\sigma_n \rangle)^{T_c/(T+T_c)} \quad (7-483)$$

Since $T_c \geq T$, the exponent $T_c/(T + T_c)$ lies between 0.5 and 1.0, which is m . The exponent m is clearly temperature dependent. This model implies that for photoconductors following the relationship of Equation 7-481, the energy distribution of the gap states in the material is exponential or at least continuous.

Several investigators have also analyzed the light intensity dependence of photoconductivity, based on the charge neutrality condition with the gap states assumed to be distributed in energy but not necessarily exponentially²¹⁷ or based on the assumption that each localized gap state has three possible charge values: neutral, negative, or positive.^{218,219} Schellenberg and Kao²²⁰ have generalized Rose's original expression in two ways:

1. By extending the contribution of gap states to recombination from those above the midgap states to any states between electron and hole quasi-Fermi levels
2. By including the effect of the asymmetry of gap state distributions and the charge neutrality condition

This analysis has shown that the odd values of the exponent m may occur in solids with gap states distributed discretely or continually.²²⁰ For more details about various analyses of the photoconductivity–light intensity relationship, see references.^{73,99,217–221} Obviously, through measuring photoconductivity as a function of light intensity at various temperatures, it is possible to determine the distribution profile of the gap-state density by carefully analyzing the experimental data.

For some materials, such as CdS , the photocurrent–light intensity ($J_{ph} - I$) relationship at a certain fixed temperature may be superlinear, that is, the exponent m becomes greater than 1 over a small range of light intensities. Beyond this range, the $J_{ph} - I$ relationship is linear, that is, $m = 1$.^{73,99,221}

The conditions necessary for superlinear photoconductivity to occur are as follows:

- There is a set of electron traps of a sufficiently large concentration; the traps are essentially unoccupied in the dark (E_{Fn} located about E_t) but become fully occupied by electrons with increasing photogeneration rate (E_{Fn} located above E_t). Then, the trapping action of these electron traps is changed to a recombination action. In other words, if all traps are occupied, the free carrier concentration, and hence the carrier lifetime, increases.
- There must not be a set of hole traps fully occupied by holes.
- There must be a set of recombination centers located above the hole demarcation level E_{Dp} to capture the holes. This set of recombination centers acts to sensitize the photoconductor, that is, to increase the electron lifetime. Rose⁷³ has called this action sensitization by “electronic doping.”

While these traps are being converted to recombination centers, the electron lifetime is continuously increasing. So, the photocurrent increases superlinearly with increasing light intensity. After the conversion is complete, the photocurrent again increases linearly with light intensity.^{73,99,221}

Light Wavelength Dependence

By measuring the steady-state photocurrent as a function of the wavelength of the exciting light, we can obtain information about trapping levels and the possible nature of traps. The photocurrent spectra usually are measured after the dark currents have reached their steady-state values. In this section, we will describe briefly a typical example to demonstrate the use of the photocurrent spectra for the study of the polyimide properties.

Kan and Kao²²² have used absorption and photoconduction spectra to study the mechanisms responsible for the four absorption peaks in ultraviolet absorption and the corresponding photocurrent quantum efficiency spectra for polyimide films fabricated at various curing temperatures. The basic experimental arrange-

ment for such measurements is shown in Figure 7-54(a). Polyimide (PI) has been one of the most important polymers, with potential for electronic applications because of its good electrical properties.^{223–225} PI is formed mainly by imidization of polypyromellitic acid (PAA). The properties of PI depend on the degree of imidization, which is determined by the curing temperature.²²⁶ The degree of imidization is defined as 100% at the curing temperature ($T_{cu} = 350^\circ\text{C}$) and 10% at the curing temperature ($T_{cu} = 135^\circ\text{C}$). So samples *a* and *A* correspond to 10% imidization ($T_{cu} = 135^\circ\text{C}$); samples *b* and *B* correspond to 67% imidization ($T_{cu} = 150^\circ\text{C}$); samples *c* and *C* correspond to 96% imidization ($T_{cu} = 200^\circ\text{C}$); and samples *d* and *D* correspond to 100% imidization ($T_{cu} = 350^\circ\text{C}$). The absorption coefficient α spectra for samples *a*, *b*, *c* and *d* are shown in Figure 7-54(b), and the corresponding photocurrent quantum efficiency g_{ph} spectra in Figure 7-54(c).

The absorption coefficient is defined as

$$\alpha = \alpha_o NM \quad (7-484)$$

where α_o is the absorption coefficient per molecule with imide rings, N is the concentration of molecules (including PI and PAA molecules), and M is the degree of imidization. The photocurrent quantum efficiency g_{ph} is defined as the ratio of the number of charge carriers producing the photocurrent generated by photoexcitation to the number of photons absorbed by the material during the same period of time. It can be expressed as

$$g_{ph} = \frac{J_{ph}}{qI_o T_r} \quad (7-485)$$

where I_o is the number of photons per unit area illuminating the sample (light intensity in joules/sec divided by photon energy $h\nu$ in joules) and T_r is the transmittance of the illuminated electrode.

Of the four ultraviolet absorption peaks and four corresponding photocurrent quantum efficiency peaks for PI, the absorption peaks at 4.35 and 6.40 eV (corresponding to the quantum efficiency peaks at 4.05 and 5.80 eV) are due to intramolecular transitions, and the

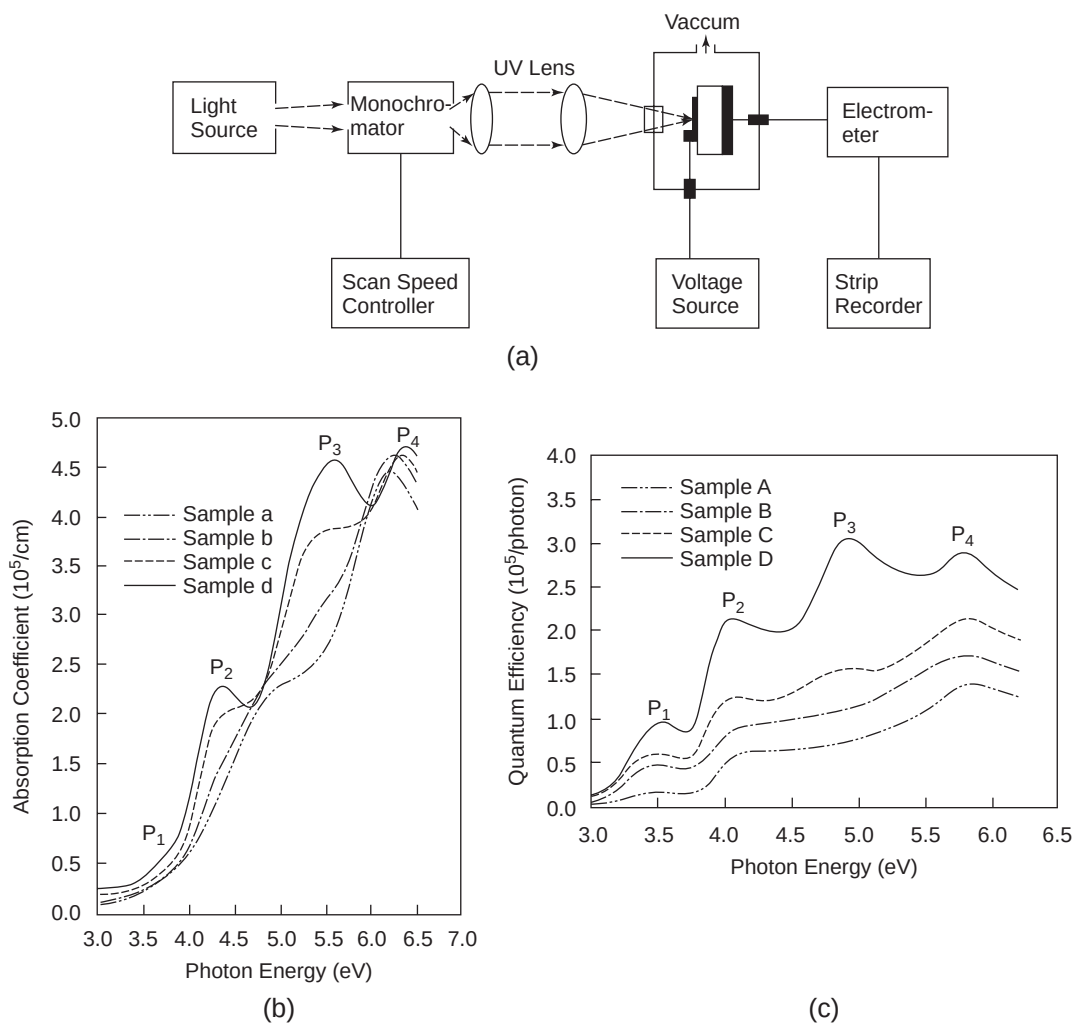


Figure 7-54 (a) Basic experimental arrangement for photocurrent measurements; (b) the absorption spectra of PI samples *a*, *b*, *c*, and *d*; and (c) the quantum efficiency spectra of PI samples *A*, *B*, *C*, and *D* for the illuminated electrode at the positive polarity and an applied field of $1.17 \times 10^5 \text{ V/cm}$.

absorption peaks at 3.65 and 5.65 eV (corresponding to the quantum efficiency peaks at 3.50 and 4.90 eV) are due to intermolecular transitions. The energy band gap of PI is about 7 eV²²⁷; the possibility for photogeneration due to band-to-band transition via a single photon process can be ruled out for the range of photon energies used for the experiment. Thus, the photogeneration is a multiphoton process associated with the formation of charge-transfer (CT) complexes and the dissociation into free carriers under applied electric fields

and thermal stress. In analyzing the experimental results, Kan and Kao have concluded that the effects of curing temperature on absorption spectra and quantum efficiency spectra are due mainly to changes of molecular orders between different polyimide chains rather than to imidization.

Temperature Dependence

Many parameters that influence the behavior of photoconduction are temperature dependent.

One such parameter is the energy band gap, which for most materials decreases with increasing temperature and increases with increasing pressure. At a constant pressure, the variation of the band gap with temperature always involves the temperature-caused expansion or contraction of the material.

The temperature dependence of the band gap may be expressed as

$$E_g = E_{g0} + \beta T \quad (7-486)$$

For most materials, the parameter β is negative (except for some materials such as PbS, PbSe, and PbTe with positive β). For example, the band gap of ZnS crystals changes from 3.6 eV at room temperature to about 3.4 eV at 300°C.^{228,229} Obviously, the optical absorption edge and hence the photoconductivity peaks will shift toward longer wavelength as the temperature is increased.

The temperature dependence of J_{ph} is associated with the temperature dependence of the carrier mobilities (μ_n and μ_p) and carrier concentrations (n and p), which are also dependent on the concentration, location, and distribution of traps or recombination centers. Equation 7-483 indicates clearly that the $J_{ph} - I$ relationship is temperature dependent.

Effects of Surface Conditions and Ambient Atmosphere

The surface states of the photoconductor can strongly influence photoconduction behavior. If the surface recombination rate R_s is small compared to the bulk recombination rate R_B , the carriers produced in a thin surface layer by strongly absorbed light will recombine slowly through the surface states, and under the action of an applied field of sufficient magnitude, excess photocarriers will be drawn to the bulk to contribute to the photocurrent. But if the surface recombination rate is large compared to the bulk recombination rate, the carriers produced in a thin surface layer by strongly absorbed light will recombine through the surface states so rapidly that they cannot contribute much to the photocurrent. In this case, only the carriers produced by weakly absorbed

light (large penetration depth) and produced at a region far from the surface can contribute to the photocurrent, since short-wavelength light is strongly absorbed in the surface region, whereas light of longer wavelength is more strongly absorbed in the bulk. For $R_s \ll R_B$, the photocurrent spectrum has a good correlation with the absorption spectrum, while for $R_s \gg R_B$ this correlation disappears.

It is well known that the presence of moisture or air, oxygen, or other gases will influence photoconductivity and its spectral response in organic and inorganic crystals.^{211,212,230} The adsorbed gases on the surface give rise to the formation of deep traps, thus increasing the surface recombination rate. O₂ and iodine, for example, would form acceptor-type electron traps, while H₂ and NH₃ would form donor-type hole traps.

7.12.3 Homogeneous and Nonhomogeneous Photoconduction

In this section we shall discuss homogeneous and nonhomogeneous photoconduction and some junction photoconduction devices.

Homogeneous Photoconduction

Homogeneous photoconduction implies that the spatial distribution of photogenerated carriers is uniform in the space between two electrodes. There are five basic types of photoconductors, depending on the freedom of the photogenerated carriers and whether the carriers are replenished at the electrodes. We shall discuss each briefly.

Both electrons and holes are mobile and are replenished at the cathode and at the anode, as shown in Figure 7-55(a). For this type of photoconductor, the lifetime is

$$\tau_n = \tau_p = \tau \quad (7-487)$$

and the photoconductive gain is

$$g_{ph} = (\mu_n \tau_n + \mu_p \tau_p) \frac{V}{d^2} = (u_n + u_p) \tau \frac{V}{d^2} \quad (7-488)$$

Electrons and holes are both mobile, but only one type of carrier (e.g., electrons) is replen-

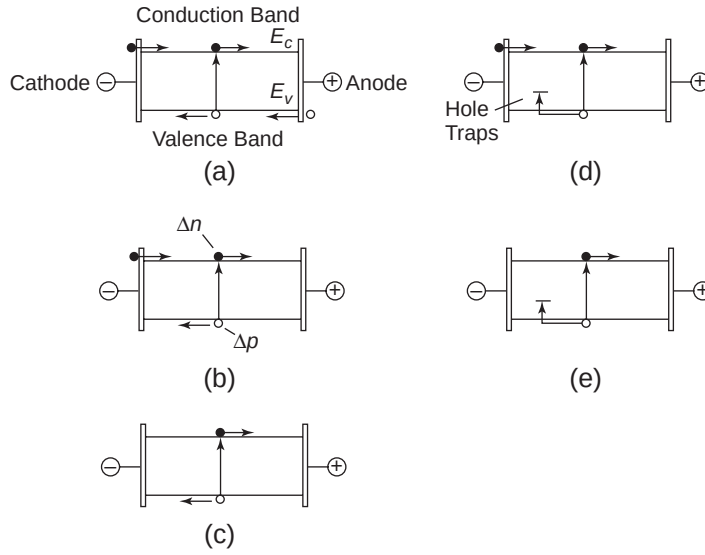


Figure 7-55 Schematic diagrams showing five basic types of photoconductors for homogeneous photoconduction: (a) both electrons and holes are replenished from electrodes; (b) only one type of carrier (e.g., electrons) is replenished from the cathode; (c) neither electrons nor holes are replenished from electrodes; (d) only one type of carrier (e.g., electrons) is mobile and replenished from the cathode, with photogenerated holes being captured in hole traps; and (e) only one type of carrier (e.g., electrons) is mobile but is not replenished from the cathode, and holes are trapped in hole traps.

ished at the cathode, as shown in Figure 7-55(b). At sufficiently high fields, the holes contribute to photocurrent until they are completely drawn off to the cathode before taking part in recombination. For applied fields larger than that required to produce saturated photocurrents, the lifetime of the holes is equal to the transit time of the holes and also equal to the lifetime of the electrons, because of the charge neutrality condition. Thus, we can write

$$\tau_p = \tau_n = t_{tp} = \frac{d^2}{\mu_p V} \quad (7-489)$$

$$t_{tn} = \frac{d^2}{\mu_n V} \quad (7-490)$$

$$g_{ph} = \frac{\tau_n}{t_{tn}} + \frac{\tau_p}{t_{tp}} = \frac{\mu_n + \mu_p}{\mu_p} \quad (7-491)$$

Electrons and holes are both mobile, but neither is replenished at the electrodes, as shown in Figure 7-55(c). In this type of photoconductor, a saturation of the photocurrent occurs when the applied field is sufficiently high to draw off both electrons and holes before they

recombine with each other. This phenomenon is generally referred to as *primary photoconduction*. Primary photoconduction is terminated when the minority carrier lifetime is terminated and is equal to the majority carrier lifetime. Thus, we can write

$$g_{ph} = \frac{G_n(\tau_n/t_{tn}) + G_p(\tau_p/t_{tp})}{G_n + G_p} \quad (7-492)$$

For primary photoconduction, $\tau_n = t_{tn}$, $\tau_p = t_{tp}$ so the maximum photoconductive gain $g_{ph} = 1$ and J_{ph} becomes saturated.

There is also the so-called *secondary photoconduction*, which involves carrier replenishment from electrodes. Secondary photoconduction is terminated only when the majority carrier lifetime is terminated, so the majority carrier lifetime may be much larger than the minority carrier lifetime. According to Equation 7-492, the photoconductive gain may be much greater than unity.

Only one type of carrier is mobile. For example, the holes are immobile and the electrons are mobile and replenished at the

cathode, as shown in Figure 7-55(d). This is a kind of secondary photoconduction. In this case, the minority holes generated by light are captured almost immediately by traps and become immobile. Only the majority electrons contribute to the photocurrent, so the photoconductive gain can be written as

$$g_{ph} = \frac{\tau_n}{t_{tr}} = \mu_n \tau_n \frac{V}{d^2} \quad (7-493)$$

The electron lifetime is terminated by recombination with the trapped holes. Such recombination via traps is much more likely than direct band-to-band recombination.

Only one type of carrier (e.g., electrons) is mobile, but the mobile electrons are not replenished at the cathode, as shown in Figure 7-55(e). In this type of photoconductor, the photocurrent decays with time even under light excitation because of polarization due to the space charge in the material. Thus, it is not possible to maintain a steady photocurrent.

Nonhomogeneous Photoconduction

Nonhomogeneous photoconduction implies that there is a barrier to regulate the flow of charge carriers. There are three basic types of nonhomogeneous photoconductors, depending on the nature of the barrier.

P-N and P-I-N Junction Photodiodes

As shown in Figure 7-56(a), the potential barrier height at the p-n junction in the dark at zero bias is given by

$$qV_D = kT \ln(n_p/n_n) = kT \ln(p_n/p_p) \quad (7-494)$$

where n_n and n_p are respectively, electron concentrations on the n-side and the p-side, and p_n and p_p are, respectively, the hole concentrations on the n-side and the p-side. If an exciting light illuminates the junction region, electron-hole pairs are generated. Under the action of the internal field at the junction, the electron-hole pairs are separated, the electrons moving to the n-side and the holes moving to the p-side. When the p-n junction is reverse-biased, photocurrent will flow. This is a good example of

primary photoconduction, because there is no carrier replenishment from the contacts and the maximum photoconductive gain $g_{ph} = 1$.

P-i-n photodiodes are similar to p-n junction diodes. The basic difference is that the p-i-n structure has a thick near-intrinsic layer sandwiched between heavily doped n and p layers, so the i region is completely depleted in the dark. With the exciting light illuminating the region, electron-hole pairs are generated inside the depletion layer and are separated by the applied field, giving rise to photocurrent. However, if a sufficiently high bias voltage is applied to the p-i-n photodiode to cause impact ionization and carrier multiplication in the depletion region, the p-i-n diode becomes an avalanche photodiode (due to avalanche multiplication resulting from impact ionization to create electrons and holes in the i regions). In this case, the photoconductive gain may be larger than 1 because of the carrier multiplication process.

NPN or PNP Phototransistors

A phototransistor can have high photoconductive gains through transistor action.¹⁶ Take an N_1PN_2 phototransistor as an example. The electron-hole pairs photogenerated in the junction regions will be separated by the action of the internal field at the junctions. At the collector junction, the electrons will flow from the p-base to the N_2 collector in a manner similar to the normal p-n junction. In this case, however, the electrons are replenished from the N_1 emitter to the base through the emitter junction, as shown in Figure 7-56(b). Furthermore, the presence of holes in the base reduces the barrier height of the emitter junction, increasing the electrons injected from N_1 emitter to the p-base. NPN and PNP phototransistors are good examples of secondary photoconduction. The photoconductive gain of phototransistors can reach as high as 10^3 if the emitter, base, and collector are properly doped.

Metal-Semiconductor Schottky Barrier Photodiodes

In Potential Barrier Height and the Schottky Effect in Chapter 6, we discussed metal-

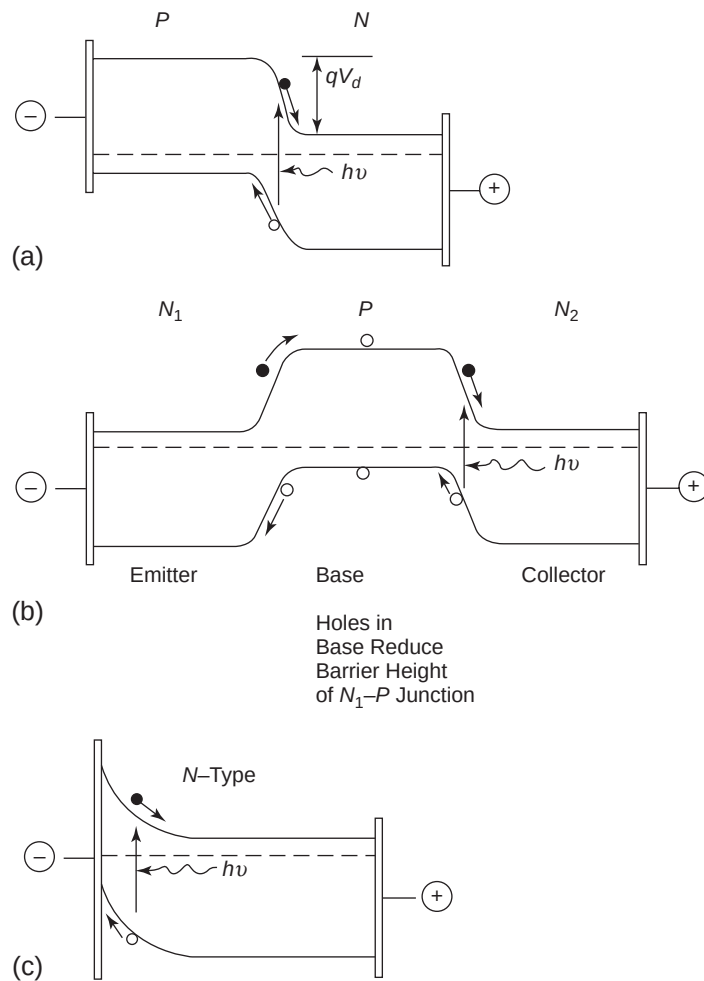


Figure 7-56 Schematic diagrams showing three basic types of photoconductors for nonhomogeneous photoconduction: (a) p-n junction photodiodes, (b) n-p-n phototransistors, and (c) metal-n-type semiconductor Schottky photodiodes.

semiconductor contacts in some detail. The basic metal-n-type semiconductor contact is shown in Figure 7-56(c). Photogenerated carriers in the barrier region will flow under the reverse-bias condition. But the presence of the barrier prevents the replenishment of carriers from the contact. So the example is of primary photoconduction, with the photoconductive gain $g_{ph} < 1$.

7.12.4 Photoresponse Times

Photoresponse time is the time required for the material to respond to the action of photoexci-

tation, usually the time τ_r required for the photocurrent to reach its steady-state value. For materials with traps, the response time is larger than the carrier lifetime but tends to approach the carrier lifetime at high temperatures. When light excitation is switched on, photogeneration not only supplies carriers to the bands, but also pours carriers to the trapping centers. When the light excitation is switched off, time is required not only for the free carriers to recombine through recombination channels, but also for the trapped carriers to be detrapped then recombine via thermal excitation and subsequent recapture and recombination processes. For

perfect crystals without traps, the photoreponse time τ_r is equal to the carrier lifetime, as discussed in Section 7.11.3. But for photoconductors or insulators with trapping and recombination centers, the response time and the carrier lifetime relationship follows Equations 7-464 and 7-465.

We will use an n-type photoconductor containing deep traps of concentration N_t located at E_b , as shown in Figure 7-57(a), to demonstrate the effect of traps on response time. It is assumed that in the dark, all donors of concentration N_d are ionized and most thermally generated electrons fall into the traps. If $N_d > N_t$, there are $n_o = N_d^+ - N_{t_o}^-$ free electrons in the conduction band and the traps are completely filled, $n_t = N_{t_o}^-$. Now, if this photoconductor is illuminated with a light pulse of photon energy $(E_c - E_t) \leq h\nu < (E_t - E_v)$, as shown in Figure 7-57(b) and (c), then the trapped electrons at E_t will be excited to the conduction band. This produces photogenerated electrons Δn and hence photocurrent, as shown in Figure 7-57(d). After the photocurrent has reached its steady-state value, $n = n_o + \Delta n$, and $N_t = N_t^- + N_t^o$, where N_t^- and N_t^o are, respectively, the occupied and empty trap concentrations. During the transient period, the change of free electron concentration can be written as

$$\begin{aligned} \frac{dn}{dt} &= \text{photogeneration rate} \\ &\quad + \text{thermal-generation rate} \\ &\quad - \text{recombination rate} \\ &= \sigma_i(N_t^-)I + \langle v\sigma_n \rangle n_t(N_t^-) - \langle v\sigma_n \rangle n(N_t^o) \end{aligned} \quad (7-495)$$

where $\sigma_i = \eta_n \alpha / N_t$, η_n and α are, respectively, the quantum yield and the absorption coefficient. In the following, we shall discuss the variation of Δn with time.

During the rise period: We use the initial condition, (i.e. when $t = 0$, $\Delta n = 0$). The solution of Equation 7-495 yields

$$\Delta n(t) = (\Delta n)_{st} [1 - \exp(-t/\tau_{on})] \quad (7-496)$$

During the decay period: After Δn has reached its steady-state value $(\Delta n)_{st}$, the light excitation is switched off at $t = t_L$ and $\Delta n(t)$ starts to decay with time. We use the bound-

ary condition when $t' = 0$, $I = 0$, and $\Delta n = (\Delta n)_{st}$. Then, the solution of Equation 7-495 gives

$$\Delta n(t') = (\Delta n)_{st} \exp(-t'/\tau_{off}) \quad (7-497)$$

where

$$\tau_{on} = [\langle v\sigma_n \rangle (n_1 + N_t^o + n_o) + \sigma_i I]^{-1} \quad (7-498)$$

$$\tau_{off} = [\langle v\sigma_n \rangle (n_1 + N_t^o + n_o)]^{-1} \quad (7-499)$$

and

$$n_1 = \frac{N_c N_t}{N_d} \exp[-(E_c - E_t)/kT] \quad (7-500)$$

The variation of Δn with time is shown in Figure 7-57(d), and the variation of τ_{on} and τ_{off} with light intensity in Figure 7-57(e). From Equations 7-498 and 7-499 we have

$$\frac{1}{\tau_{on}} - \frac{1}{\tau_{off}} = \sigma_i I \quad (7-501)$$

So, σ_i can be interpreted as the photoionization cross-section. For details, see reference.²⁰⁸

If this photoconductor is illuminated with a light pulse of photon energy $h\nu \geq E_g$, as shown in Figure 7-57(f), then the situation is quite different. In this case, there are two portions of the relaxation curves: one fast and the other slow. During the rise period, the light excitation produces Δn and Δp very fast, before the free electrons start to fall into traps. This short period of time is termed the *fast portion*. As soon as the electrons fall into the traps and recombine with holes, the net increase rate of the photogenerated electrons becomes smaller, and the combined trapping and detrapping processes make the change of Δn with time slow, resulting in the formation of the slow portion, as shown in Figure 7-57(g), (h), and (i).

Similarly, during the decay period, free electrons and free holes will recombine very fast, forming the fast portion. Afterward, the trapped electrons will gradually be released by thermal excitation and recombine with holes, and the whole system will gradually return to its original thermal equilibrium state. This process is slow, and this is why this period of time is termed the *slow portion*. However, relaxation time is usually longer during the decay period than during the rise period. For more details,

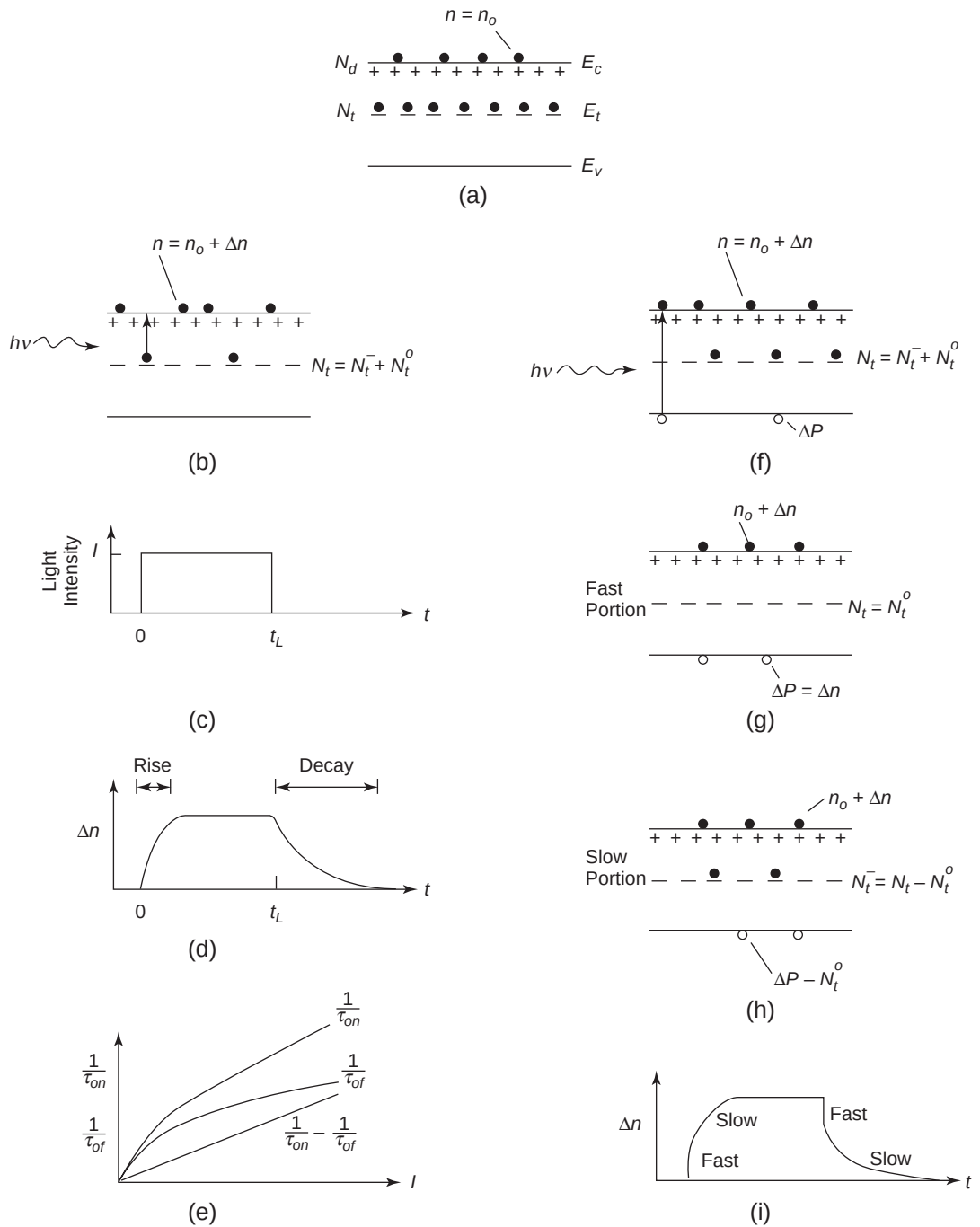


Figure 7-57 Schematic diagrams illustrating the variations of Δn with time during the rise and decay periods for exciting light with photon energy $E_c - E_t \leq h\nu < E_t - E_v$, and that with photon energy $h\nu \geq E_g$. The description is given in the text.

see references.^{73,99,208} From this example, it can be seen that relaxation time and response time are dependent on the exciting light intensity and on the photon energy.

7.13 Photosensitization

Photosensitivity is defined as the photoconductivity per unit exciting light intensity. Photosensitization is a process that increases the photosensitivity of a photoconductor or converts an insensitive photoconductor to a sensitive one. This can be done by incorporating suitable impurities into the photoconductor to form a set of localized states, which have the capability of capturing minority carriers and decreasing the probability of capturing the majority carriers, thus increasing the sensitivity of the photoconductor. Such impurities are generally referred to as the *activators* or *activating* (or *sensitizing*) *centers*. Generally, activators increase photosensitivity but tend to decrease photoresponse speed—in other words, they increase the response time.⁷³ Here, we shall discuss briefly the basic concept of sensitization.

Consider an insensitive photoconductor, as shown in Figure 7-58(a). In this photoconductor, there is only one type of fast recombination center (type I) located at E_{r1} below E_c and between the electron and hole demarcation levels E_{Dn} and E_{Dp} . It is obvious that the presence of such type I recombination centers will make the photogenerated carriers recombine fast through the centers, thus reducing the carrier lifetimes. This implies that this photoconductor is insensitive to photoexcitation; hence, its photoconductive gain is small.

Now, if suitable impurities were incorporated into this photoconductor, creating type II impurity centers located at E_{r2} , these type II centers might not change the behavior of the photoconductor if the centers were located below the hole demarcation level E_{Dp} , as shown in Figure 7-58(b). However, if the impurities were so chosen such that the type II centers were located at E_{r2} but above the hole demarcation level E_{Dp} , as shown in Figure 7-58(c), then the

type II centers would act as sensitizing centers. This is because the holes captured by the type II centers have a longer lifetime in the centers than the holes captured by the type I centers and the type II centers have a small capture cross-section for recombining with free electrons. Furthermore, the type II centers would become occupied primarily by holes, implying that the electrons initially in type II centers are effectively transferred to type I centers, so the lifetime of free electrons would be increased, because they would encounter mostly centers with a small capture cross-section and only a few centers with a large capture cross-section. Since the incorporation of type II centers sensitizes the photoconductor, this incorporation is called *electronic doping*.

However, it should be noted that the sensitization is effective only when the concentration of both type I and type II centers is much larger than the concentration of free carriers. If this is not the case, the incorporation of the Type II centers only provides additional recombination centers without sensitizing effects. It should also be noted that the demarcation levels are dependent on temperature and exciting light intensity. So the location of E_{Dp} , originally located above E_{r2} , can be changed to a level below E_{r2} either by increasing light intensity or by decreasing temperature.

At a fixed temperature, an increase of the exciting light intensity will lower the hole demarcation level E_{Dp} from the energy level E_{r2} of type II centers. This will, in turn, increase the photosensitivity, resulting in the superlinear behavior of the photocurrent–light intensity ($J_{ph} - I$) relationship, that is, the exponent m becomes greater than 1. However, at a fixed light intensity, when the temperature of the photoconductor is increased to such a level that the hole demarcation level E_{Dp} is above E_{r2} , then the type II centers will reduce the sensitivity of the photoconductor. This phenomenon is called the *thermal quenching* of the photoconduction.

At a fixed temperature, if a second light source with photon energy $h\nu \geq (E_{r2} - E_v)$, usually in the infrared region, is used to illuminate the photoconductor, producing more

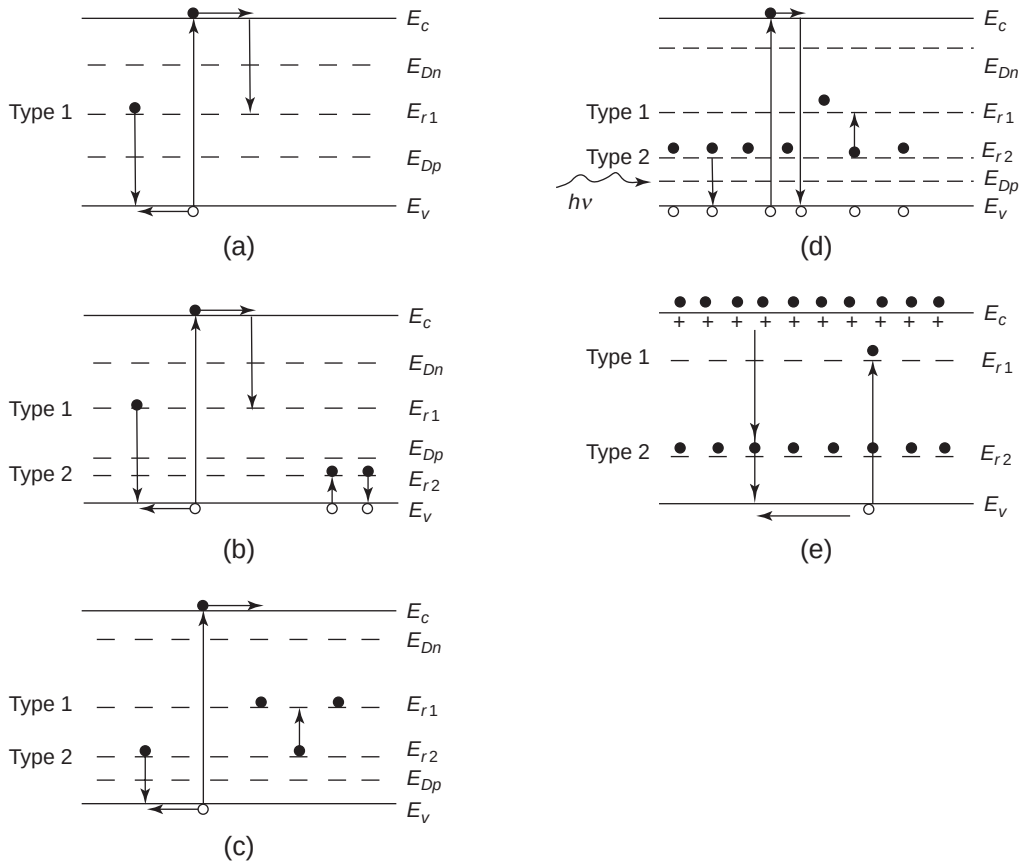


Figure 7-58 Schematic illustration of the concept of sensitization: (a) an insensitive photoconductor containing only type 1 fast recombination centers; (b) the introduction of type 2 centers located below E_{Dp} does not affect the condition of (a); (c) when the location of type 2 centers becomes above E_{Dp} , these centers will become activators, sensitizing the photoconductor and increasing its photoconductivity; (d) a second light of the infrared region may create more holes for recombination with the majority electrons, then type 2 centers become quenching (or poisoning) centers tending to desensitize the photoconductor; and (e) when the light can excite only electrons from E_v to E_{r1} , it produces mainly trapped electrons at E_{r1} and free holes for recombination with the majority electrons. If the thermal excitation rate of trapped electrons is smaller than the recombination rate, the conductivity will decrease, resulting in negative photoconductivity.

holes for recombination with free electrons, then the second light in fact reduces the lifetime of the free electrons and hence the photoconductivity, as shown in Figure 7-58(d). This second light converts the type II centers into *quenching* or *poisoning* centers, thus reversing the sensitizing effect. This phenomenon is referred to as *optical quenching* or *infrared quenching* of the photoconduction.

Light excitation may also produce negative photoconductivity, depending on the photon energy. Let us consider a simple n-type semi-

conductor consisting of naturally thermally generated free electrons of concentration n_o , with type I and type II centers, as shown in Figure 7-58(e). If now we use the exciting light of photon energy $(E_{r1} - E_v) \leq h\nu < (E_c - E_v)$ to illuminate the semiconductor, the light can excite only the electrons from E_v to E_{r1} , creating free holes to recombine with free electrons. This reduces the concentration of the majority electrons, making the conductivity lower than the original dark conductivity. This is why photoconductivity under such light excitation is

called *negative photoconductivity*. For more details, see references.^{73,99}

Sensitization includes not only the sensitizing of photoconductivity at a specific photon energy, but also the sensitizing of spectral sensitivity. In most large-bandgap insulators, such as polymers, only high-energy radiations such as x-rays (or at least ultraviolet light) can be absorbed to produce photocurrent. For most applications, such as electrophotography or xerography, the photoconductors chosen must have a high photosensitivity at visible light wavelengths. For example, poly(N-vinylcarbazole) (PVK) is typically only photosensitive to ultraviolet light. However, in mixtures of PVK and 2, 4, 7 trinitro-9-fluorenone (TNF), the TNF molecules form charge-transfer complexes with the monomer carbazole units in the PVK, with TNF as the electron-acceptors. This complex exhibits an additional absorption band and photoconductivity in the visible region of the spectrum at energies below the absorption edges of the individual components.^{231–235} In inorganic photoconductors, the incorporation of suitable impurities can also provide spectral sensitization. For example, the photoconductivity peak for CdS occurs at a photon wavelength around 5000 Å, but in CdS incorporated with a high concentration of Cu, the photoconductivity peak shifts to a longer wavelength, around 6000 Å, and is higher than the undoped CdS.^{99,236}

7.14 Transient Photoconduction

In Section 7.8, we discussed the current transient phenomena resulting from the injection of a thin sheet of photogenerated carriers under a steady electric field, and also the time-of-flight technique for the measurements of carrier mobilities. In fact, these transient phenomena are photocurrent transient phenomena. For such measurements, a short, strongly absorbed light pulse is usually used to produce a photocurrent in the specimen. The measuring circuit is shown in Figure 7-39(a). Measurements of transient photocurrent, in which the time of photoexcitation is very short compared to the carrier transit time, yield important information

about the behavior of the photoconductor—in particular, the time dependence of the in-transit packet of carriers, which is directly related to the structure of the material.

Modern photocopying machines tend to employ inorganic or organic amorphous materials for the photoconductor, such as *a*-As₂Se₃, *a*-Se, 1:1 TNF-PVK, and ZnO. In such materials, the mutual orientation of and intermolecular distances between the constituent molecules exhibit a distribution that creates the so-called *diagonal* and *off-diagonal disorders*. The carrier transport processes do not follow conventional band conduction law; the electrical transport must be treated as a sequence of hoppings in an assembly of localized sites with fluctuating site energies ΔE and intra- and intermolecular transition matrix elements $\Delta \Gamma$. If $\Delta E/\Delta \Gamma > 1$, charge carriers will be localized. In this case, the tail of the photocurrent $J_{ph}(t)$ is long, indicating a dispersion of the carrier transit time. Furthermore, the shape of $J_{ph}(t)$ curve is invariant with electric field and specimen thickness. This feature, called *universality*, is incompatible with the traditional concept of statistic spreading, that is, moving carriers spread as a propagating Gaussian packet, as shown in Figure 7-39(d).

Scher and Montroll²³⁷ were the first to analyze this dispersive transport behavior using the formalism of continuous-time random walk (CTRW), in which the carriers hop among a random set of localized sites. Non-Gaussian dispersive transport is not restricted to the hopping motion. Several investigators^{238,239} have shown that non-Gaussian dispersion behavior also occurs in conventional multiple-trapping transport. Here, we shall discuss only briefly the concept of dispersive transport processes. For detailed mathematical treatments, see references.^{237,240–244}

The measured photocurrent $J_{ph}(t)$ is simply the space-averaged conduction current, which can be expressed as

$$J_{ph}(t) = \frac{1}{d} \int_0^d j_{ph}(x, t) dx \quad (7-502)$$

Unlike the $J_{ph}(t) - t$ curves shown in Figure 7-40 (based on the Gaussian spreading of the

moving charge sheet), the $J_{ph}(t) - t$ curve for amorphous materials such as α -As₂Se₃ and 1:1 TNF-PVK exhibits a current spike immediately after the onset of the short light pulse. This spike is followed by a soft plateau, a *shoulder*

or *transition region*, and finally a ubiquitous long tail, as shown in Figure 7-59(a). The experimental results are from Scharfe.²⁴⁵ A similar curve has been observed in PVK²⁴⁶ and 1:1 TNF-PVK.^{247,248} The $J_{ph}(t) - t$ relation on a

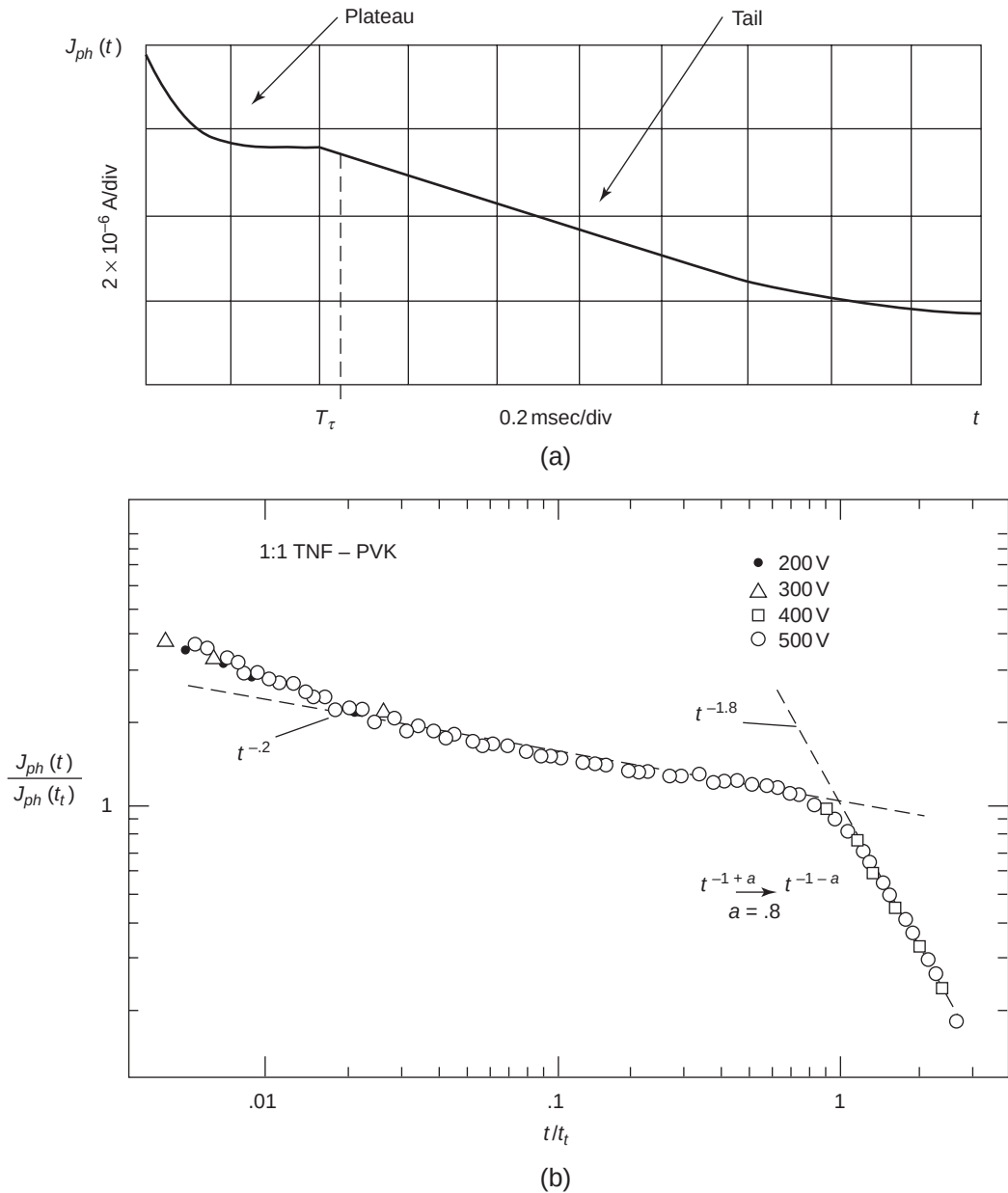


Figure 7-59 (a) The transient photocurrent $J_{ph}(t)$ as a function of time for amorphous α -As₂Se₃ and (b) the $\log(J_{ph}(t)/J_{ph}(t_i)) - \log(t/t_i)$ plot for 1:1 TNF-PVK, for which $J_{ph}(t)$ is normalized by $J_{ph}(t_i)$ and t is normalized by t_i , where t_i is the transit time.

$\log J_{ph} - \log t$ plot for 1:1 TNF-PVK is shown in Figure 7-59(b). The $J_{ph}(t) - t$ curve, plotted in logarithmic units, exhibits features of dispersive transport processes. Next, we shall summarize the essential features of dispersive non-Gaussian electrical transport based on the CTRW treatment by Scher and Montroll.²³⁷

The hopping of carriers in a set of localized sites with a random distribution is governed by a hopping-time distribution function $\Psi(t)$. This distribution function determines the statistics of the carrier transport. For classical Gaussian transport, as shown in Figure 7-40 and Figure 7-60(a), $\Psi(t) \propto \exp(-\lambda t)$, involving only a single transition rate or a single time constant λ^{-1} . In amorphous materials, there is a dispersion in the separation distance between nearest-neighbor localized sites for hopping carriers, and also in the potential barriers between those sites. This is why $J_{ph}(t)$ and $\Psi(t)$ have a long tail. Scher and Montroll have proposed that the transient photocurrent $J_{ph}(t)$ and the hopping-time distribution function $\Psi(t)$ can be expressed in the form

$$\begin{aligned} J_{ph}(t) &\propto t^{-(1-\alpha)} & \text{for } t < t_t \\ J_{ph}(t) &\propto t^{-(1+\alpha)} & \text{for } t > t_t \end{aligned} \quad (7-503)$$

where α is a constant with a value within $0 < \alpha < 1$. The parameter α is directly related to the shape of the transient current-time characteristics.

The carrier transit time t_t can be expressed as²⁴¹

$$t_t = \left[\frac{d}{\ell(F)} \right]^{1/\alpha} \exp(\Delta E_o/kT) \quad (7-504)$$

where d is the specimen thickness, $\ell(F)$ is the average displacement in the field direction between consecutive jumps of the hopping carriers, and ΔE_o is the activation energy.²⁴¹ The traditional definition of the transit time is

$$t_t = d/\mu F \quad (7-505)$$

The traditional t_t is proportional to d , while the transit time based on Equation 7-504 is proportional to $d^{1/\alpha}$ which is clearly superlinear, completely different from Gaussian statistics. Generally, the more the disordered the system

is, the smaller the value of α , the more dispersive the transient photocurrent profile, and the stronger the thickness dependence of t_t .²⁴³

Figure 7-59(b) shows that for 1:1 TNF-PVK, J_{ph} is proportional to $t^{-0.2}$ for $t < t_t$ and J_{ph} is proportional to $t^{-1.8}$ for $t > t_t$. From Equation 7-503, α is equal to 0.8. Extrapolating the two dashed lines in Figure 7-59(b) shows the meeting point that is at $t = t_t$. The experimental results agree well with the theoretical prediction and also provide evidence of the universality of dispersive transport, that is, the $J_{ph}(t)t$ curve is independent of applied field and specimen thickness.

To illustrate the difference between the classical model and the random-walk model for transient photoconduction, we will use two sets of schematic diagrams, shown in Figure 7-60 to show the basic difference between these two models. The illustrations are self-explanatory. In the random walk model (a), the position of the representative carriers in the specimen at $t \approx 0^+(o)$ which are just injected at $x = 0$ as a narrow packet of carriers. The carrier packet spreads and propagates to the right at $t < t_t$ the positions of the carriers are represented by ($\bullet$) as soon as some carriers arrive at the other electrode, the positions of the carriers are represented by ($*$) for $t > t_t$ (b) the spreading of the carrier packet in which 1, 2, 3 corresponding to the positions of the carriers at the time 1, 2, 3 in (a); (c) the recorded transient photocurrent (in linear units), again, 1, 2, 3 correspond to the time $t = 0^+$, $t < t_t$ and $t > t_t$, respectively; and (d) the same transient photocurrent plotted in logarithmic units. As the transit time increases when the applied field is lowered, the $J_{ph} - t$ curves in Figure 7-60(c) and (d) shift to the right, that is, a longer transit time is required for lower applied field.

Several investigators have reported that the transient current profile for electrons generated by an electron beam by injection from Al or Au electrodes to 1:1 TNF-PVK specimen violates the universality criterion for transit-time dispersion.²⁴⁹ They have attributed this disagreement with the Scher-Montroll theory to a field-dependent detrapping mechanism. There are many factors that can affect the transient

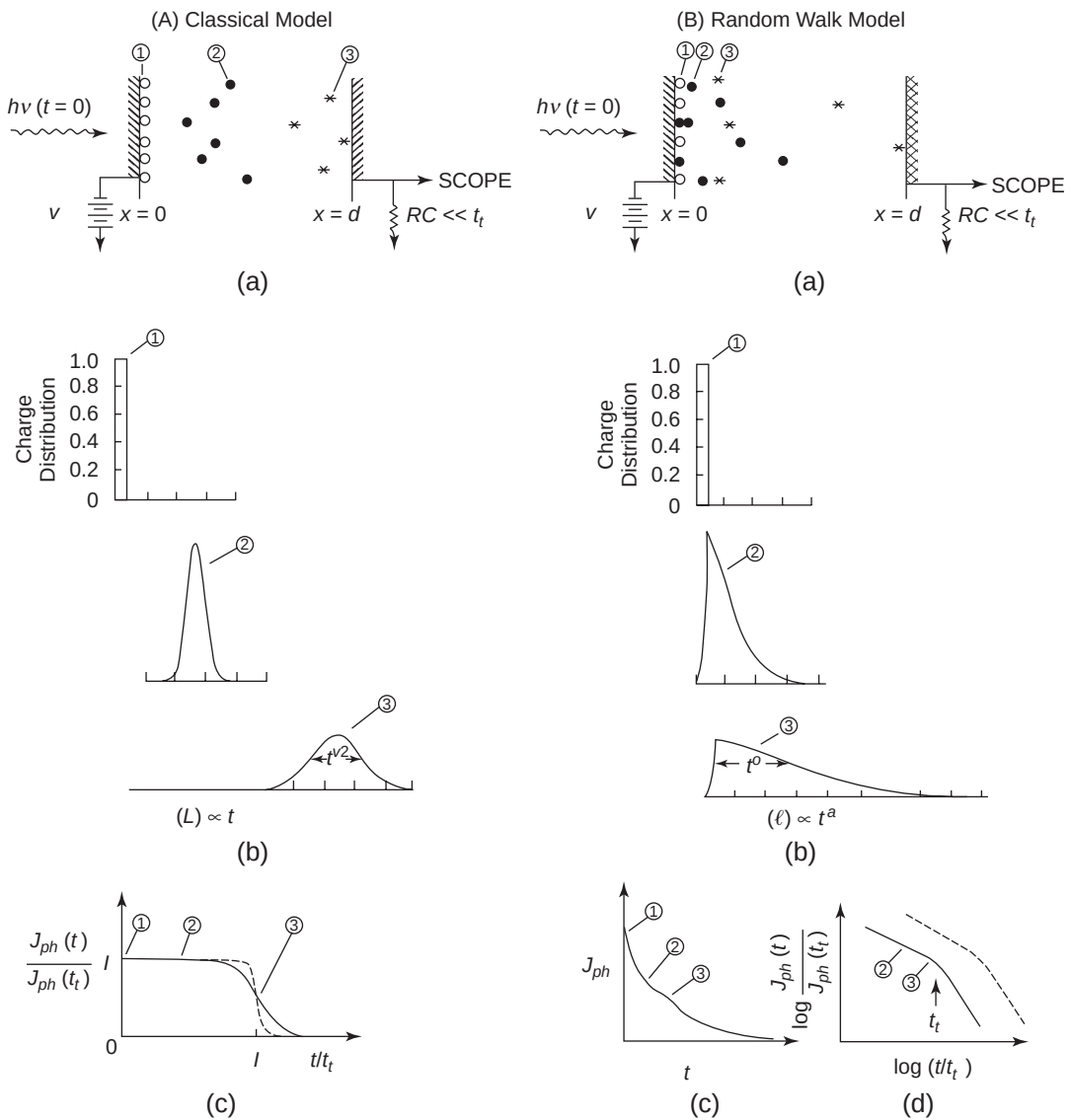


Figure 7-60 The basic differences between (A) the classical model and (B) the random walk model: (a) the positions of the representative carriers in the specimen at $t = 0$ (○), at $0 < t < t_t$ (●), and at $t > t_t$ (*); (b) the spreading of the carrier packet at different times; (c) the transient photocurrent as a function of time in linear units; and (d) the same transient photocurrent plotted in logarithmic units. (1), (2), and (3) denote three different times. Dashed lines represent J_{ph} for a lower applied field (longer transit time).

current profile. The one considered by Scher and Montroll is based mainly on the spreading of the sheet of charge carriers to a width comparable to the specimen thickness. In fact, the spreading can arise from statistical fluctuation associated with a variety of processes, includ-

ing multiple trapping, hopping, dispersion due to material inhomogeneity, etc. The spreading can result in a decreasing current level and considerable smearing of the transit edge.

Obviously, more experimental work is required to examine the validity of the Scher-

Montroll theory for organic disordered systems, such as undoped and doped polymers. The Scher–Montroll theory is based on a set of reasonable assumptions that make an analytical solution possible. However, microscopic description of carrier motion in an amorphous system, such as a polymer with pendent groups containing important elements, is quite complex and may render an analytical solution impossible. In such cases, computer simulation of the experimental data can be informative.^{250–252}

References

1. IEEE Dielectrics and Electrical Insulation Society, "Digest of Literature on Dielectrics," published annually since 1940 by the National Academy of Science, Washington, D.C.
2. J. O'Dwyer, *The Theory of Electrical Conduction and Breakdown in Solid Dielectrics*, (Clarendon, Oxford, 1973).
3. D. A. Seanor (Ed.), *Electrical Properties of Polymers*, (Academic Press, New York, 1982).
4. N. F. Mott and E. A. Davis, *Electronic Processes in Non-Crystalline Materials*, 2nd edition, (Clarendon, Oxford, 1979).
5. H. J. Wintle, "Conduction Processes in Polymers," in *Engineering Dielectrics*, vol. IIA, edited by R. Bartnikas and R. M. Eichhorn, (ASTM, Philadelphia, 1983), p. 239.
6. K. C. Kao and W. Hwang, *Electrical Transport in Solids*, (Pergamon, Oxford, 1981).
7. A. D. Liddiard, "Ionic Conductivity," in *Handbook of Physics*, vol. 20, (McGraw Hill, New York, 1957).
8. A. R. von Hippel, *Molecular Science and Molecular Engineering*, (Wiley, New York, 1959).
9. S. Saito, H. Sasabe, T. Nakajima, and K. Yada, *J. Polym. Sci.*, A-2, 6, 1297 (1968).
10. K. Tihira and K. C. Kao, *J. Phys. D: Appl. Phys.*, 18, 2247 (1985).
11. A. B. Blythe, *Electrical Properties of Polymers*, (Cambridge University Press, Cambridge, 1979).
12. L. Brillouin, *Wave Propagation in Periodic Structures*, (Dover, New York, 1953).
13. E. J. Blatt, *Physics of Electronic Conduction in Solids*, (McGraw-Hill, New York, 1968).
14. J. R. Chelikowsky and M. L. Cohen, *Phys. Rev.*, B14, 556 (1976).
15. R. A. Smith, *Semiconductors*, (Cambridge University Press, Cambridge, 1964).
16. S. M. Sze, *Physics of Semiconductor Devices*, (Wiley, New York, 1981).
17. A. C. Smith, J. F. Janak, and R. B. Adler, *Electronic Conduction in Solids*, (McGraw-Hill, New York, 1967).
18. J. M. Andre and J. Ladik (Eds.), *Electronic Structures of Polymers and Molecular Crystals*, (Plenum, New York, 1975).
19. D. Eley and M. R. Willis, "The Electrical Conduction of Solid Free Radicals and the Electron Tunneling Mechanism," in *Symposium on Electrical Conductivity in Organic Solids*, edited by H. Kallmann and M. Silver, (Wiley Interscience, New York, 1961), pp. 257–276.
20. G. Kemeny and B. Rosenberg, *J. Chem. Phys.*, 53, 3549 (1970).
21. R. A. Keller and H. E. Rast, *J. Chem. Phys.*, 36, 2640 (1962).
22. J. L. Katz, S. A. Rice, S. I. Choi, and J. Jortner, *J. Chem. Phys.*, 39, 1683 (1963).
23. N. F. Mott and W. D. Twose, *Adv. Phys.*, 10, 107 (1961).
24. N. F. Mott, *Can. J. Phys.*, 34, 1356 (1956).
25. E. M. Conwell, *Phys. Rev.*, 103, 51 (1956).
26. H. Fritzsche, *J. Phys. Chem. Solids*, 6, 69 (1958).
27. H. Fritzsche, *Phys. Rev.*, 119, 1899 (1960).
28. H. Fritzsche and M. Cuevas, *Phys. Rev.*, 119, 1238 (1960).
29. C. A. Mead, *Phys. Rev.*, 128, 2088 (1962).
30. J. G. Simmons and R. R. Verderber, *Radio Electron. Eng.*, 34, 81 (1967).
31. A. J. Bosman and C. Crevecoeur, *Phys. Rev.*, 144, 763 (1966).
32. A. J. Springthorpe, J. G. Austin, and B. A. Smith, *Solid State Commun.*, 3, 143 (1965).
33. A. P. Schmidt, *J. Appl. Phys.*, 39, 3140 (1968).
34. A. K. Jonscher and R. M. Hill, *Physics of Thin Films*, vol. 8, (1975) pp. 169–249.
35. F. Gutmann and L. E. Lyon, *Organic Semiconductors*, (Wiley, New York, 1967).
36. H. Frohlich, H. Pelzer, and S. Zienau, *Phys. Mag.*, 41, 221 (1950).
37. H. Frohlich, *Adv. Phys.*, 3, 325 (1954).
38. D. C. Langreth, *Phys. Rev.*, 159, 717 (1967).

39. J. Appel, "Polarons," in *Solid State Physics*, vol. 21, edited by F. Seitz, D. Turnbull, and H. Ehrenreich (Academic Press, New York, 1968), (1968), pp. 193-391.
40. I. G. Austin and N. F. Mott, *Adv. Phys.*, **18**, 41 (1969).
41. D. Adler, *Solid State Physics*, vol. 21, (Academic Press, New York, 1968), pp. 1-113.
42. V. N. Bogomolov, E. K. Kudinov, and Yu A. Firsov, *Soviet Phys.—Solid State*, **9**, 2502 (1968).
43. A. J. Bosman and H. J. van Daal, *Adv. Phys.*, **19**, 1 (1970).
44. G. R. Allcock, *Adv. Phys.*, **5**, 412 (1956).
45. T. Holstein, *Ann. Phys. (New York)*, **8**, 325 and 343 (1959).
46. W. Siebrand, *J. Chem. Phys.*, **41**, 3574 (1965).
47. G. Kemeny and B. Rosenberg, *J. Chem. Phys.*, **53**, 3549 (1970).
48. J. Vilfan, *Phys. Stat. Sol. (B)*, **59**, 351 (1973).
49. R. R. Heikes and D. W. Johnston, *J. Chem. Phys.*, **26**, 582 (1957).
50. H. J. Queisser, "Semiconductors in the Relaxation Regime," in *Solid State Devices 1972*, edited by P. N. Robson, Conference Series no. 15, (Institute of Physics, London, 1972), pp. 145-168.
51. H. J. Queisser, *J. Appl. Phys.*, **41**, 3892 (1972).
52. E. A. Silinsh, *Organic Molecular Crystals*, (Springer-Verlag, Berlin, 1980).
53. W. Helfrich, "Space Charge Limited and Volume Controlled Currents in Organic Solids," in *Physics and Chemistry of the Organic Solid State*, vol. 3, (Wiley Interscience, New York, 1967), pp. 1-65.
54. M. A. Lampert and P. Mark, *Current Injection in Solids*, (Academic Press, New York, 1970).
55. J. M. Thomas, J. O. Williams, and L. M. Turton, *Trans. Faraday Soc.*, **64**, 2496 (1968).
56. J. M. Caywood, *Mol. Cryst. and Liq. Cryst.*, **12**, 1 (1970).
57. D. C. Hoesterey and G. M. Letson, *J. Phys. Chem. Solids*, **24**, 1609 (1963).
58. J. Sworakowski, *J. Appl. Phys.*, **41**, 292 (1970).
59. P. J. Reucroft and F. D. Mullins, *J. Chem. Phys.*, **58**, 2918 (1973).
60. H. P. D. Lanyon, *Phys. Rev.*, **130**, 134 (1963).
61. K. Unger, *Phys. Stat. Sol.*, **2**, 1279 (1962).
62. A. Sussman, *J. Appl. Phys.*, **38**, 2738 (1967).
63. J. G. Simmons and W. G. Taylor, *J. Phys. C.*, **6**, 3706 (1973).
64. E. A. Silinsh, *Phys. Stat. Sol. (A)*, **3**, 817 (1970).
65. G. P. Owen and A. Charlesby, *J. Phys. C.*, **7**, L400 (1974).
66. A. Nespurek and P. Semejtek, *Czech H. Phys.*, **B. 22**, 160 (1972).
67. J. S. Bonham, *Aust. J. Chem.*, **26**, 927 (1973).
68. J. Grenet, C. Vautier, D. Carles, and J. J. Chabrier, *Phil. Mag.*, **28**, 1265 (1973).
69. W. Hwang and K. C. Kao, *Solid State Electron.*, **15**, 523 (1972).
70. M. A. Nicolet, *J. Appl. Phys.*, **37**, 4224 (1966).
71. D. W. Corington and D. C. Ray, *J. Appl. Phys.*, **45**, 2616 (1974).
72. N. F. Mott and R. W. Gurney, *Electronic Processes in Ionic Crystals*, (Dover, New York, 1940).
73. A. Rose, *Concepts in Photoconductivity and Allied Problems*, (Wiley Interscience, New York, 1963).
74. M. A. Lampert, *Phys. Rev.*, **103**, 1648 (1956).
75. A. Rose, *Phys. Rev.*, **97**, 322 and 1538 (1955).
76. R. W. Smith and A. Rose, *Phys. Rev.*, **97**, 1531 (1955).
77. H. T. Henderson, K. L. Ashley, and M. K. L. Shen, *Phys. Rev.*, **B6**, 4079 (1972).
78. P. Mark and W. Helfrich, *J. Appl. Phys.*, **33**, 205 (1962).
79. D. F. Williams and M. Schadt, *J. Chem. Phys.*, **53**, 3480 (1970).
80. J. M. Thomas, J. O. Williams, and G. A. Cox, *Trans. Faraday Soc.*, **64**, 2496 (1968).
81. P. J. Reucroft and F. D. Mullins, *J. Phys. Chem. Solids*, **35**, 347 (1974).
82. W. Hwang and K. C. Kao, *J. Chem. Phys.*, **60**, 3845 (1974).
83. W. Hwang and K. C. Kao, *Solid State Electron.*, **19**, 1045 (1976).
84. J. G. Simmons, W. G. Taylor, and M. C. Tam, *Phys. Rev.*, **B7**, 3714 (1973).
85. R. S. Muller, *Solid State Electron.*, **6**, 25 (1963).
86. P. N. Murgatroyd, *Thin Solid Films*, **17**, 335 (1973).
87. I. G. Paritskii and A. I. Rozental, *Soviet Phys.-Semicon.*, **1**, 210 (1967).
88. G. G. Roberts, *Phys. Stat. Sol.*, **27**, 209 (1968).
89. A. Touraine, C. Vautier, and D. Carles, *Thin Solid Films*, **9**, 229 (1972).

90. M. Schadt and D. F. Williams, *J. Chem. Phys.*, **50**, 4364 (1969).
91. D. F. Barbe and C. R. Westgate, *J. Chem. Phys.*, **52**, 4046 (1970).
92. H. Baessler, G. Herrmann, N. Riehl, and G. Vaubel, *J. Phys. Chem. Solids*, **30**, 1579 (1969).
93. M. Lax, *Phys. Rev.*, **119**, 1052 (1960).
94. A. G. Milnes, *Deep Impurities in Semiconductors*, (Wiley, New York, 1973).
95. G. A. Dussel and R. H. Bube, *J. Appl. Phys.*, **37**, 2797 (1966).
96. B. K. Ridley and T. B. Watkins, *Proc. Phys. Soc. (London)*, **78**, 710 (1961); also *J. Phys. Chem. Solids*, **22**, 155 (1961).
97. H. C. Law and K. C. Kao, *Solid State Electron.*, **13**, 659 (1970).
98. A. R. Elsharkawi and K. C. Kao, *Solid State Electron.*, **16**, 1355 (1973).
99. R. H. Bube, *Photconductivity of Solids*, (Wiley, New York, 1960).
100. R. H. Bube, *Electronic Properties of Crystalline Solids: An Introduction to Fundamentals*, (Academic Press, New York, 1974).
101. W. Shockley and W. T. Read, *Phys. Rev.*, **87**, 835 (1952).
102. J. S. Blakemore, *Semiconductor Statistics*, (Pergamon Press, Oxford, 1962).
103. R. H. Parmenter and W. Ruppel, *J. Appl. Phys.*, **30**, 1548 (1959).
104. M. A. Lampert, *Phys. Rev.*, **125**, 126 (1962).
105. M. A. Lampert and R. B. Schilling, "Current Injection in Solids: the Regional Approximation Method," in *Semiconductors and Semimetals*, vol. 6, edited by R. K. Willardson and A. C. Beer, (Academic Press, New York, 1970), pp. 1-96.
106. K. L. Ashley and A. G. Milnes, *J. Appl. Phys.*, **35**, 369 (1964).
107. R. Baron and J. W. Mayer, "Double Injection in Semiconductors," in *Semiconductors and Semimetals*, vol. 6, edited by R. K. Willardson and A. C. Beer, (Academic Press, New York, 1970), pp. 201-314.
108. P. Migliorato, G. Margaritondo, and P. Perfetti, *J. Appl. Phys.*, **47**, 656 (1976).
109. H. J. Deuling, *J. Appl. Phys.*, **41**, 2179 (1970).
110. Yu. A. Bykovskii, K. N. Vinogradov, and V. V. Zuev, *Soviet Phys.-Semicon.*, **1**, 1295 (1968); also **3**, 1442 (1970).
111. Y. Otani, K. Matsubara, and Y. Nishida, *J. Appl. Phys.*, **41**, 4711 (1970).
112. I. Melingailis and R. H. Rediker, *J. Appl. Phys.*, **33**, 1883 (1962).
113. O. S. Mortensen, R. W. Munn, and D. F. Williams, *J. Appl. Phys.*, **42**, 1192 (1971).
114. W. P. Dumke, "Theory of the Negative Resistance in p-i-n Diodes, Physics of Semiconductors," in *Proceedings of the International Conference, Paris, 1964*, (Academic Press, New York, 1964), p. 611.
115. O. J. Marsh, R. Baron, and J. W. Mayer, *Appl. Phys. Lett.*, **7**, 120 (1965).
116. E. M. Conwell, *High Field Transport in Semiconductors*, (Academic Press, New York, 1967).
117. J. G. Simmons, *D.C. Conduction in Thin Films*, (Mills and Boon, London, 1971).
118. N. F. Mott, *Phil. Mag.*, **24**, 911 (1971).
119. B. R. Nag, *Theory of Electrical Transport in Semiconductors*, (Pergamon, Oxford, 1972).
120. A. K. Jonscher, "AC Conductivity and High Field Effects," in *Electronic and Structural Properties of Amorphous Semiconductors*, edited by P. G. LeComber and J. Mort (Academic Press, New York, 1973), pp. 329-362.
121. J. M. Marshall and G. R. Miller, *Phil. Mag.*, **27**, 1151 (1973).
122. A. M. Barnett, "Current Filaments in Semiconductors," *IBM J. Res. Dev.*, **13**, 522-528 (1969).
123. A. M. Barnett, "Current Filament Formation," in *Semiconductors and Semimetals*, vol. 6, edited by R. K. Willardson and A. C. Beer, (Academic Press, New York, 1970), pp. 141-200.
124. A. M. Barnett and A. G. Milnes, *IEEE Trans. Electron Devices*, **ED-13**, 816 (1966).
125. A. M. Barnett and A. G. Milnes, *J. Appl. Phys.*, **37**, 4215 (1966).
126. K. C. Kao, *IEEE Trans. Electr. Insul.*, **EI-11**, 121 (1976).
127. M. Saji and K. C. Kao, *J. Non-Cryst. Solids*, **22**, 223 (1976).
128. W. P. Ballard and R. W. Christy, *J. Non-Cryst. Solids*, **17**, 81 (1975).
129. J. Frenkel, *Phys. Rev.*, **54**, 647 (1938); also *Tech. Phys. U.S.S.R.*, **5**, 685 (1938).
130. M. Ieda, G. Sawa, and S. Kato, *J. Appl. Phys.*, **42**, 3737 (1971).
131. R. M. Hill, *Phil. Mag.*, **23**, 59-86 (1971).
132. G. A. N. Connell, D. L. Camphasen, and W. Paul, *Phil. Mag.*, **26**, 541 (1972).
133. J. Antula, *J. Appl. Phys.*, **43**, 4663 (1972).
134. W. Vollmann, *Phys. Stat. Sol. (A)*, **22**, 195 (1974).
135. J. L. Hartke, *J. Appl. Phys.*, **39**, 4871 (1968).
136. V. Adamic and J. H. Calderwood, *J. Phys. D: Appl. Phys.*, **8**, 551 (1975).

137. L. Onsager, *J. Chem. Phys.*, **2**, 599 (1934).
138. L. Onsager, *Phys. Rev.*, **54**, 554 (1938).
139. D. M. Pai, *J. Appl. Phys.*, **46**, 5122 (1975).
140. D. M. Pai and R. C. Enck, *Phys. Rev.*, **B11**, 5163 (1975).
141. R. H. Batt, C. L. Braun, and J. F. Horning, *J. Chem. Phys.*, **49**, 1967 (1968).
142. R. R. Chance and C. L. Braun, *J. Chem. Phys.*, **59**, 2269 (1973).
143. P. J. Melz, *J. Chem. Phys.*, **57**, 1694 (1972).
144. N. E. Geacintov and M. Pope, "Intrinsic Photoconductivity in Organic Crystals," in *Proceedings of the 3rd International Conference on Photoconductivity*, edited by E. M. Pell, (Pergamon, Oxford, 1971), pp. 289–295.
145. G. Pfister and D. F. Williams, *J. Chem. Phys.*, **61**, 2516 (1974).
146. J. Bardeen and W. Shockley, *Phys. Rev.*, **80**, 72 (1950).
147. E. Conwell and V. F. Weisskopf, *Phys. Rev.*, **77**, 388 (1950).
148. H. Ehrenreich, *Phys. Rev.*, **120**, 1951 (1960).
149. M. A. Lampert, *J. Appl. Phys.*, **29**, 1082 (1958).
150. J. L. Wagener and A. G. Milnes, *Solid State Electron.*, **8**, 495 (1965).
151. J. B. Gunn, *Solid State Commun.*, **1**, 88 (1963).
152. B. K. Ridley and T. B. Watkins, *Proc. Phys. Soc. (London)*, **78**, 293 (1961).
153. P. N. Butcher, *Phys. Lett.*, **19**, 546 (1965).
154. P. N. Butcher, W. Fawcett, and C. Hilsum, *British J. Appl. Phys.*, **17**, 841 (1966).
155. K. C. Kao, *Solid State Commun.*, **9**, 599 (1971).
156. H. Kroemer, *IEEE Spectrum*, **5**, 47 (1968).
157. G. S. Kino and I. Kuru, *IEEE Trans. Electron. Devices*, **ED-16**, 735 (1969).
158. K. C. Kao, *J. Phys. D: Appl. Phys.*, **17**, Pt. I. (Solids without Defects), 1433–1448; Pt. II. (Solids with Defects), 1449–1467 (1984).
159. A. M. Goodman and A. Rose, *J. Appl. Phys.*, **42**, 2823 (1971).
160. R. I. Frank and J. G. Simmons, *J. Appl. Phys.*, **38**, 832 (1967).
161. J. C. Schug, A. C. Lilly, and D. A. Lowitz, *Phys. Rev.*, **B1**, 4811 (1970).
162. G. R. Johnston and L. E. Lyons, *Aust. J. Chem.*, **23**, 2187 (1970).
163. M. Stuart, *Phys. Stat. Sol.*, **23**, 595 (1967).
164. W. Hwang and K. C. Kao, *J. Chem. Phys.*, **58**, 3521 (1973).
165. H. P. Kunkel and K. C. Kao, *J. Phys. Chem. Solids*, **37**, 863 (1976).
166. J. Sworakowski, J. M. Thomas, D. F. Williams, and J. O. Williams, *J. Chem. Soc. Faraday Trans.*, **II70**, 676 (1974).
167. K. C. Kao and D. M. Tu, *J. Appl. Polymer Science*, **90**, 1864 (2003).
168. D. M. Tu, Y. Yin, X. Wang, Y. Fan, J. Wang, and Q. Lei, "Space Charge Distribution, TSC and TSL Spectra in Polyethylene Aged by Electric Stress," *Proceedings of the 6th International Conference on Properties and Applications of Dielectric Materials (ICPADM)* June 20–26, 2000 (IEEE Dielectrics and Electrical Insulation Society, New York, 2000), pp. 46–50.
169. T. Mizutani, *High Voltage DC Insulation and Space Charge*, *ibid.*, pp. 18–23.
170. J. R. Haynes and W. Shockley, *Phys. Rev.*, **81**, 835 (1951).
171. J. R. Haynes and W. C. Westphal, *Phys. Rev.*, **85**, 680 (1952).
172. R. G. Kepler, *Phys. Rev.*, **119**, 1226 (1960).
173. R. G. Kepler and D. C. Hoesterey, *Phys. Rev.*, **B9**, 2743 (1974).
174. O. H. LeBlanc, Jr., *J. Chem. Phys.*, **33**, 626 (1960).
175. F. C. Brown, *Phys. Rev.*, **97**, 355 (1955).
176. W. E. Spear, *Proc. Phys. Soc. (London)*, **B70**, 669 (1957).
177. W. E. Spear, *J. Non-Cryst. Solids*, **1**, 197 (1969).
178. J. A. Cooper, Jr., and D. F. Nelson, *J. Appl. Phys.*, **54**, 1445 (1983).
179. A. G. R. Evans and P. N. Robson, *Solid State Electron.*, **17**, 805 (1974).
180. A. Many and G. Rakavy, *Phys. Rev.*, **126**, 1980 (1962).
181. R. Baron, M. A. Nicolet, and V. Rodriguez, *J. Appl. Phys.*, **37**, 4156 (1966).
182. R. B. Schilling and H. Schachter, *J. Appl. Phys.*, **38**, 1643 (1967).
183. W. Helfrich and P. Mark, *Z. Phys.*, **168**, 495 (1962).
184. L. M. Schwartz and J. F. Hornig, *J. Phys. Chem. Solids*, **26**, 1821 (1965).
185. R. M. Blakney and H. P. Grunward, *Phys. Rev.*, **159**, 658 (1967).
186. I. P. Batra and H. Seki, *J. Appl. Phys.*, **41**, 3409 (1970).
187. I. P. Batra, B. H. Schechtman, and H. Seki, *Phys. Rev.*, **B2**, 1592 (1970).
188. S. Z. Weisz, A. Cobas, S. Trester, and A. Many, *J. Appl. Phys.*, **39**, 2296 (1968).
189. A. Many, S. Z. Weisz, and M. Simhony, *Phys. Rev.*, **126**, 1989 (1962).
190. C. Bogus, *Z. Phys.*, **184**, 219 (1965); *ibid.*, **207**, 218 (1970).
191. A. C. Papadakis, *J. Phys. Chem. Solids*, **28**, 641 (1967).

192. M. D. Tabak and M. E. Scharfe, *J. Appl. Phys.*, **41**, 2114 (1970).
193. D. Liu and K. C. Kao, *J. Appl. Phys.*, **69**, 2486 (1991).
194. D. Liufu, X. S. Wang, D. M. Tu, and K. C. Kao, *J. Appl. Phys.*, **83**, 2209 (1998).
195. J. Lindmayer, *J. Appl. Phys.*, **36**, 196 (1965).
196. R. H. Walden, *J. Appl. Phys.*, **43**, 1178 (1972).
197. H. J. Wintle, *J. Appl. Phys.*, **44**, 3514 (1973).
198. H. J. Wintle, *J. Non-Cryst. Solids*, **15**, 471 (1974).
199. H. J. Wintle, *IEEE Trans. Electr. Insul.*, **EI-12**, 97 and 424 (1977).
200. D. K. Schroder, *Semiconductor Material and Device Characterization*, (Wiley Interscience, New York, 1990).
201. D. Liufu and K. C. Kao, *J. Appl. Phys.*, **85**, 1089 (1999).
202. W. Smith, *Nature*, **7**, 303 (1873).
203. R. Pohl, *Ber Physik Ges.*, 961 (1911).
204. B. Gudden and R. Pohl, *Z. Physik*, **1**, 365 (1920); **3**, 98 (1920); **4**, 206 and 262 (1921); and **5**, 176 (1921).
205. J. Wilson and J. F. B. Hawkes, *Optoelectronics*, (Prentice Hall, New York, 1983).
206. E. Uiga, *Optoelectronics*, (Prentice Hall, Englewood Cliffs, New Jersey, 1995).
207. C. L. Chen, *Elements of Optoelectronics and Fiber Optics*, (Irwin, Chicago, 1996).
208. S. M. Ryvkin, *Photoelectric Effects in Semiconductors*, English translation by A. Tybulewicz (Consultants Bureau, New York, 1964).
209. F. C. Strome, Jr., *Phys. Rev. Lett.*, **20**, 3 (1968).
210. R. G. Kepler, *Pure Appl. Chem.*, **27**(3), 515 (1970).
211. M. Pope and C. E. Swenberg, *Electronic Processes in Organic Crystals*, (Clarendon, Oxford, 1982).
212. H. Meier, *Organic Semiconductors—Dark and Photoconductivity of Organic Solids*, (Verlag Chemie, Weinheim, Germany, 1974).
213. S. I. Choi, *J. Chem. Phys.*, **40**, 1691 (1964); also *J. Chem. Phys.*, **43**, 1818 (1965).
214. A. Rose, *RCA Rev.*, **12**, 362–414 (1951).
215. H. Kiess and A. Rose, *Helv. Phys. Acta*, **46**, 434 (1973); also *Phys. Rev. Lett.*, **31**, 153 (1973).
216. C. Popescu and H. K. Henisch, *J. Phys. Chem. Solids*, **37**, 47 (1976).
217. F. Evangelisti, P. Fiorini, G. Fortunato, A. Frova, C. Giovanella, and R. Peruzzi, *J. Non-Cryst. Solids*, **55**, 191 (1983).
218. V. Halpern, *Phil. Mag.*, **B54**, 473 (1986).
219. F. Vaillant and D. Jousse, *Proc. Symp. Materials Issues in Amorphous Semiconductor Technology*, vol. 70, edited by D. Adler, Y. Hamakawa, and A. Madann, (MRS, Pittsburgh, 1986), p. 143.
220. J. J. Schellenberg and K. C. Kao, *J. Phys. D: Appl. Phys.*, **21**, 1764 (1988).
221. F. Stockmann, *Phys. Stat. Sol.*, **34**, 741 (1969); also **34**, 351 (1969).
222. L. Kan and K. C. Kao, *J. Chem. Phys.*, **98**, 3445 (1993).
223. R. A. Larson, *IBM J. Res. Dev.*, **24**, 286 (1980).
224. A. Endo, M. Takada, K. Adachi, H. Takasago, T. Yada, and Y. Onishi, *J. Electrochem. Soc.*, **134**, 2522 (1987).
225. E. Sugimoto, *IEEE Electr. Insul. Mag.*, **5**, 15 (1989).
226. A. Endo and T. Yada, *J. Electrochem. Soc.*, **132**, 155 (1985).
227. S. A. Kafafi, J. P. LaFemina, and J. L. Nauss, *J. Am. Chem. Soc.*, **112**, 8742 (1990).
228. C. Z. Van Doorn, *Physica*, **20**, 1155 (1954).
229. C. Z. Van Doorn and D. de Nobel, *Physica*, **22**, 338 (1956).
230. P. Görlich, *Photoconductivity in Solids*, (Routledge and Kegan Paul, London, 1967).
231. R. M. Schaffert, *IBM J. Res. Dev.*, **15**, 75 (1971).
232. R. M. Schaffert, *Electrophotography*, (Halstead Press, New York, 1975).
233. E. M. Williams, *The Physics and Technology of Xerographic Processes*, (Wiley Interscience, New York, 1984).
234. G. Weiser, *J. Appl. Phys.*, **43**, 5028 (1972).
235. K. Meier, *Spectral Sensitization*, (Focal Press, New York, 1968).
236. W. Veith, *Z. Angew Physik*, **7**, 1 (1955).
237. H. Scher and E. W. Montroll, *Phys. Rev.*, **B12**, 2455 (1975).
238. F. W. Schmidlin, *Solid State Commun.*, **22**, 451 (1977); also *Phys. Rev.*, **B16**, 2362 (1977).
239. J. Nooland, *Solid State Commun.*, **24**, 477 (1977); also *Phys. Rev.*, **B16**, 4466, 4474 (1977).
240. G. Pfister and H. Scher, *Phys. Rev.*, **B15**, 2026 (1977); also *Adv. Phys.*, **27**, 747 (1978).
241. G. Pfister, *Phys. Rev.*, **B16**, 3676 (1977).
242. G. Pfister and C. H. Griffiths, *Phys. Rev. Lett.*, **40**, 659 (1978).
243. J. Mort and G. Pfister (Eds), *Electronic Properties of Polymers*, (Wiley, New York, 1982).

- 244. H. Scher, "Theory of Time-Dependent Photoconductivity in Disordered Systems," in *Photoconductivity and Related Phenomena*, edited by J. Mort and D. M. Pai, (Elsevier, New York, 1976), pp. 71–115.
- 245. M. E. Sharfe, *Phys. Rev.*, *B2*, 5026 (1970); also *Bull. Am. Phys. Soc.*, *18*, 454 (1973).
- 246. M. Mort and A. J. Lakatos, *J. Non-Cryst. Solids*, *4*, 117 (1970).
- 247. W. D. Gill, *J. Appl. Phys.*, *43*, 5033 (1972).
- 248. H. Seki, *Proc. 5th Conf. on Amorphous and Liquid Semiconductors*, Garnish-Partenkirchen, 1973 (Taylor and Francis, London, 1974).
- 249. S. M. Godson and J. Hirsch, *Solid State Commun.*, *20*, 285 (1976).
- 250. J. M. Marshall, *Phil. Mag.*, *36*, 959 (1977); *ibid.*, *38*, 335 (1978); *ibid.*, *43*, 401 (1980).
- 251. M. Silver and L. Cohen, *Phys. Rev.*, *B12*, 2455 (1975); also *Phys. Rev.*, *B15*, 3276 (1977).
- 252. M. Silver, G. Schönherr, and H. Bassler, *Phil. Mag.*, *43*, 943 (1981).

8 Electrical Aging, Discharge, and Breakdown Phenomena

The tao begot one, one begot two, two begot three, and three begot the ten thousand things. The ten thousand things carry yin and embrace yang. They achieve harmony by combining these forces. The ten thousand things are brought into existence by the interaction of the three basic elements, heaven, earth and human; and this interaction comes from the emptiness.

Lao Tsu (600BC)

The sayings of Lao Tsu may be interpreted as follows: The tao means the universal law of nature and virtue, which creates the spirit (one). The spirit creates the yin and yang (two). (Yin and yang may mean female and male, negative and positive charges, forward and reverse directions etc.) The yin and yang create the heaven, the earth, and the human (three). The heaven, the earth and the human create all things. (*The ten thousand things* means a great number of things, that is, all things.)

All things possess yin and yang. They can reach a state of dynamic equilibrium and stability because of the collective interaction of the yin and the yang that makes the combined effort of the heaven, the earth, and the human productive. All things are brought into existence by the collective interaction of the yin and the yang in the heaven, the earth, and the human. This interaction comes from energy. (The emptiness may mean something very important, but invisible and untouchable. In today's language, it may mean the energy that is intangible. Because of the limitations of ancient language, people might not have known how to describe energy in terms of something invisible and untouchable. This may be why it was termed *the emptiness*.)

Most electrical failures in electrical or electronic engineering systems are caused by electrical aging, partial discharge, or breakdown in insulating materials. The failure of power cables in the power industry and the failure of electronic elements involving thin insulating

films, such as silicon dioxide or silicon nitride thin films, in the microelectronics industry are good examples. In the previous chapter, we mentioned that electrical conduction at high fields is caused by carrier injection from electrical contacts. So we can consider carrier injection the starting point for dealing with the topics covered in this chapter.

8.1 Electrical Aging

When an insulating material is subjected continually to an electrical stress, the material will be in a nonequilibrium state and its properties will change with time. In this state, the material is said to undergo *electrical aging*. Electrical aging is a gradual degradation process leading to destructive breakdown of the material. Obviously, the lifetime of an electrically stressed material depends on the magnitude of the electric stress applied to the material and the length of the time it has been subjected to such a stress. This implies that the lifetime of a material depends mainly on the electrical contacts, which control carrier injection, and the kind and concentration of carrier traps, which control the degradation process. Electrical aging is always an important problem of concern to industry.

8.1.1 Theory

The general features of insulating materials are as follows:

- Carrier mobility is low, usually lower than $10^{-1} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$.
- Dielectric relaxation time is much greater than carrier lifetime.
- The energy band gap is large, generally larger than 4 eV.
- The localized gap state concentration is much larger than the thermal equilibrium carrier concentration.
- The mean free path is small, usually of the order of several molecular radii (5–20 Å).

Since the work function of most metals is smaller than 4 eV and the work function of most insulating materials, including polyethylene, silicon dioxide, etc., is larger than 8 eV, the potential barrier height of the metal–insulator contact for electron injection is generally smaller than that for hole injection. For most electrical insulation systems, the applied

average field F is usually less than 1 MV cm^{-1} , so the carriers injected into an insulator are mainly electrons. Assuming that the electrical contacts are neutral contacts (see Types of Electrical Contacts in Chapter 6), the electrons may be injected into the insulator by a thermionic emission process or by an electron tunneling process, as shown in Figure 8-1. However, if the applied field is much larger than 1 MV cm^{-1} , the hole current may become predominant, particularly when hole mobility is larger than electron mobility, even though the probability for hole injection is not better than for electron injection, as in polyethylene.¹

At normally high fields, it is most likely that electrons are injected into the conduction band of an insulator specimen from the injecting contact by Fowler–Nordheim tunneling.² As shown in Figure 6-20(a), these injected electrons will quickly become trapped after a few scatterings because of the small mean free path.

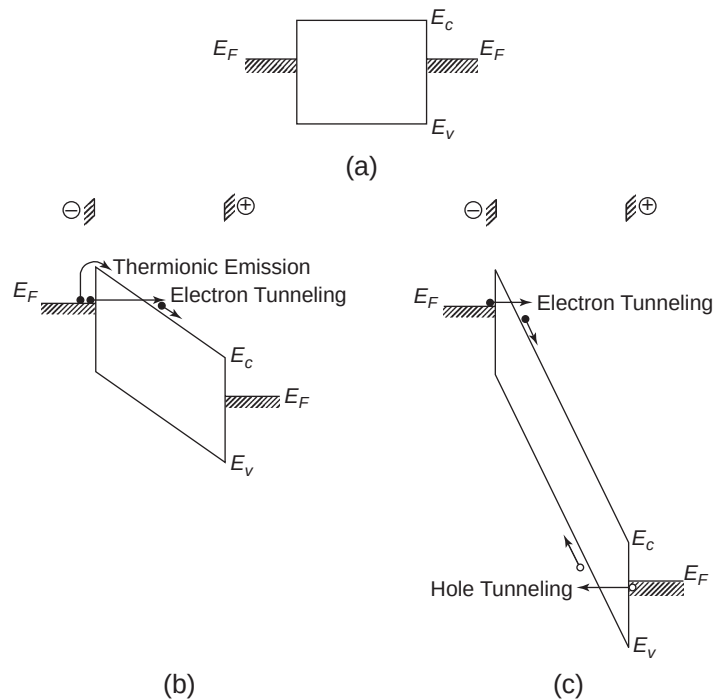


Figure 8-1 Possible processes of charge carriers injected from neutral electrical contacts to an insulating material (a) at zero field $F = 0$, (b) at high fields $F < 1 \text{ MV cm}^{-1}$, and (c) at very high fields $F > 1 \text{ MV cm}^{-1}$.

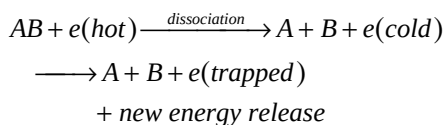
Electron trapping will result in the following two important phenomena:

1. After being trapped, the trapped electrons will form a negatively charged homo-space charge near the electron-injecting contact (cathode) and hence create an internal field F_i opposite to the applied field, reducing the effective field for and rate of electron injection. This space charge also enhances the field toward the anode and may switch on hole injection.
2. In the transition from an upper to a lower energy state due to trapping (or recombination), an energy equal to the energy difference between the two states will be evolved, which is mainly nonradioactive for non-crystalline insulating materials. The energy evolved at each trapping or recombination event is of the order of 3–4 eV or greater for deep traps and recombination centers.

No complete theory is currently available to describe quantitatively how such a large quantity of energy, evolved due to nonradioactive transition, is dissipated. Similarly, no aging or breakdown theory so far put forward has mentioned the importance of this energy to the breakdown process. We believe that this energy may be dissipated in two possible ways:

It may be dissipated directly, causing structural damage in the microregion around the trap site in which the electron is being trapped.

The energy evolved from one trapping event may be transferred to another electron and make it become a hot electron via an Auger-type process, as shown in Figure 8-2(a). This second electron can now have sufficient energy to bombard a molecule and break its bonds. In other words, the energy will be used up by dissociating a molecule into free radicals, which create traps.



This new energy release from the second electron due to trapping will be transferred to a

third electron, making it another hot electron. This process will go on in a chain action to produce more and more radicals to form low-density regions and create traps. The energy of the hot electron depends on the location of traps in the energy band gap. For condensed insulating materials, it can be larger than most bond dissociation energies E_d , which are generally lower than 4 eV.

For example, E_d for C—C, C—H, CH₃—H, CH₃—CH₃, Si—H, and Si—O bonds are 3.50, 3.55, 4.40, 3.60, 3.05, and 3.80 eV, respectively.^{3,4} These energies are much lower than the energy band gap, which is about 9 eV for polyethylene and SiO₂ and which is the energy required for impact ionization. The energy of the hot electrons is generated by an Auger-type process, which is completely different from the conventional concept: the energy gained from the applied field. It is important to note that the probability for the creation of hot electrons by an Auger-type process increases with increasing concentration of injected electrons.⁵ Obviously, the concentration of injected electrons is higher near the injecting contact; it also increases with increasing applied electric field.

The dissipation of the energy evolved due to nonradioactive transition of electron trapping or recombination in causing damage to the material structure is analogous to the dissipation of the energy of water drops dripping from a higher level to a stone, creating a recess on the stone's surface, as shown in Figure 8-3.

We have just mentioned two important phenomena. One is the formation of a negative homo-space charge near the electron-injecting contact (cathode) by trapped electrons. After a prolonged period of electrical stressing, short-circuiting the two electrodes will result in a discharging current flow. In this case, the trapped carriers (mainly trapped electrons in this case) will be thermally emitted to the conduction band and then discharged at both electrodes. At the time of short-circuiting, the concentration of trapped electrons is higher near the electron-injecting contact and decreases with x toward the noninjecting contact, and at point x_1 the

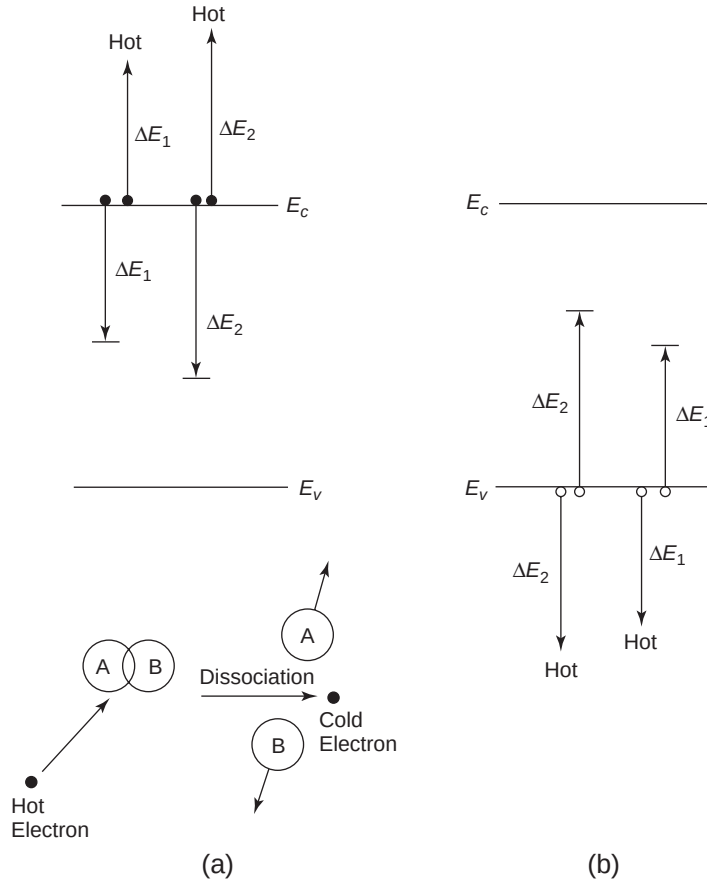


Figure 8-2 Schematic diagrams illustrating (a) hot electron generation by an Auger-type process due to electron trapping, and the dissociation of a molecule by the bombardment of a hot electron and (b) hot hole generation by a similar process due to hole trapping.

internal field $F_{in} = 0$, as shown in Figure 8-4. The field distribution at $x < x_1$ and $x > x_1$ can be expressed as

$$F = -(q/\epsilon) \int_{x_1}^x n_t(x) dx \quad (8-1)$$

where $n_t(x)$ is the trapped electron density distribution function. Since the two short-circuiting electrodes are at the same potential, the position of x_1 can be determined from the following equation:

$$\begin{aligned} V &= -\int_0^d F dx \\ &= (q/\epsilon) \int_0^d \int_{x_1}^x n_t(x') dx' dx = 0 \end{aligned} \quad (8-2)$$

The internal field F_i is controlled by $n_t(x)$. The larger the value of $n_t(x)$ near the injecting contact, the closer point x_1 is to the injecting contact and hence the higher the field is toward the injecting contact (i.e., the larger the slope of the potential between $x = 0$ and $x = x_1$), as shown in Figure 8-4(b). It is likely that x_1 is very close to the injecting contact. Depending on the density of the accumulated trapped electrons $n_t(x)$, it is possible that after prolonged electrical stressing, $n_t(x)$ may reach such a level that F_i becomes high enough to cause an internal discharge, leading to final, destructive breakdown of the specimen when short-circuiting the electrodes. This is one of the major disadvantages of polymeric-insulated cables for DC

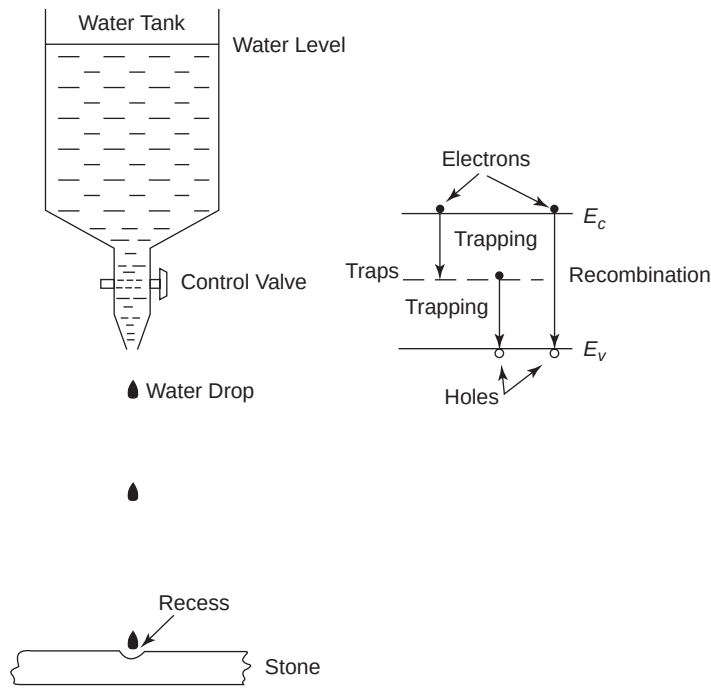


Figure 8-3 The continual dripping of water on a stone surface will produce a recess on the surface after a prolonged period due to the dissipation of the energy of the water drops from the gravitation force. This is analogous to the dissipation of energy evolved due to nonradiative transition (trapping and recombination) in causing damage to the material structure.

power transmission. For polymeric-insulated cables, trap concentration is likely to be higher near the metal–polymer contacts.

An increase in applied field not only causes an increase in carrier injection and trap filling but also results in the creation of more new traps.^{6,7} It can be imagined that each electron (or hole) trapping event will evolve an energy of the order of 2–5 eV, depending on the trap energy level. This energy, if not converted to light emission, will be dissipated in the material in the form of breaking bonds, creating defects or traps.⁶ Electrical aging is due to this gradual degradation process. The number of newly created traps should increase with increasing injection of carriers into the insulator. In other words, the number of newly created traps increases with increasing electric stress for a fixed length of stressing period, or with increasing length of stressing period at a fixed electric stress. The increase in structural degradation and trap concentration after

prolonged electric stressing reflects the degree of electrical aging and hence the lifetime of the electrically stressed material.

On the basis of this concept, we have derived a theoretical formula for the prediction of the lifetime of electrically stressed insulating materials. The derivation is now briefly described.

Supposing that the probability for the creation of a new trap under an average field F is γ , then the rate of the new trap creation can be written as

$$\frac{dN'_t}{dt} = \frac{\gamma J(t)}{q} \quad (8-3)$$

where N'_t is the newly created trap concentration and $J(t)$ is the injected current density.⁸ The probability γ can be considered a field-activated process, expressed as

$$\gamma = \gamma_0 \exp(BF) \quad (8-4)$$

where γ_0 and B are constants related to the chemical structure of the material.^{9–11}

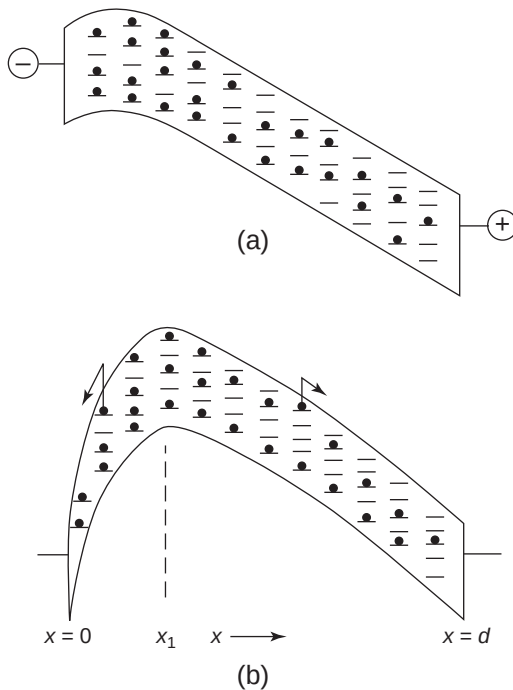


Figure 8-4 Schematic energy band diagrams for an insulating material (a) immediately after a period of being electrically stressed and (b) immediately after short-circuiting the two electrodes with electrons thermally emitted from traps to the conduction band (during discharging).

The dark conduction current decay with time after the application of a step-function electric field has been proved experimentally to be associated with the trap-filling process.¹² This phenomenon was discussed in Electronic Conduction in Chapter 7. Several investigators have studied this current decay phenomenon theoretically.^{13–17} Their analyses and experimental evidence indicate that J is proportional to t^{-m} , with m close to unity. Therefore, J can be reasonably expressed as

$$J = J_0(1 + Kt^{-m}), \quad t > 0, 0 < m < 1 \quad (8-5)$$

where K and m are constants depending on the material structure and the potential barrier profile of the injecting contact, and J_0 is the quasi-steady state current after the current decay transient period, which is field dependent.

Substituting Equations 8-4 and 8-5 into Equation 8-3 and solving it with the boundary condition $N'_t = 0$ at $t = 0$, we obtain

$$N'_t = \frac{\gamma_0 J_0 t}{q} \left(1 + \frac{K}{1+m} t^{-m} \right) \exp(BF) \quad (8-6)$$

For prolonged electric stressing (i.e., large t), the term $Kt^{-m}/(1+m)$ becomes much smaller than 1. Then, Equation (8-6) can be simplified to

$$N'_t = \frac{\gamma_0 J_0 t}{q} \exp(BF) \quad (8-7)$$

Assuming that J_0 is proportional to F and that destructive breakdown occurs when the concentration of the field-created new traps reaches a certain critical value $N'_{t(\text{crit})}$, then the lifetime t of an insulating material subjected to a stressing field F can be predicted by

$$N'_{t(\text{crit})} = AtF \exp(BF) \quad (8-8)$$

where A is a constant depending on the material structure and the potential barrier profile of the injecting contact.⁸ Thus, for a fixed applied stressing field F , the lifetime of the polymer is t , and for a fixed stressing time t , an average stressing field F is required to create $N'_{t(\text{crit})}$, that is, to cause destructive breakdown.

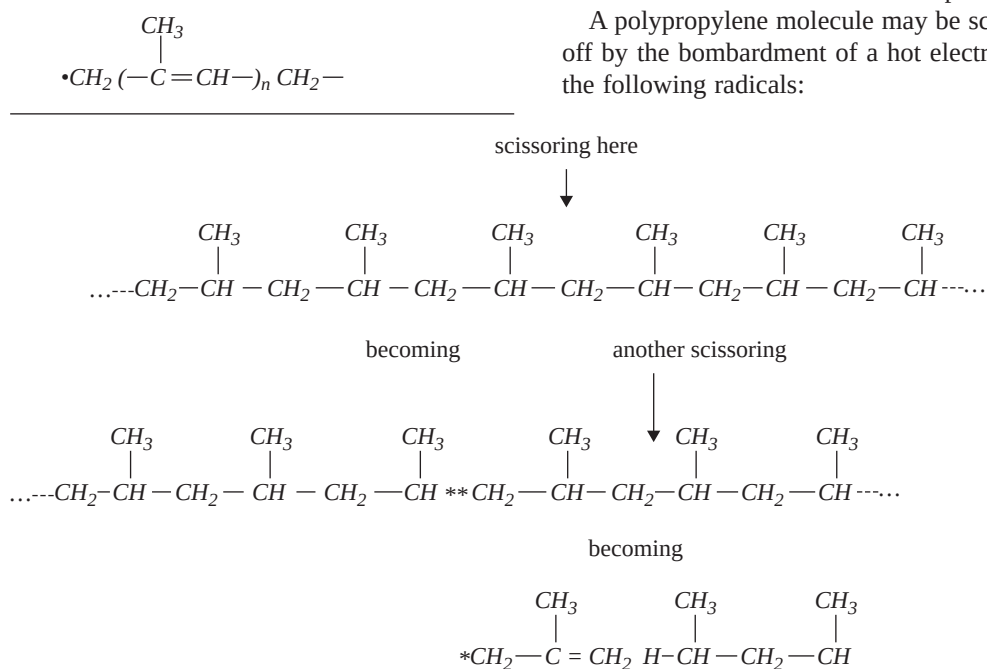
8.1.2 Measurements of Electrical Aging

The degree of electrical aging of an electrically stressed insulating material can be considered the degree of structural degradation,⁸ which can be measured as the rate of the change in the properties of the material—in other words, the rate of the increase in the concentration of the stress-created new traps. For insulating polymers, we have attributed structural degradation to the bombardment of hot electrons on macromolecules.⁶ Thus, the lifetime depends on the electrically stressing condition. The lifetime can be simply defined as the time required for the concentration of the stress-created new traps to reach a certain critical value. This section presents some typical experimental results showing the degree of electrical aging for widely used insulating polymers, such as polypropylene and polyethylene, and correlates these results with the theory given previously.^{8,18} In polymers, structural degradation can be diagnosed by measuring the concentra-

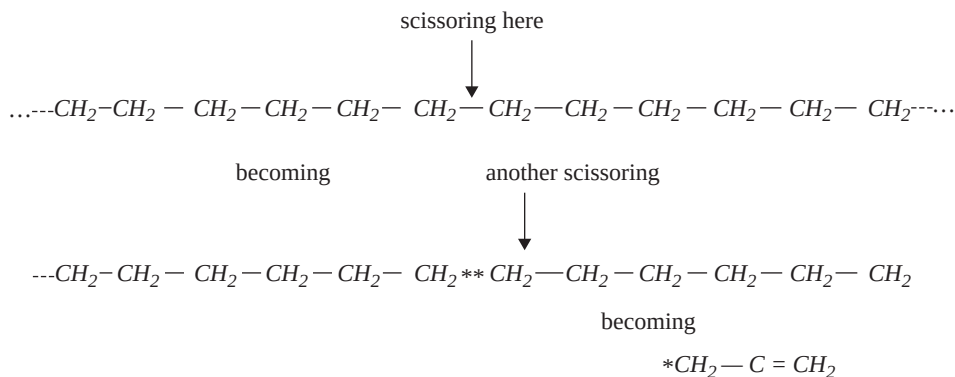
tion of free radicals formed in the polymer due to continuous electrical stressing.

Electron Paramagnetic Resonance (EPR) Spectroscopy

The basic principle of EPR was described in Magnetic Resonance in Chapter 1. EPR spectroscopy can be used to identify free radicals and estimate their concentration. Some experimental results show that the free radicals formed in polypropylene (PP) after the PP specimen is electrically aged are in the form shown here, based on absorption peaks in EPR spectrograms¹⁹:



Similarly, a polyethylene molecule may be scissored into the following radicals.



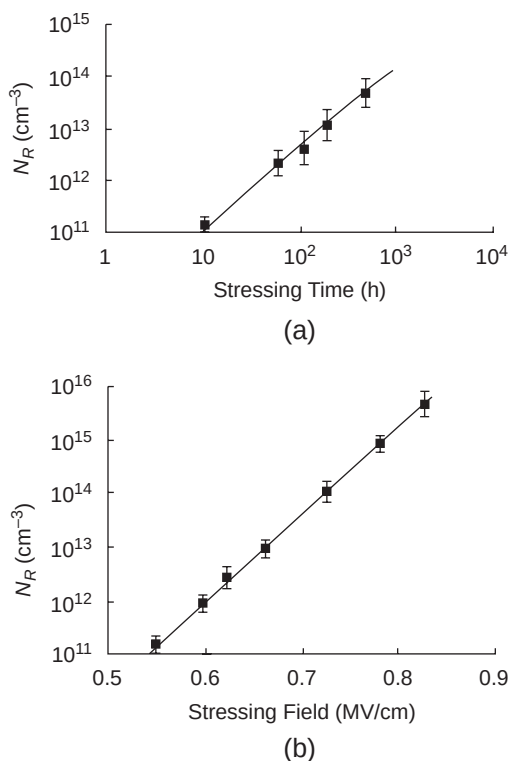


Figure 8-5 The concentration of free radicals N_R in polypropylene specimens after being subjected to electrical stressing (a) at a fixed field of 833 kV cm^{-1} as a function of stressing time and (b) for a fixed stressing time of 250 hours as a function of stressing field.

The two broken parts have free valences in the form of unpaired electrons, which are generally referred to as *free radicals* with the sign (*). The radicals react readily with other atoms, molecules, or radicals. They also act as acceptorlike electron traps to capture electrons.²¹

The concentration of free radicals N_R in PP that has been subjected to an electric stress of 833 kV cm^{-1} as a function of stressing time is shown in Figure 8-5(a). That for a fixed stressing time of 250 hours as a function of stressing field is shown in Figure 8-5(b). These results are in good agreement with the theory (see Equation 8-8).

Infrared (IR) Absorption Spectroscopy

Infrared (IR) absorption spectra have been measured before and after the PP specimens

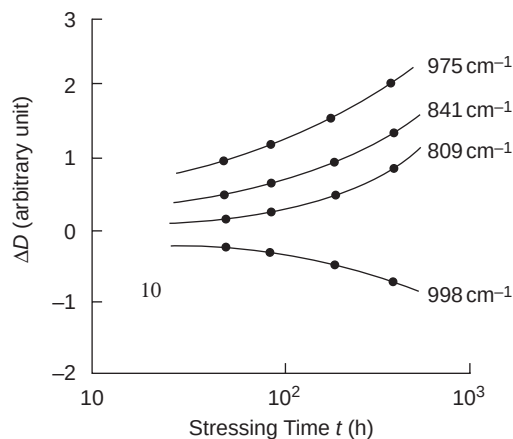


Figure 8-6 The difference in IR absorption peaks ΔD at 809, 841, 975, and 998 cm^{-1} between the absorption peaks D after and D_0 before the polypropylene specimens were subjected to electrical stressing under an AC field of 833 kV cm^{-1} as a function of stressing time.

were subjected to electrical stressing under an AC field of 833 kV cm^{-1} for 75, 100, 250, and 500 hours. [The changes in the absorption peaks occur mainly at 809, 841, 975 and 998 cm^{-1} .⁸ The difference between absorption peaks D after and D_0 before the specimen was subjected to electrical stressing is denoted by ΔD ; it changes with the stressing time, as shown in Figure 8-6. The peak at 809 cm^{-1} is due to the movement of irregular parts of the structure; the peak at 841 cm^{-1} is due to the vibration of double-bond vinyl groups; the peak at 975 cm^{-1} is due to the vibration of tertiary methyl groups; and the peak at 998 cm^{-1} is due to the skeletal vibration of macromolecular chains.^{8,18}

These results indicate clearly that macromolecules have been dissociated into low-weight molecules or radicals by electron trapping and hot electron bombardment. This is why the absorption peak at 998 cm^{-1} decreases and the peaks at 809, 841, and 975 cm^{-1} increases with stressing time, as shown in Figure 8-6. Electrical aging is caused by the incessant dissociation of macromolecules and the incessant creation of new traps in an electrically stressed polymer.

If the tertiary methyl group, $-\text{CH}_2-\text{CH}_2^*$, and the double-bond vinyl group, $R_1R_2-\text{C} =$

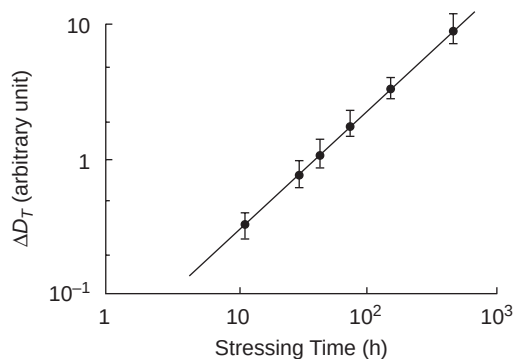


Figure 8-7 The sum of the relative IR absorption peaks $\Delta D_T = \Delta D_{975} + \Delta D_{841}$ for polypropylene specimens subjected to electrical stressing under an AC field of 833 kV cm^{-1} as a function of stressing time.

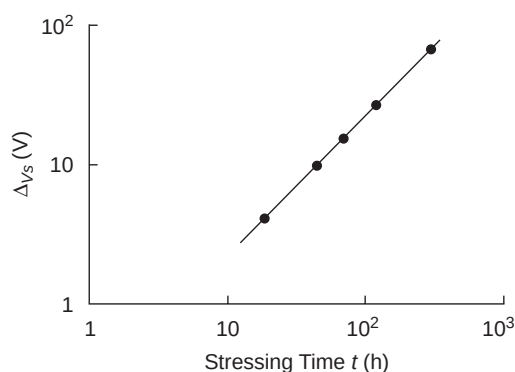


Figure 8-8 The increment of the surface potential for polypropylene specimens after being electrically stressed at 833 kV cm^{-1} as a function of stressing time.

CH- R_2 (R_1 and R_2 are radicals), represent low-weight radicals separated from the macromolecules,²² the sum of the relative peaks at 975 cm^{-1} and 841 cm^{-1} (i.e., $\Delta D_T = \Delta D_{975} + \Delta D_{841}$) reflects the total amount of free radicals formed. Figure 8-7 shows ΔD_T as a function of stressing time for PP specimens electrically stressed at 833 kV cm^{-1} . The results are in good agreement with the theory (see Equation 8-8).

Surface Potential Measurements

The principle and the method for surface potential measurements were described in Measurements of Surface Potentials in Chapter 7. Surface potential can be used to determine the concentration of traps. The surface potential measured for PP specimens prior to electrical stressing, termed V_{so} , should reflect the concentration of originally existing traps. After the specimens are subjected to electrical stressing, the surface potential, measured again (termed V_s), should reflect the concentration of total traps, including originally existing traps and newly stress-created traps. Thus, the increment in surface potential $\Delta V_s = V_s - V_{so}$ is directly related to the concentration of the stress-created traps. Figure 8-8 shows the increment of the surface potential ΔV_s for PP after it was electrically stressed at 833 kV cm^{-1} , as a function of stressing time.

To ensure that electrical aging is caused by the increase of stress-created traps with increasing stressing time, we have also studied the effects of ultraviolet light-created traps (optical aging) on the surface potential. The results are similar, indicating that electrical aging and optical aging produce the same effects, due to the creation of new traps in the material. Again, for a given electrical stress, N'_t increases linearly with stressing time, as predicted by Equation 8-7.

Small Angle Scattering of X-Rays (SASX) Spectroscopy

Structural defects created by electron trapping and recombination events will produce microvoids. To detect microvoids, small angle scattering of X-rays (SASX) spectra have been measured for PP specimens after the specimens were subjected to electrical stressing at 833 kV cm^{-1} for various stressing times. There are two major causes of small angle scattering: boundaries between the crystalline and the noncrystalline regions, and the existence of microvoids in the specimen. In general, SASX due to the second cause is much greater than that due to the first cause, so SASX due to the first cause may be ignored.¹⁹

Based on the analysis of SASX data using the standard method,^{23,24} the concentration of

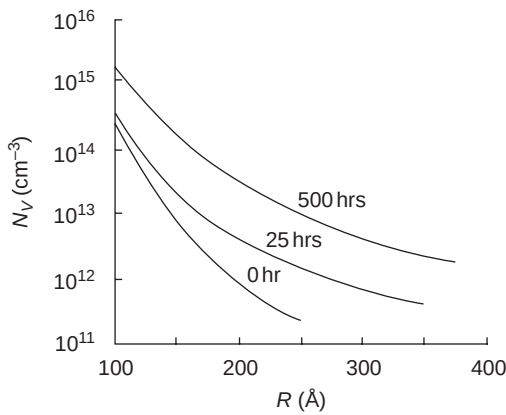


Figure 8-9 The variation of microvoid concentration N_v with microvoid size in radius R (microvoid distribution) in polypropylene specimens after being electrically stressed at 833 kV cm^{-1} for 0 hours, 25 hours, and 500 hours.

microvoids and the distribution of their sizes have been determined.¹⁸ Results for PP are shown in Figure 8-9. The number and the size of microvoids both increase with increasing stressing time, indicating that the material undergoes electrical aging. The large microvoids may be the so-called *low-density domains* referred by Kao.⁶ He has considered the formation of low-density domains a necessary prelude to the occurrence of electrical discharge and breakdown, which are initiated by impact ionization. Only in low-density domains where the electron free path is large can impact ionization take place. Naturally, the more structural degradation, the smaller the microcrystallinity is. As a result, free volume increases and the density of the material decreases with increasing stressing time, as expected.

Lifetime Evaluation

The lifetime of an electrically stressed insulating material can be considered the time required for the occurrence of destructive breakdown at a given electrical stress. A typical relationship between the breakdown strength F_b and the stressing time (lifetime) for PP specimens is shown in Figure 8-10(a). The higher the electrical stress applied to the material, the

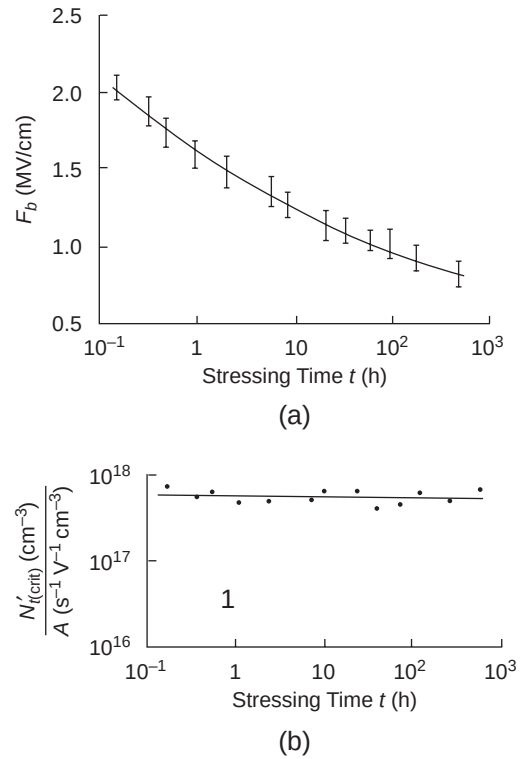


Figure 8-10 (a) Variation of the time to breakdown (lifetime) with the applied stressing field (breakdown strength F_b) and (b) the value of $N'_{t(\text{crit})}/A$ polypropylene specimens aged at various stressing fields (F_b) until the occurrence of destructive breakdown at t (lifetime) for each set of F_b and t from Figure 8-10(a).

shorter its lifetime is. By calculating $N'_{t(\text{crit})}$ based on Equation 8-8 for each set of F_b and t from the data given in Figure 8-10 and then plotting $N'_{t(\text{crit})}/A$ as a function of stressing time in Figure 8-10(b), it can be seen that the value of $N'_{t(\text{crit})}/A$ remains practically unchanged irrespective of F_b and t values chosen for breakdown measurements, where A is a normalizing factor, which is a constant. It can be concluded, therefore, that the lifetime of an electrically stressed insulating material is the time required for the concentration of new traps created by electron trapping and hot electron bombardments in the material to reach a certain critical value $N'_{t(\text{crit})}$.

Other Measurements

Thermally stimulated current (TSC), thermally stimulated luminescence (TSL), and pulsed electro-acoustic (PEA) methods were discussed in Relaxation Times of Dipoles and the Thermally Stimulated Discharge Current (TSDC) Technique and Spatial Distributions of Trapped Real Charges in Chapter 5. These methods can be used to investigate electrical aging problems. Several investigators have used the TSC and TSL methods to study the concentrations and energy levels of electron traps and luminescence centers, and the PEA method to study space charge distributions in electrically aged low-density polyethylene (LDPE).²⁵ These investigators have reported that the total concentrations of electron traps and luminescence centers created by the applied electrical stress of 500 kV cm^{-1} increase with increasing stressing time. Also, the net space charge in LDPE specimens that have been electrically aged at 500 kV cm^{-1} DC increases with increasing stress time. Their findings on electrical aging phenomena are very similar to those mentioned previously.

8.1.3 Remedy for Electrical Aging

To reduce the effects of electrical aging (in other words, to increase the lifetime of an electrically stressed material), we must inhibit the gradual structural degradation process. As mentioned, the major causes of this damaging process are carrier injection from electrical contacts and the subsequent energy evolved due to carrier trapping and recombination. The following two methods would serve the remedy purpose.

Emission Shields

The major function of emission shields is to suppress carrier injection from the injecting contact. Carrier injection efficiency depends mainly on the potential barrier and the applied electrical field near the injecting contact. The applied field tends to reduce the height and the width of the potential barrier, thus enhancing carrier injection. An emission shield should be chosen so that its potential barrier height is not lower than the insulating material with respect

to the injecting contact and its permittivity is higher than that of the insulating material. An emission shield with a high potential barrier and a high permittivity should be able to suppress carrier injection and increase the treeing initiation voltage and breakdown strength.

It has been found that the coating of emission shields in polyethylene with a point-plane electrode configuration, the initiation voltage for electrical treeing in polyethylene is much higher for the iron point electrode tip oxidized (i.e. the tip covered with a thin oxide layer) than that without.²⁶ The iron oxide acts as an emission shield because the relative permittivity of iron oxide is higher than that of polyethylene.

It has also been found that the effect of emission shields in polyethylene with a quasi-uniform field spherical electrode configuration using a mixture of BaTiO_2 and polyethylene (PE) as emission shields increases the breakdown strength.²⁷ The relative permittivity of this mixture can be adjusted by adjusting the proportion of the two components: BaTiO_2 and PE. Table 8-1 shows the effect of this emission shield. The experiment was carried out with both spherical electrodes coated with emission shields. The short-circuit breakdown with DC prestressing means that the polyethylene specimen was first prestressed at a predetermined field for three minutes. After that, the specimen was short-circuited and then examined for breakdown. Each of the breakdown data was the average of at least 15 measurements.²⁷

Silicon-incorporated polyethylene (SiPE) has also been used as an emission shield. The

Table 8-1 The increase in percent of the breakdown strength of PE due to the action of emission shields.

<i>Emission shield coated on PE surfaces</i>	<i>50 Hz AC normal breakdown</i>	<i>Short-circuit breakdown with DC prestressing</i>
$\text{BaTiO}_2 + \text{PE}$ mixture with $\epsilon_{\text{ES}} = 3.4$	37.9%	34.5%
$\text{BaTiO}_2 + \text{PE}$ mixture with $\epsilon_{\text{ES}} = 7.9$	103.0%	46.0%

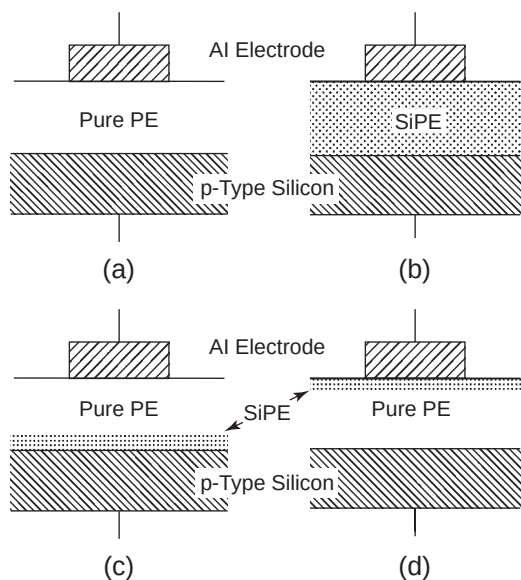


Figure 8-11 Metal-polymer-semiconductor configurations (a) Al-PE-Si, (b) Al-SiPE-Si, (c) Al-PE-SiPE-Si, and (d) Al-SiPE-PE-Si.

SiPE films were fabricated by plasma polymerization in a gas mixture of ethylene and silane. Four MIS configurations, shown in Figure 8-11, were used for this study, the electrode area and the specimen thickness being $1.13 \times 10^{-2} \text{ cm}^2$ and $1000 \text{ \AA}$, respectively.²⁸ For the two-layer configurations, the thicknesses of PE and SiPE were, respectively, $860 \text{ \AA}$ and $140 \text{ \AA}$, and the SiPE was fabricated in a gas mixture with a volume ratio of $\text{SiH}_4/\text{C}_2\text{H}_4$ of 0.066. Using linear ramp voltages and with the Al gate electrode negatively biased, the I - F curve shifts very little from that of the pure polyethylene when the SiPE layer is in contact with the Al electrode, as shown in Figure 8-12. But with the SiPE layer in contact with the p -Si substrate, which is positively biased, the I - F curve shifts a great deal toward high fields.²⁸ It can be seen that the SiPE layer acts as an emission shield suppressing hole injection from the p -Si, thus increasing the threshold field for hole injection and the breakdown strength.

Obviously, the materials chosen for emission shields depend on the insulating material and the type of carrier injection to be suppressed.

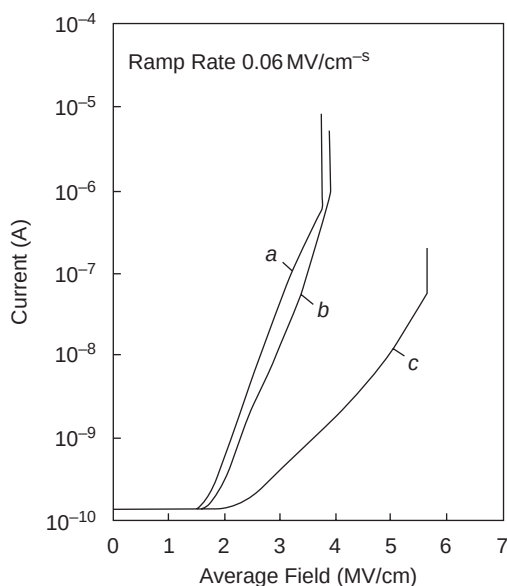


Figure 8-12 Current-average field (I - F) characteristics for the configurations (a) Al-PE-Si, (b) Al-SiPE-Si, and (c) Al-PE-SiPE-Si with the SiPE layer deposited at the $\text{SiH}_4/\text{C}_2\text{H}_4$ volume ratio of 0.066.

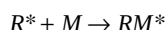
However, based on the concept given here, an emission shield with both the potential barrier height and the relative permittivity higher than those of the insulating material should effectively suppress carrier injection. It has been reported that an emission shield with a relative permittivity of 4.16 made of a compound of ethylene-vinyl acetate (EVA) and TiO_2 effectively suppresses electron injection to cross-linked polyethylene (XLPE) power cables.²⁹

Radical Scavengers

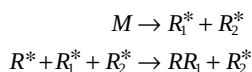
To reduce the continuous structural degradation associated with the energy evolved due to trapping and recombination of carriers injected from electrical contacts, it is necessary to reduce the concentration of deep traps or to convert deep traps to shallow traps to reduce the energy evolved. This can be achieved by incorporating suitable dopants into the insulating material. In this section, we take insulating polymers as typical insulating materials. The dissociation of macromolecules by electron

bombardment will form free radicals, leaving behind broken molecular bonds that act as acceptorlike electron traps. If the traps are deep traps, an energy of the order of 3–4 eV may be evolved at each trapping event. This energy may be large enough to break the molecular bonds to form free radicals, which act as traps.

An electrically stressed insulating polymer contains a large number of molecules with free valencies in the form of unpaired electrons at their endings. These unpaired electrons act as acceptorlike electron traps. If a dopant can satisfy such unpaired electrons, it may serve the purpose of killing a deep trap and possibly producing a new shallow trap. Denoting the molecule with an unpaired electron (radical) as R^* and the dopant as M , the reaction between R^* and M may be as follows:



This may kill the deep trap due to R^* and create a new shallow trap due to RM^* . A dopant may also be dissociated into two new radicals R_1^* and R_2^* when it is dissolved into the polymer. If this is the case, we may have



It is possible that R_1^* may combine chemically with R^* to form a compound RR_1 , leaving R_2^* to act as a new shallow trap. Dopants that can perform such a function would inhibit the activity of R^* . This is why they can be called the *radical scavengers*. Incorporation of such radical scavengers into insulating polymers would improve their dielectric properties and extend their life expectancy. Several investigators have reported that the incorporation of suitable dopants, such as halogen comonomers, into polyethylene increases the breakdown strength.^{30,31} Here, we do not intend to review all possible dopants for this purpose. Rather we will use a simple example to demonstrate the concept of radical scavengers.

One kind of dopant containing mainly phenolic hydroxyl can serve this purpose. This material consists of nonsymmetrical radicals needing only a small activation energy to move

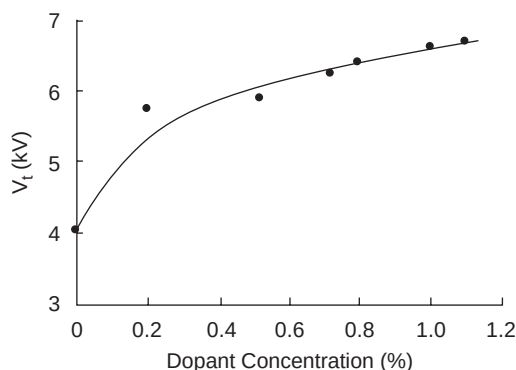


Figure 8-13 The mean tree initiation voltage as a function of dopant concentration of phenolic hydroxyl in percentage of the weight of the polyethylene specimens at 80°C.

and to react with broken macromolecular bonds to inhibit their activity. This implies that the action of the dopants tends to stabilize the disturbed region involving electron trapping and recombination, inhibiting the broken molecular bonds from acting as deep electron traps. For polyethylene doped with phenolic hydroxyl, the initiation voltage V_t for electrical treeing increases with increasing dopant concentration, as shown in Figure 8-13 (Electrical treeing will be discussed in Section 8.2.) But such an increase tends to become saturated. Further increase in dopant concentration may even produce a reverse effect. There is an optimal concentration, which would give an optimal result. It has also been reported that for low-density polyethylene, the concentration of new traps created by applied electrical stress of 500 kV cm^{-1} for a period of time is lower for the specimens doped with 1% of phenolic hydroxyl than that for the pure specimens.²⁵

These remedies for electrical aging are analogous to the remedy for a person under a lot of stress. To reduce the effects stress and extend life expectancy, this stressed person should reduce worries or convert big worries to small ones. To do this, special drugs may be required. Similarly, for electrically stressed insulating materials, special materials are required to act as emission shields and radical scavengers.

8.2 Electrical Discharges

Electrical discharges involve processes by which atoms or molecules become electrically charged due to ionization by avalanches of hot carriers, mostly starting in the medium of gas state. Electrical discharges that do not bridge the electrodes or any pair of electrical contacts in an electrical system, are called *partial discharges*. In fact, all partial discharges involve gas discharges.³²

In general, partial discharges can be classified into four types:

1. **Corona discharges:** These generally refer to discharges occurring in the vicinity of a sharp point or edge of a metallic contact, or in the vicinity of a conducting particle whose surrounding field is extremely high due to divergent or inhomogeneous field distribution, as shown in Figure 8-14(a).
2. **Surface discharges:** These discharges occur on the surface of a dielectric material, as shown in Fig 8-14(b).
3. **Internal discharges:** These discharges occur in inclusions or cavities originally existing in a dielectric material, or in low-

density domains or channels created due to electrical stressing at high fields,⁶ as shown in Figure 8-14(c).

4. **Electrical treeing:** Shown in Figure 8-14(d), electrical treeing may be considered a combination of corona and internal discharges.

All electrical discharges are detrimental to dielectric materials. They may cause permanent changes in chemical structure or in constituent elements of the material's molecules. In the following sections, we shall discuss the mechanisms leading to partial discharges.

8.2.1 Internal Discharges

To aid the understanding of internal partial discharges, we begin with a composite dielectric system consisting of a gas layer and a dielectric solid layer, as shown in Figure 8-15. In this system, the dielectric constant of the gas (e.g., air) is lower than that of the solid, so the electric stress in the gas is higher than in the solid, while the breakdown strength of gas is much lower than the solid's. Therefore, partial discharges occur in the gas at a critical voltage much lower than the breakdown voltage of the

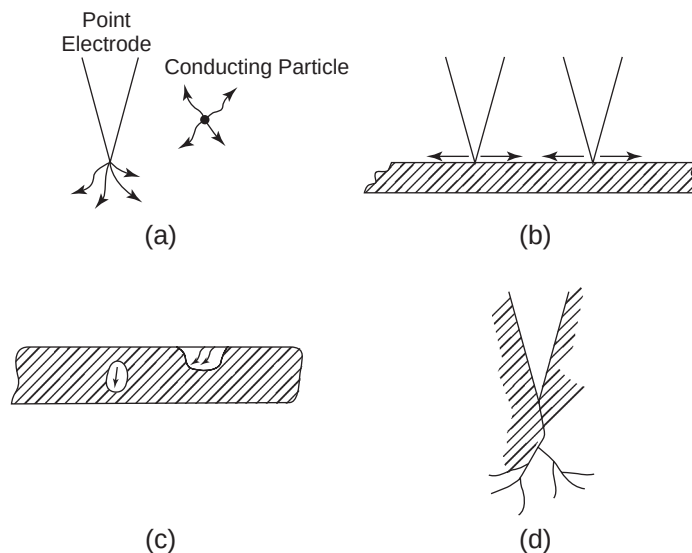


Figure 8-14 Four types of partial discharges: (a) corona discharges, (b) surface discharges, (c) internal discharges, and (d) electrical treeing involving corona and internal discharges.

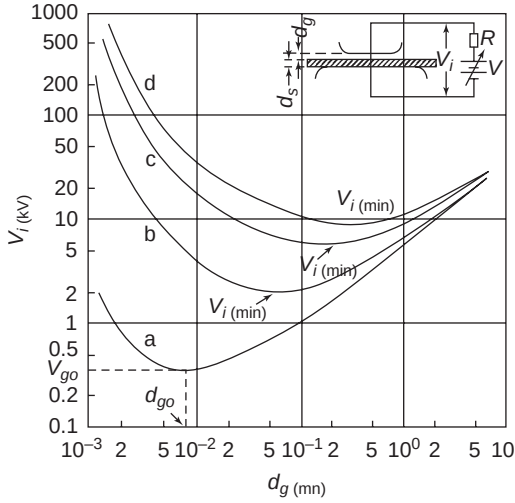


Figure 8-15 Discharge inception voltage V_i as a function of air layer thickness d_g at atmospheric pressure for various solid layer thicknesses d_s . (a) $d_s/\epsilon_r = 0$, $d_g = d_{g0} = 7.4 \times 10^{-3}$ mm, $V_{i(\min)} = V_{g0} = 0.327$ kV; (b) $d_s/\epsilon_r = 0.1$ mm, $d_g = 7.4 \times 10^{-2}$ mm, $V_{i(\min)} = 1.98$ kV; (c) $d_s/\epsilon_r = 0.5$ mm, $d_g = 2.0 \times 10^{-1}$ mm, $V_{i(\min)} = 5.8$ kV; (d) $d_s/\epsilon_r = 1$ mm, $d_g = 2.3 \times 10^{-1}$ mm, $V_{i(\min)} = 8.6$ kV.

whole system. The voltage applied to the composite dielectric system initiating electrical discharge in the gas layer is called the *discharge inception voltage* of the system V_i , which is given by

$$V_i = V_g \left(1 + \frac{d_s/\epsilon_r}{d_g} \right) \quad (8-9)$$

where V_g is the breakdown voltage of the gas layer; d_g and d_s are, respectively, the thicknesses of the gas and the solid layers; ϵ_r is the relative permittivity (or the dielectric constant) of the solid layer; and the dielectric constant of the gas layer is assumed to be unity. For $d_s/\epsilon_r = 0$, we have $V_i = V_g$; the system reduces to a gas gap between two metallic electrodes. The relationship between the breakdown voltage V_g and the gap layer thickness d_g follows Paschen's law, as shown in Figure 8-15. There is a minimum breakdown voltage V_{g0} corresponding to an optimal gas gap d_{g0} . For the air gap, V_{g0} and d_{g0} are, respectively, about 325 V and 7×10^{-3} mm at atmospheric pressure.³³ For $d_s/\epsilon_r \neq 0$, the system is composite. For a fixed value

of d_s/ϵ_r , V_i varies with V_g and d_g , following Equation 8-9. The optimal value of d_g , for which V_i becomes minimal, can be determined by setting $dV_i/dd_g = 0$. Thus, we obtain

$$d_{g(\min)} = \frac{d_s}{\epsilon_r} \left[\frac{V_g/d_g}{dV_g/dd_g} - 1 \right] \quad (8-10)$$

For a fixed solid-layer thickness, there is an optimal value of d_g , denoted by $d_{g(\min)}$, which corresponds to a minimum inception voltage, denoted by $V_{i(\min)}$. According to Equations 8-9 and 8-10, both $d_{g(\min)}$ and $V_{i(\min)}$ move toward higher values as the solid-layer thickness is increased.

Fundamental Features of Internal Discharges in a Cavity

The gas layer in Figure 8-15 is analogous to a gas cavity in a dielectric solid, as shown in Figure 8-16, in which the region I is in the form of a column enclosing the cavity. The actual electric field (and hence the discharge inception voltage) inside such a cavity depends on the size and the shape of the cavity, the ratio of the dielectric constant of the solid to that of the cavity, and the content of the cavity. However, cavity size and shape in practical insulation vary to a great extent, and the cavity location is quite random. Therefore, the minimum value of the discharge inception voltage $V_{i(\min)}$ for a given dielectric solid thickness should be used as the criterion for the design of insulation systems to ensure that internal discharge does not occur at normal operating voltages.

Although it has been generally accepted that the breakdown strength of a gas in a cavity bounded by dielectric material is of about the same order of magnitude as that between equally spaced metallic electrodes,³⁴⁻³⁷ there are many factors that can affect the breakdown voltage of a gas cavity in insulation.³⁸ Apart from factors such as type of gas, gas pressure, cavity wall conditions, etc., surface conductivity over cavity walls plays an important role in the magnitude of discharge inception voltage. Ionization in the gas cavity occurring at a voltage across the cavity $V_{AB'}$ smaller than the breakdown voltage of the cavity may give rise

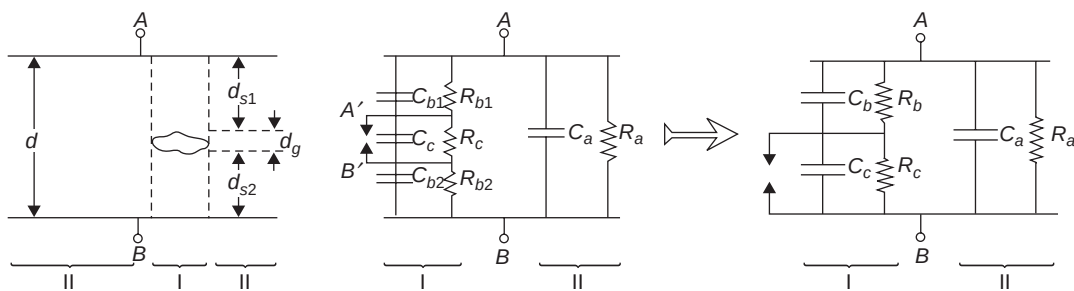


Figure 8-16 Equivalent circuits of a dielectric solid containing a cavity. I: the column with cavity; II: the remaining solid without cavities $d_s = d_{s1} + d_{s2}$, $d = d_g + d_s$.

to charge accumulation on the surface of the cavity. This tends to enhance the field in the gas cavity in subsequent half-cycles when the insulation system is subjected to AC field.³⁸ A slow rate of rise of applied voltage usually gives a lower discharge inception voltage than a fast rate of rise. It has also been reported that superposition of an impulse on alternating voltage may cause a considerable decrease in discharge inception voltage compared to impulses alone in polyethylene specimens.³⁹

Other factors may also influence the breakdown voltage of the cavity, such as static charges on the cavity walls produced by the previous discharges, a semiconducting layer on the cavity walls, contamination of gas content in the cavity due to radicals evolved by previous discharges, etc. Metal splinters embedded in the dielectric may affect the inception voltage, because a metal splinter has sharp edges around which the local field may be high enough to initiate electrical discharges.

For cavities with sizes of the order of a few microns and dielectric solid thicknesses larger than 1 mm, the discharge inception voltage due to cavities may be higher than the initiation voltage for electrical treeing, because the latter involves a very high field at the point electrode. Formulas for the ratio of the field inside the cavity F_i to that outside the cavity F_o under a uniform applied field for various, commonly encountered cavity shapes are given in Table 8-2.

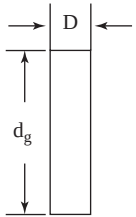
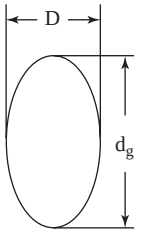
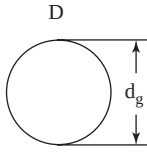
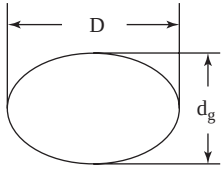
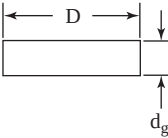
Discharge Current Impulses, Recurrence, Discharge Magnitudes, Discharge Energy, and Power Losses

A commonly seen electrical discharge is the lightning discharge across a huge gas gap between two thick insulating air layers or between a thick insulating air layer and the ground. This discharge can penetrate a considerable depth into insulating materials and possibly cause tremendous damage. This phenomenon is analogous to the internal discharge in a gas cavity in a dielectric solid, as shown in Figure 8-16. The energy involved in a discharge in a gas cavity depends on the energy stored in the whole system before the occurrence of the discharge. Its damaging effect is similar to that of lightning discharge, but on a much smaller scale.

The internal discharge, which involves energetic particles (electrons and ions) continuously bombarding the walls of the cavity and ultraviolet photon irradiation, gradually causes the material to deteriorate. This deleterious effect is directly related to chemical transformation due to prolonged bombardment and irradiation by energetic particles (e.g., burning the polymer material and changing it into gas, carbon, or another low-molecular weight substance).

The most common types of partial discharges in polymeric insulation systems are internal discharges in cavities (or voids), which are

Table 8-2 The electric field inside the cavity for some common cavity shapes.

Item	Cylinder $D \ll d_g$	Prolate Spheroid $D < d_g$	Sphere $D = d_g$	Oblate Spheroid $D > d_g$	Disc $D \gg d_g$
Cavity Shape					
Stress Factor F_1/F_o	1	$\frac{\epsilon_r}{\epsilon_r - (\epsilon_r - 1)G}$ with $G < 1/3$	$\frac{3\epsilon_r}{2\epsilon_r + 1}$ with $G = 1/3$	$\frac{\epsilon_r}{\epsilon_r - (\epsilon_r - 1)G}$ with $G > 1/3$	ϵ_r

F_1 is the electric field inside the cavity.

F_o is the electric field outside the cavity.

ϵ_r is the relative permittivity of the solid with the assumption that $\epsilon_r = 1$ inside the cavity.

$$G = \frac{(d_g/2)(D/2)^2}{2} \int_0^\infty \frac{ds}{[s + (d_g/2)^2]^{3/2} [s + (D/2)^2]}$$

almost unavoidable in the material due to imperfect manufacturing and handling processes, and electrical treeing in low-density domains or channels created by carrier injection from electrical contacts (electrodes) before the occurrence of the electrical discharges.⁶ It is likely that all internal discharges take place in a cavity containing particles in gas form or in a low-density region so that the electrons in such a cavity or a region can have a mean free path much larger than any in the dielectric solid surrounding it.

To understand partial discharges, it is important to understand electrical signals produced by discharges in a gas cavity, because similar signals may be produced by all types of discharges. Since the breakdown strength in a gas cavity or a low-density region is lower than in the dielectric solid surrounding it, and since the field in such a gas cavity or low-density region is larger than that in the surrounding medium, the breakdown of the cavity or the low-density region occurs at a much lower applied field.

A discharge creates hetero-space charges near the two opposite cavity surfaces perpendi-

cular to the field direction, so an internal inverse field will be created in this direction opposite to the applied field, making the net field inside the cavity less than the discharge inception field or equal to zero, thus extinguishing the discharge. If the applied field continues to increase, reaching such a value that applied field F_a minus the inverse field F_r is greater than the discharge inception field F_b of the cavity, then the discharge will occur again.

On the other hand, if the applied voltage does not continue to increase but remains constant, the discharge will not occur again until all space charges leak out and the inverse field reduces to such a value that $F_a - F_r = F_b$ and the voltage across the cavity again reaches discharge inception voltage. This is why the frequency of discharge recurrence in DC cases is much smaller than in AC cases for the same magnitude of the voltage. The applied voltage is constant for the DC cases, but it increases with time in one half-cycle and reverses its polarity in the following half-cycle in AC cases. However, whether the applied voltage is DC, AC, or impulse, the discharge current is in

pulse form. Usually, a discharge develops into a spark if sufficient energy is available.

Next, we shall discuss partial discharges under DC and AC voltage conditions.

Case 1: DC Voltage Conditions

In analyzing discharge current pulses, voltage distribution, and discharge recurrence frequency, we assume that the cavity shown in Figure 8-16 contains gas and that the breakdown voltage follows Paschen's law. The cavity has a capacitance C_c and a parallel effective resistance R_c , which may be the combination of cavity surface resistance and discharge channel resistance, both of which vary from time to time during discharge. C_b is the combination of C_{b1} and C_{b2} : $C_b = C_{b1}C_{b2}/(C_{b1} + C_{b2})$. R_b is the combination of R_{b1} and R_{b2} , ($R_b = R_{b1} + R_{b2}$) which are, respectively, the capacitance and the resistance of the portion of the dielectric specimen in series with the cavity. C_a and R_a , are, respectively, the capacitance and the resistance of the remaining dielectric specimen. R_a and R_b are usually very large, and R_c may become quite small after repeated discharges. Since the size of a cavity is very small compared to the whole specimen, C_a and R_a are usually considered to be the capacitance and the resistance of the whole specimen.

For simplicity, we will assume that the cavity is a cylindrical cavity with a diameter D which is larger than its length d_g , and that it contains a gas with the relative dielectric constant of unity. If this cavity is embedded in a dielectric material having a relative dielectric constant ϵ_r between two parallel plane electrodes with the total thickness d , the voltage across the cavity is given by

$$V_c(t) = V \left[\frac{R_c}{R_b + R_c} + \left(\frac{C_b}{C_b + C_c} - \frac{R_c}{R_b + R_c} \right) \times \exp \left(- \frac{R_b + R_c}{R_b R_c (C_b + C_c)} t \right) \right] \quad (8-11)$$

where V is the applied voltage across the two metallic parallel plane electrodes. Under DC voltage conditions, the voltage distribution is determined initially by capacitances C_b and C_c , and finally by R_b and R_c . During the transient

period, the voltage distribution is determined by the combination of both the capacitances and the resistance, as given by Equation 8-11. At time $t = 0^+$ just after the application of a DC voltage with a finite rise time, Equation 8-11 becomes

$$V_c = V \left(\frac{C_b}{C_b + C_c} \right) = V \left(\frac{\epsilon_r d_g}{d_s + \epsilon_r d_g} \right) \quad (8-12)$$

which is the same as those cases under AC voltage conditions. But, finally for $t \rightarrow \infty$, Equation 8-11 becomes

$$V_c = V_{dc} \left(\frac{R_c}{R_b + R_c} \right) \quad (8-13)$$

Since the time lag of breakdown in a short air gap is very short compared to the variation of the AC voltage at power or low frequencies, there is practically no difference in breakdown voltage of a gas cavity under AC and DC voltage conditions. Suppose that a DC voltage is applied across a dielectric specimen, which is raised linearly to a voltage V_{dc} , as shown in Figure 8-17, and that V_{dc} is just the threshold voltage for the onset of a discharge, which is referred as the *discharge inception voltage* V_i . The voltage across the cavity V_c increases gradually, following Equation 8-11. As soon as it reaches the final value V'_{dc} , the cavity breaks down, because in this case $V'_{dc} = V_g$, the breakdown voltage of the cavity. Once the discharge occurs at $t = t_1$, the discharge current i_d increases rapidly to a peak value I_d within a short time τ_r , which is of the order of 0.2–1.0 ns.

This portion of current is contributed mainly by the flow of electrons generated due to the discharge. After the current reaches its peak value, it decays slowly to a value I_e . The decay time τ_d is of the order of 100 ns, and the portion of current during the decay time is due mainly to ionic movement. At $t = t_1$, the inverse voltage V_r created by the separation of positive and negative charges increases rapidly, until its value reaches a value $V_g - V_r$ equal to the discharge extinction voltage V_e . Then the discharge extinguishes.

After discharge extinction, the charges left behind in the cavity, particularly those on the walls, can only be neutralized through conduc-

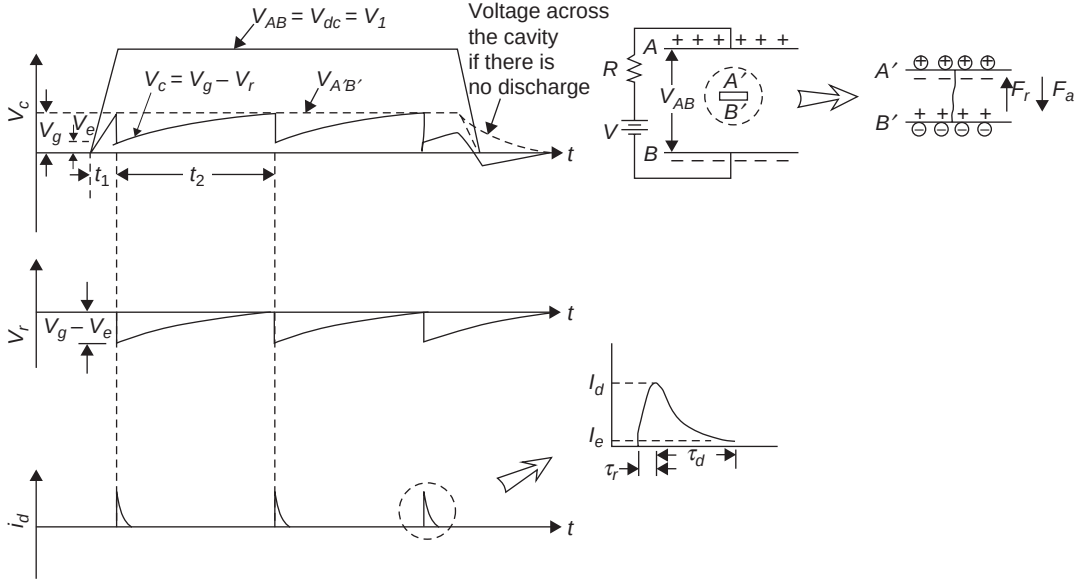


Figure 8-17 The variation of the voltage across the cavity V_c , the inverse voltage across the cavity V_r , and the discharge current i_d with time for a long-duration DC voltage applied across a dielectric solid specimen with a cavity. V_{dc} is the voltage across the specimen, which is equal to the discharge inception voltage V_i , V_g is the breakdown voltage of the cavity, which is the voltage across the cavity $V_{A'B'}$; V_e is the discharge extinction voltage in the cavity; F_a is the applied field, and F_r is the inverse field.

tion processes. Since the conductivity of dielectric solids is usually very small, it takes a long time for the charges to leak out and for the inverse voltage to reduce to zero, so that the voltage across the cavity can regain its level V_g for a second discharge to take place and the whole process to repeat, as shown in Figure 8-17. After the discharge extinguishes, the voltage across the cavity will build up again, following the equation^{39,40}

$$V_c(t) = V_{dc} \left(\frac{R_c}{R_b + R_c} \right) - \left(\frac{V_{dc} R_c}{R_b + R_c} - V_e \right) \exp\left(-\frac{t}{\tau}\right) \quad (8-14)$$

in which

$$\tau = \frac{R_b R_c}{R_b + R_c} (C_b + C_c) \approx R_b C_c = \frac{\epsilon_o}{\sigma} \frac{d_s}{d_g} \quad (8-15)$$

In general, $d_s > d_g$, $C_c > C_b$ and $R_c > R_b$ because R_c is mainly the surface resistance of the cavity when there is no discharge and all charges have leaked out.

From Equation 8-14, if $V_{dc} = V_i$, then theoretically it would take a very long time for V_c to reach $V_g = V_i [R_c / (R_b + R_c)] \approx V_{A'B'}$ again. Practically, t_2 , as shown in Figure 8-17, does not go to infinity before the occurrence of the second discharge, because the threshold voltage for the occurrence of subsequent discharges is usually smaller than for the first discharge, or V'_{dc} may sometimes go to a value slightly higher than V_g . However, it takes a very long time (which could be as long as a few hours to several weeks), and under certain conditions it may not happen again. This means that the discharge recurrence frequency f_r under DC stressing conditions is very low if $V_{dc} = V_i$ and usually many orders of magnitude smaller than that under AC stressing conditions at similar stressing fields. This is why partial discharges under DC voltage conditions have a much less damaging effect on the life of the insulation systems than under AC voltage conditions.

If the insulation system is subjected to over-voltages, $V_{dc} > V_i$, or to a high temperature, then

the time required for the occurrence of subsequent discharges t_2 becomes much shorter. In this case, from Equation 8-14, the discharge recurrence frequency (the number of discharges per second) is given by

$$\begin{aligned} f_r &= \left[\tau \ln \left(\frac{\frac{V_{dc} R_c}{R_b + R_c} - V_e}{\frac{V_{dc} R_c}{R_b + R_c} - V_g} \right) \right]^{-1} \\ &\approx \left[\tau \ln \left(\frac{V'_{dc}}{V'_{dc} - V_g} \right) \right]^{-1} \\ &\approx \left[-\tau \ln \left(1 - \frac{V_g}{V'_{dc}} \right) \right]^{-1} \\ &\approx \frac{\sigma}{\epsilon_o} \frac{d_g}{d_s} \left[-\ln \left(1 - \frac{V_g}{V'_{dc}} \right) \right]^{-1} \end{aligned} \quad (8-16)$$

where

$$V'_{dc} = \frac{V_{dc} R_c}{(R_b + R_c)} > V_g \quad (8-17)$$

and σ is the conductivity of the dielectric solid. V_e is assumed to be very small compared to V'_{dc} and may be neglected in Equation 8-16. From Figure 8-18 and Equation 8-16, it can be seen that f_r increases with increasing overvoltage.

The higher the dielectric solid's conductivity is, the higher the recurrence frequency. This is why f_r increases with increasing temperature, because σ increases with temperature.

The discharge magnitude Q_d is given by

$$\begin{aligned} Q_d &= \left[C_c + \frac{C_a C_b}{C_a + C_b} \right] (V_g - V_e) \\ &= \left[C_c + \frac{C_a C_b}{C_a + C_b} \right] \left[\frac{R_c}{R_b + R_c} \right] V_i \\ &\approx (C_c + C_b) \left(\frac{R_c}{R_b + R_c} \right) V_i \end{aligned} \quad (8-18)$$

since V_e is negligibly small and $C_a \gg C_b$. In Equation 8-18, C_b and C_c cannot be measured, so the partial discharge magnitude cannot be evaluated. However, we can measure C_a , which is practically equal to the capacitance of the whole test specimen because the cavity is very small. The voltage drop across C_a during discharge can be estimated by

$$\delta V_a = \frac{C_b}{C_a + C_b} (V_g - V_e) \approx (C_b / C_a) V_g \quad (8-19)$$

The apparent discharge magnitude generally used for the comparison of discharge magnitudes is

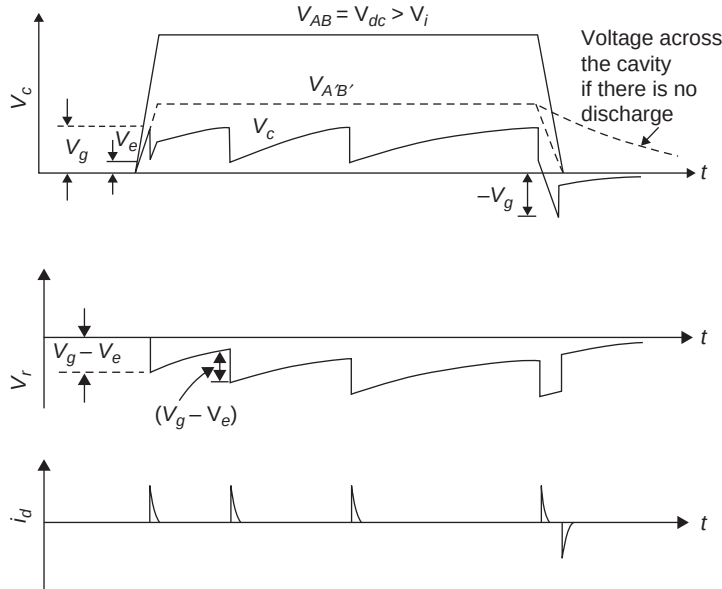


Figure 8-18 The variation of V_c , V_r , and i_d with time for a long-duration DC voltage applied across a dielectric solid specimen with a cavity for $V_{dc} > V_i$.

$$Q_{ad} = C_a \delta V_a = C_b V_g \quad (8-20)$$

which can be measured. The actual discharge magnitude can be much larger than the apparent one according to

$$Q_d \approx \frac{C_b + C_c}{C_b} Q_{ad} \quad (8-21)$$

which depends on the value of C_c .

The discharge current can be expressed as

$$i_d = \sum_{k=1}^n f_{rk} Q_{dk} \quad (8-22)$$

which is the sum of individual discharges inside the cavities; k denotes the particular discharge at site k .

The energy dissipated per discharge is

$$\begin{aligned} W_d &= \frac{1}{2} \left(C_c + \frac{C_a C_b}{C_a + C_b} \right) (V_g^2 - V_e^2) \\ &\approx \frac{1}{2} \left(C_c + \frac{C_a C_b}{C_a + C_b} \right) V_g^2 \\ &\approx \frac{1}{2} \left(C_c + \frac{C_a C_b}{C_a + C_b} \right) \left(\frac{R_c}{R_b + R_r} \right)^2 V_i^2 \end{aligned} \quad (8-23)$$

and the mean power dissipated per discharge is

$$P_d = W_d f_r \quad (8-24)$$

Case 2: AC Voltage Conditions

The equivalent circuit for a dielectric solid containing a cylindrical disk-shaped cavity is the same as that shown in Figure 8-16. The voltage across the cavity without discharges is simply given by

$$V_c = V_{ac} \left(\frac{C_b}{C_b + C_c} \right) = V_{ac} \left(\frac{\epsilon_r d_g}{d_s + \epsilon_r d_g} \right) \quad (8-25)$$

The voltage distribution is determined by the capacitance distribution; the resistance distribution may be ignored for AC voltage conditions. In this case study, we shall confine ourselves to cases involving only power frequency (50 or 60 Hz) or low frequency AC voltages.

If the peak value of the AC voltage is the discharge inception voltage of the specimen $V_{ac(\text{peak})} = V_i$ and $V'_{ac(\text{peak})} = V_g$, then the discharge will occur at the peak. During the discharge and after discharge extinction, an inverse voltage

will develop, opposing the applied voltage due to the separation of negative and positive charges left behind (in a manner similar to that under DC voltage conditions). When the voltage of the positive half-cycle decreases below zero and just turns to the negative half-cycle, the voltage across the cavity may become $-V_r - V'_{ac} = -V_r - V_e = -V_g$, and a reverse discharge will occur. This reverse discharge will neutralize the space charge left over by the first discharge during the positive half-cycle, reducing the inverse voltage V_r to zero. After that, the voltage across the cavity is practically equal to V'_{ac} , so as soon as V'_{ac} reaches the peak value of the negative half-cycle, a third discharge will occur, and so on, as shown in Figure 8-19(a).

On the basis of the assumption that the inverse field created by the separation of positive and negative charges produced by discharges does not decay with time between discharges, the minimum discharge repetition rate (or the recurrence frequency) should be four per cycle.^{32,41} The observed minimum discharge repetition rate, however, is two per cycle when $V_{ac(\text{peak})} = V_i$.⁴² The inverse field F_r , whose direction is opposite to that of the applied field, depends on the net charges accumulated on the upper and lower cavity surfaces. Since these charges will diminish, V_r will decay, and the assumption that F_r is unchanged between discharges is not physically realistic, particularly for AC stressing voltages.

There are two ways for these charges to diminish. One is due to leakage through conduction in the bulk of the dielectric solid, and the other is due to recombination (or neutralization) through diffusion of positive and negative charges in opposite directions along the cavity wall and inside the cavity, as shown in Figure 8-20(c). If the decay of V_r is taken into account, the minimum discharge repetition rate is two per cycle, as shown in Figure 8-20(b). This is in good agreement with experimental observation.

If the peak value of the applied voltage is higher than the discharge inception voltage $V_{ac} > V_i$, and $V'_{ac} > V_g$, then as soon as V_c reaches the breakdown voltage of the cavity V_g , a dis-

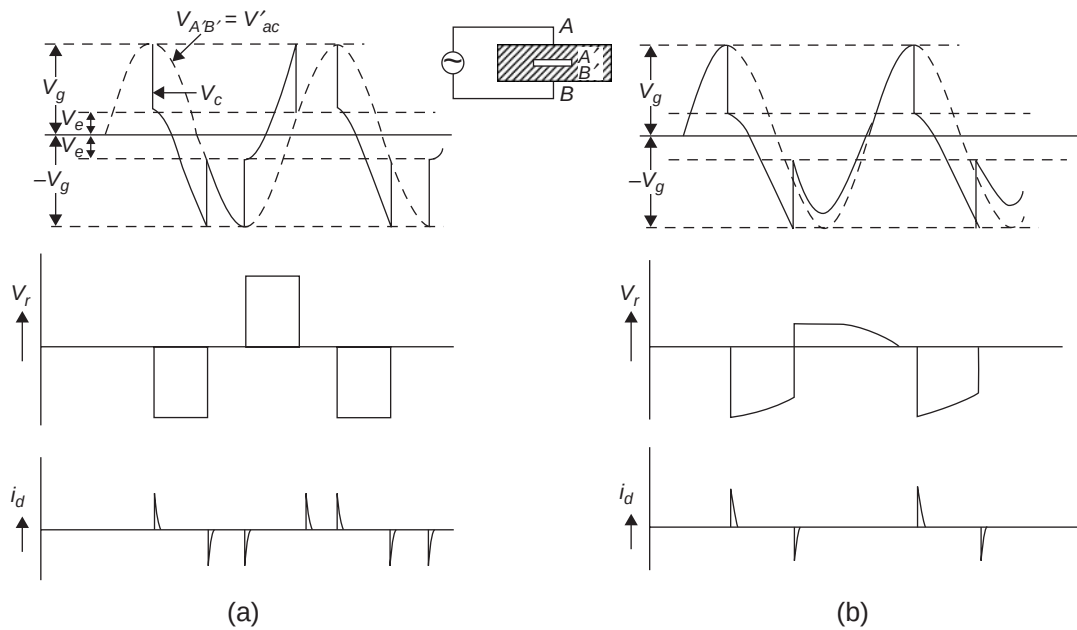


Figure 8-19 The variation of the applied AC voltage $V_{AB'}$, the actual voltage V_c , and the inverse voltage V_r across the cavity and the discharge current impulse i_d with time for the case with the applied voltage (peak value) across the cavity equal to the breakdown voltage of the cavity V_g (a) with the decay in cavity surface charge not taken into account, $\alpha = 0$ and (b) with the decay in cavity surface charge taken into account, $\alpha > 0$.

charge occurs. Then V_c drops to the discharge extinction voltage V_e within a time interval of about 10^{-7} s.⁴¹ This time is so short that it may be considered negligible compared to the duration of a half-cycle of the AC voltage.

After the discharge extinction at point *a*, as shown in Figure 8-20(a) and (b), the positive and the negative charges produced by this discharge will create an inverse field F_r . Obviously, the distribution of these charges on the upper and lower cavity surfaces varies following the variation of the AC voltage. The charge distribution at *a* is qualitatively illustrated in Figure 8-20(a). In the rising portion of the positive half-cycle, the charges tend to move to the center of the upper and lower cavity surfaces as the voltage rises. In this case, the major channel for the charges to diminish is leakage through conduction in the bulk. The bulk conductivity of the dielectric solid, such as polyethylene, is extremely low. Thus, it would take a very long time for the charges to diminish. During the time from *a* to *b*, shown in Figure 8-20(a), the

total charges and hence the inverse field can be assumed to be practically unchanged between discharges. This is very similar to what happens under DC stressing voltages. The charges under a DC stressing voltage will be held on the upper and lower cavity surfaces, and the only effective channel for them to diminish is leakage through conduction in the bulk.

The situation is completely different on the falling portion of the positive half-cycle and the rising portion of the negative half-cycle just prior to the occurrence of a discharge. Charges accumulated due to discharges in the rising portion of the positive half-cycle will diminish much faster by recombination due to the movement of positive and negative charges toward each other along the cavity wall and inside the cavity. The rate of charge diminution depends on the rate of change of voltage in these portions of the cycle. Therefore, V_r remains practically constant from *a* to *b* but decreases with time from *b* to *e* just prior to the occurrence of a discharge, as shown in Figure 8-20(a) through

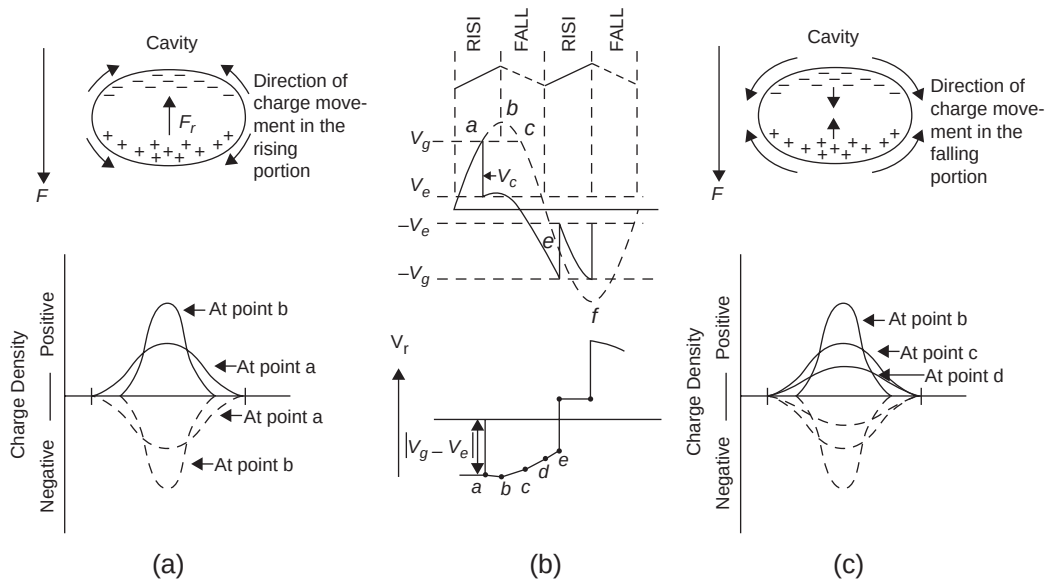


Figure 8-20 The mechanism responsible for the decay in cavity surface charge and the decrease in inverse field during one cycle of the AC stressing voltage when the peak value of the applied AC voltage across the cavity is larger than the breakdown voltage of the cavity V_g : (a) the cavity surface charge distribution in the rising portion of the AC cycle, (b) the variation of V_c and V_r with time, and (c) the cavity surface charge distribution in the falling portion of the AC cycle.

(c). This may also be the reason that most experimental results show discharges occurring mainly in the first and the third quarter-cycles, that is, on the rising portion of the positive half-cycle and on the rising portion of the negative half-cycle.⁴³

When the peak value of the applied AC voltage is larger than the discharge inception voltage of the specimen, after the discharge extinguishes at point 1, the voltage across the cavity V_c increases again because V'_{ac} increases, as shown in Figure 8-21. When V_c reaches point 2, $V_c = V_g$ again, and a second discharge occurs. At point 2, $V_c = V'_{ac} - V_r$, in which $V_r = V_g - V_e$ is the inverse voltage created by the cavity surface charges produced by the first discharge. Now, the second discharge also produces an inverse voltage $V_r = V_g - V_e$, so at and beyond point 2, the total inverse voltage $V_r = 2(V_g - V_e)$. Since V_r is so high, a third discharge will occur when $(-V_r + V'_{ac}) = -V_g$. This occurs at point 3 with the discharge current in reverse direction, as shown in Figure 8-21. Following the same process, we can predict when the fourth, fifth,

sixth, and so on discharges will occur and what polarity of the discharges will be.

Once a discharge occurs, the positive and negative charges move separately in opposite directions, following the field toward the upper and lower cavity surfaces. The amount of the charges that can be deposited firmly on the surfaces or trapped in surface states is time-dependent, because the first layer of charges arriving at the surfaces tends to repel the subsequent oncoming layers of charges, slowing their landing on the surfaces. Charge deposition is a time-dependent process, similar to dipole alignment and internal charge separation.⁴⁴ Before being trapped, the charges might remain relatively free entities on the surfaces for a short period. These charges should be quite free to move either tangentially along the cavity surfaces or normally in the space inside the cavity. It is likely that within the time between two discharges, most charges created by the discharges are still relatively free to move.

The literature on surface charge transport in polymers is scarce. However, it should be noted

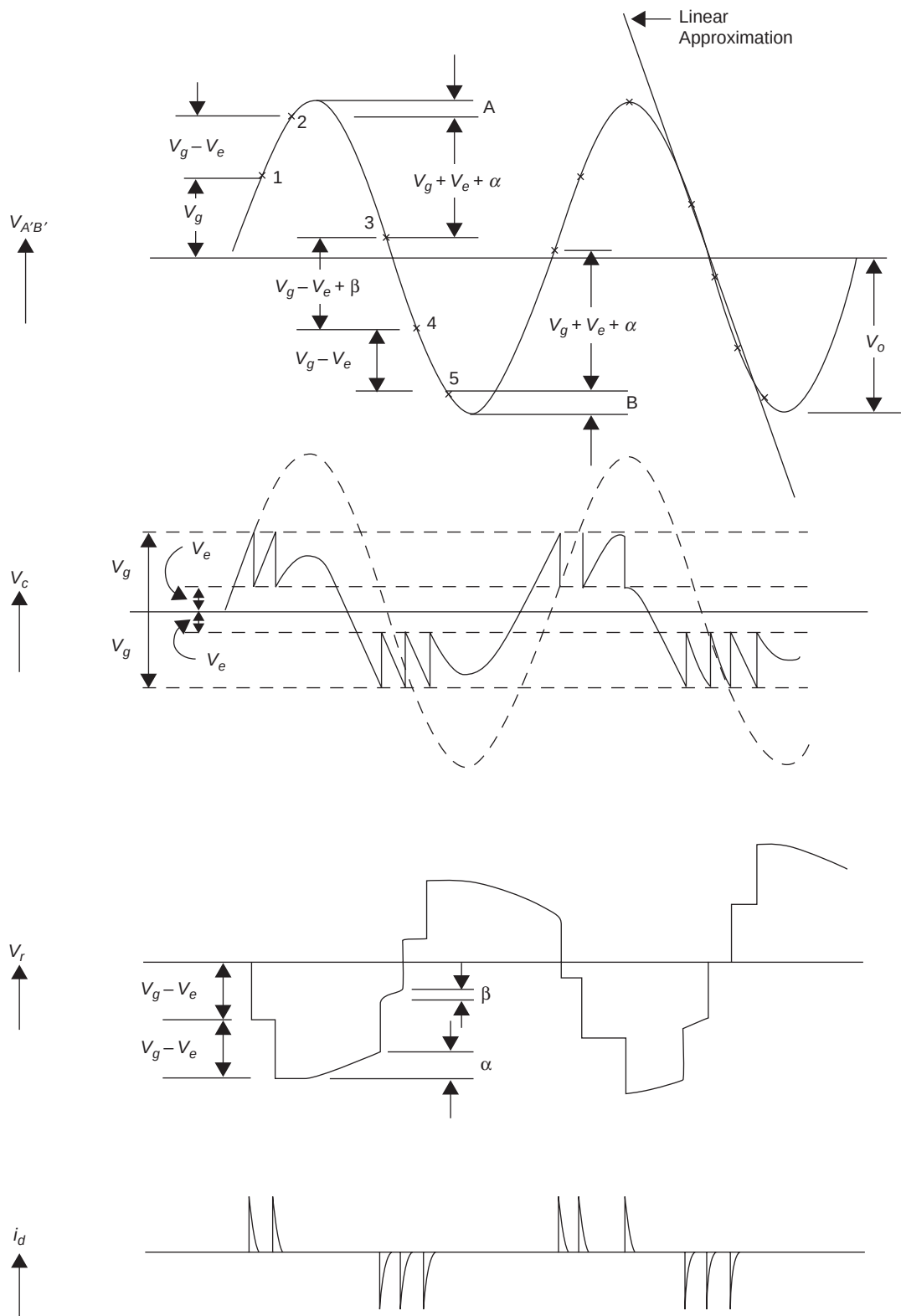


Figure 8-21 The variation of $V_{A'B'}$ (V_{ac}), V_c , V_r , and i_d with time when the peak value of $V_{A'B'}$ is much larger than V_g .

that the behavior of the charges produced by a discharge in the cavity under a time-varying AC field is different from the behavior of those produced by a corona discharge near a polymer surface without bias or under DC fields. Although there is currently no direct experimental evidence for surface charge behavior under time-varying fields, it is reasonable to believe that time is required for the discharge-produced charges to land on the cavity surfaces and to become trapped in surface states, and that within the time between two discharges a certain amount of charges is free to move.

It is assumed that the charge distribution remains practically unchanged from a to b and that the rate of charge diminution or the rate of inverse voltage decay during the time interval Δt from b to e , shown in Figure 8-20(b), follows a simple relation

$$V_r = V_{ro}(1 - Kt^2) \quad (8-26)$$

where V_{ro} is the inverse voltage at the point V_r starts to decay and K is a constant determining the rate of decay. For mathematical simplicity, the applied AC voltage during the time interval Δt is assumed to vary linearly with time by approximation and $A + B < V_g - V_e$, as shown in Figure 8-21. On the basis of these assumptions, we have derived an approximate expression for the discharge repetition rate f_r (the number of discharges per cycle for power frequency of 60 Hz):

$$f_r = \frac{4(V_o - V_e) - 2(A + B + \alpha)}{V_g - V_e} \quad (8-27)$$

where

$$\alpha \approx \frac{\left[377V_o - \sqrt{(377V_o)^2 - 2f_r K(V_g - V_e)} \right]^2}{2f_r K(V_g - V_e)} \quad (8-28)$$

and V_o is the peak value of the applied voltage across the cavity. For convenience, Equation 8-27 can be rewritten as

$$f_r \approx f_{ro} - \frac{2\alpha}{V_g - V_e} \quad (8-29)$$

where

$$f_{ro} \approx \frac{4(V_o - V_e) - 2(A + B)}{V_g - V_e} \quad (8-30)$$

When $K = 0$ and hence $\alpha = 0$, V_r does not decay with time and Equation 8-29 reduces to Equation 8-30, which is the conventional expression generally used.³² It can be seen from Equation 8-30 that if $\alpha = 0$, f_{ro} is at least equal to or larger than four per cycle. However, when $K > 0$, f_r is always less than f_{ro} .

For simplicity, we will take a case with $A + B = 0$ as an example. In this case, Equation 8-29 can be approximated to

$$f_r \approx \frac{f_{ro}}{1 + Kt^2} \quad (8-31)$$

where

$$t \approx (377)^{-1} \sin^{-1}(V_g + V_e)/V_o \quad (8-32)$$

Equation 8-31 clearly indicates that as K increases, f_r decreases. For $0 < Kt^2 < 1$, we have $f_{ro} > f_r > f_{ro}/2$. This implies that the minimum repetition rate can be only one-half the rate predicted by the conventional expression (Equation 8-30) and explains why for this case, with $V_o - V_e \approx 2(V_g - V_e)$, the steady-state repetition rate is six per cycle^{41,42} rather than eight per cycle, as predicted by Equation 8-30. This can also explain why the number of discharges occurring in the rising portions of an AC cycle is larger than that occurring in the falling portions.⁴⁴

Cavity surface resistance will change after prolonged exposure to discharges due to chemical contamination by the radicals produced by discharges.^{45,46} This change in surface resistance tends to reduce the voltage across the cavity if the applied voltage across the test specimen remains unchanged. This will reduce f_r , because f_r increases with increasing applied voltage across the cavity. However, the change of surface resistance after prolonged exposure to discharges is quite different from that caused by charge movement and recombination, which varies with time in the falling portions of the cycles. It is the charge movement and recombination that really controls the discharge repetition rate.

The discharge magnitude Q_d is

$$\begin{aligned} Q_d &= \left(C_c + \frac{C_a C_b}{C_a + C_b} \right) (V_g - V_e) \\ &\approx (C_c + C_b) \left(\frac{C_b}{C_b + C_a} \right) V_i \\ &\approx C_b V_i \end{aligned} \quad (8-33)$$

since V_e is negligibly small and $C_a \gg C_b$. In practice, the extinction voltage in most dielectric solids is from 0 to about 25% of the breakdown voltage of the cavity V_g . Since C_b cannot be measured and C_a can be measured, the apparent discharge magnitude is generally used for comparison purposes. During discharge, the voltage drop across C_a can be evaluated by

$$\begin{aligned} \delta V_a &= \frac{C_b}{C_a + C_b} (V_g - V_e) \\ &\approx \frac{C_b}{C_a} V_g \approx \left(\frac{C_b}{C_a} \right) \left(\frac{C_b}{C_b + C_a} \right) V_i \end{aligned} \quad (8-34)$$

since $V_g > V_e$ and $C_a > C_b$. The apparent discharge magnitude is thus

$$\begin{aligned} Q_{ad} &= C_a \delta V_a = \frac{C_b^2}{C_b + C_c} V_i \\ &\approx \left(\frac{C_b}{C_c} \right) C_b V_i \end{aligned} \quad (8-35)$$

From Equations 8-34 and 8-35, it can be seen that the actual discharge magnitude can be much larger than the apparent discharge magnitude.

The energy dissipated per discharge is given by

$$\begin{aligned} W_d &= \frac{1}{2} \left(C_c + \frac{C_a C_b}{C_a + C_b} \right) (V_g^2 - V_e^2) \\ &\approx \frac{1}{2} (V_g - V_e) C_b V_i \\ &\approx \frac{1}{2} Q_{ad} V_i \end{aligned} \quad (8-36)$$

The discharge inception voltage V_i is usually assumed to be the peak value of the AC voltage. Thus, in terms of the effective value (r.m.s. [root-mean-square] value), Equation 8-36 becomes

$$\begin{aligned} W_d &\approx \frac{1}{2} Q_{ad} [1.41 V_{i(rms)}] \\ &= 0.7 Q_{ad} V_{i(rms)} \end{aligned} \quad (8-37)$$

Discharges take place not just at one site, but at many different sites in the cavity, thus creating transverse fields along the cavity surfaces. Discharge recurrence processes will be affected by the charge transfer from one discharge site to another. Furthermore, consecutive discharges may also differ in size.

8.2.2 Electrical Treeing

Point-plane electrode configuration for the study of prebreakdown phenomena is used mainly to simulate the surface irregularities in high-voltage insulation systems, such as power cables, for quality-control evaluation. Internal partial discharge is a prebreakdown phenomenon. Partial discharge progresses through the material, producing damage paths, which look somewhat like a tree. This is why this phenomenon is sometimes called *electrical treeing*. We have mentioned that partial discharges may not start in the already existing voids inside the insulating material, particularly when the voids are very small. Partial discharges in this case may be initiated by other mechanisms. Here, we shall discuss the possible mechanisms.

Prebreakdown Disturbances and Light Emission

Using a Schlieren optical system,⁴⁷ several investigators have observed a prebreakdown disturbance phenomenon in condensed dielectric materials using a divergent field point-plane electrode configuration.^{48–52} For a point-plane electrode configuration, the electric field at the point in the absence of space charge may be estimated from the relation

$$F_p = \frac{2V}{r \ell n(4d/r)} \quad (8-38)$$

where V is the applied voltage, r is the tip radius of the point electrode, and d is the gap length between the point and the plane.^{38,53}

Point-plane electrode configuration is a good configuration for studying the development and behavior of disturbances at various fields from the highest field at the point to the lowest field at the plane. The Schlieren technique is a simple method for detecting any phenomenon

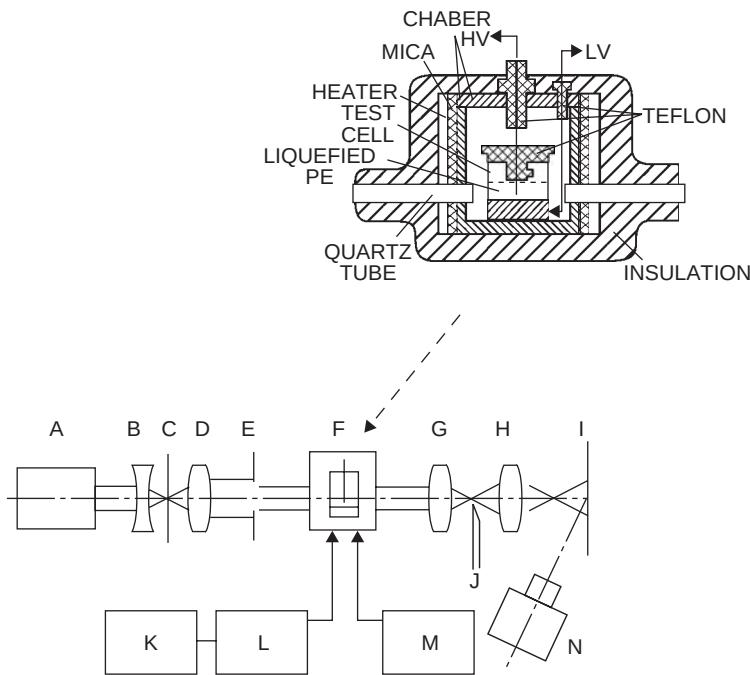


Figure 8-22 The Schlieren electro-optical system for observing the development of low-density regions in liquefied polyethylene. (A) He-Ne laser, (B) condenser lens, (C) pin hole, (D) collimating lens, (E) slit to control the size of the light beam entering the test region, (F) test chamber, (G) Schlieren lens, (H) magnifying lens, (I) screen, (J) knife edge, (K) storage oscilloscope, (L) current amplifier, (M) HV power supply, (N) camera.

involving a change in refractive index.⁴⁷ Thus, any prebreakdown phenomena can be observed using this technique, because these phenomena always involve changes in refractive index. Figure 8-22 shows the experimental arrangement of the basic Schlieren optical system and the test sample chamber for the observation of prebreakdown phenomena.

In low-viscosity dielectric liquids, electrical discharge or breakdown is always preceded by the growth of a disturbance whose refractive index is lower than that of the surrounding medium.⁴⁸⁻⁵² A great deal of work on prebreakdown phenomena has concentrated on liquids of low viscosities. However, a few studies have reported on high-viscosity epoxy fluid⁵¹ and on polyethylene at 145°C (liquefied polyethylene).⁵² Polyethylene is a typical insulating polymer widely used for insulation in many engineering systems and apparatus. Here, we will present only the results on liquefied poly-

ethylene as an example to demonstrate the consistency of observations with theory. These results may hint that a similar phenomenon may occur in all solid insulating polymers.

Figure 8-23 shows photographs of disturbances in electrically stressed polyethylene at 145°C (liquefied polyethylene). The refractive index of such disturbances has been measured experimentally to be smaller than that of the surrounding medium, indicating that the disturbances could be low-density regions (domains) which may be formed by large microvoids caused by electron trapping and recombination, as was discussed fully in Section 8.1. To develop an observable disturbance requires time, because such disturbances would not happen until the concentration of free radicals or the degree of structural degradation has reached a certain critical level. This is why t_0 time is required after the application of the stressing voltage for the appearance of a

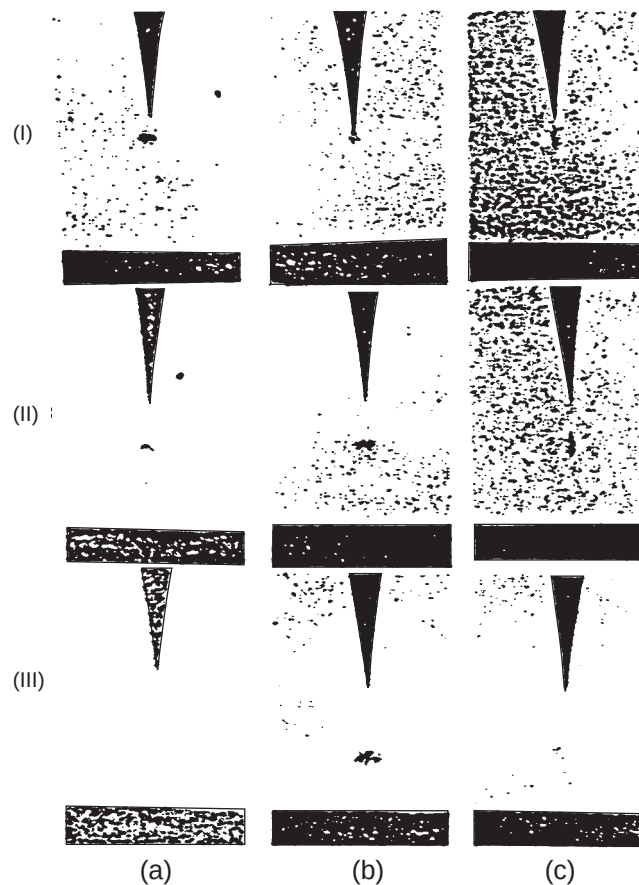


Figure 8-23 Photographs of the low-density regions in polyethylene at 145°C taken at t after the first appearance of the disturbance, which requires time t_0 after the application of the stressing voltage: (a) 6.5 kV AC (60 Hz, rms), $t_0 = 31$ min, (I) $t = 1$ s, (II) $t = 3$ s, (III) $t = 5$ s; (b) 10 kV DC (negative point), $t_0 = 25$ min, (I) $t = 0.5$ s, (II) $t = 4$ s, (III) $t = 7$ s; (c) 10 kV DC (positive point), $t_0 = 23$ min, (I) $t = 1$ s, (II) $t = 5$ s, (III) $t = 9$ s.

disturbance at the point electrode, as shown in Figure 8-23. The disturbance can be considered an entity with a dielectric constant (or refractive index) lower than its surrounding medium. Thus, there is a tendency for this entity to move from the high-field region at the point to a low-field region at the plane (see Electromechanical Effects in Chapter 2).

Using an image intensifier, investigators have observed light emission near the tip of the point electrode for a point-plane electrode configuration in electrically stressed insulating materials, such as hydrocarbon liquids,^{54,55} alkali halide crystals,⁵⁶ and polymers.^{52,57-60} This phenomenon can be attributed to the fact

that after the formation of low-density regions, impact ionization may occur in these regions because the electrons can have a larger mean free path to gain sufficient energy from the field near the point electrode. Once impact ionization occurs, recombination of positive and negative charges will follow. It is the radiative recombination of positive and negative charges that produces light emission.

Mechanisms and Characterization of Electrical Treeing

Electrical treeing has two distinct phases: the tree initiation phase, during which no

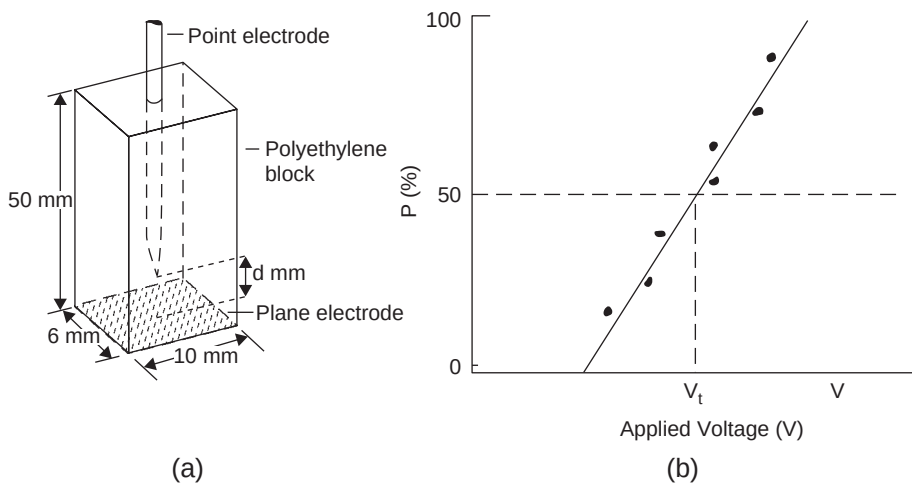


Figure 8-24 (a) Divergent field point-plane electrode configuration and (b) definition of the mean 50% initiation voltage for electrical treeing, V_t .

detectable partial discharge appears, and the growth phase, during which the tree propagates. It has been reported that luminescence occurs during the initiation phase.⁵⁸⁻⁶⁰ This indicates that a low-density region has been formed in the initiation stage because of the very high field at and near the point electrode tip. Heavy carrier injection from the point electrode to its vicinity results in the creation of the low-density region and the occurrence of impact ionization there, leading to the appearance of luminescence due to carrier trapping and recombination. This process will finally develop thermal instability in that region and extend from the electrode tip to a point where the electric field becomes so low that impact ionization ceases. This is the process by which a tree is formed.

The voltage required to initiate the formation of a tree, referred to as the *tree initiation voltage*, is usually measured as the mean 50% initiation voltage V_t . The value of V_t is determined by the following test procedure: The test is carried out with groups, each consisting of 10 or more specimens, using the electrode configuration shown in Figure 8-24(a). Different groups are electrically stressed at different applied voltages—one group at one voltage—for a period of one hour. After the stressing

period, each specimen of each group is optically examined with a microscope. The percentage of the specimens exhibiting treeing (P) at various stressing voltages is plotted versus the applied stressing voltage, as shown in Figure 8-24(b). When a line is drawn at $P = 50\%$, the interception of this line with the plotted curve defines the mean 50% initiation voltage for electrical treeing V_t , as shown in Figure 8-24(b). V_t depends on the material and the radius of curvature r of the point electrode tip, the electrode separation d , and the form of the applied stressing voltage.

Some Special Features of Electrical Treeing

After the formation of a low-density region, impact ionization may occur in the region only when the field, the mean free path, and the ionization energy of the radicals in the region are in appropriate values. Electrons travel from the electron-injecting point electrode in a divergent field in the low-density region and produce impact ionization. The avalanche may propagate to a point at which the field is not high enough for the electrons to gain sufficient energy to cause impact ionization. There, discharge ceases to form partial discharge

(electrical treeing). The breaking of bonds by hot electrons also creates more electron traps prior to the formation of low-density regions, because of the creation of a high density of structural defects. Harari²⁰ has reported that high field and high electron injections into SiO₂ films just prior to breakdown result in the generation of a very high density of defects, which behave electrically as stable electron traps, close to the injecting contact.

It is well known that electrical treeing is a main cause of breakdown of cables insulated with polyethylene (PE) or with chemically cross-linked polyethylene (XLPE). Electrical treeing has been observed in many insulating materials other than the polyethylene-based polymers, including rubber, resins, polymethyl methacrylate (PMMA), etc. Because of the ease of direct visual observation, however, transparent or translucent materials (such as PMMA and in particular PE) are always chosen for experimental studies.

The shape of an electrical tree depends on the magnitude of the applied stressing voltage, the waveform of the stressing voltage, the frequency of the AC voltages, and the structure and treatment of the material under investigation. Electrical trees may look like dendrites, branchlike or bushlike trees, spikes, strings, or bow ties. There is a great volume of literature dealing with various patterns of trees, techniques for investigating the initiation, growth and inhibition of treeing, and in particular, technical data for industrial use.^{61–68} Here, we shall discuss briefly some significant features of electrical treeing in PE.

Table 8-3 shows the mean 50% tree initiation voltage V_t and breakdown voltage V_b for PE. It can be seen that both V_t and V_b are much lower for AC voltages (50 Hz) than for DC voltages. For DC voltages, both V_t and V_b for the positive point are lower than for the negative point.⁶⁹ When the point is negative, electrons injected from the point move in a divergent deceleration field and quickly become trapped. These trapped electrons will immediately form a homo-space charge, tending to reduce the effective field at the injecting contact, suppressing further electron injection. At the same time, free radicals are created due to the release

Table 8-3 V_t and V_b of polyethylene under DC and AC voltages between point-plane electrodes.

	DC (kV)		AC (kV)
	+ Point	– Point	
V_t	37	54	6
V_b	39	100	33

The radius of the tip of the point electrode $r = 4 \mu\text{m}$, the gap length $d = 4 \text{ mm}$, and the rate of rise of the applied voltage = 2 kV/s.

of the energy from the trapping. This process will continue until the concentration of free radicals reaches a critical value. Then, a low density region will be formed and electron impact ionization will take place there, leading to the appearance of luminescence due to recombination and initiating the growth of electrical trees.

When the point is positive, the holes injected from the point also move in a divergent deceleration field, but they can travel a longer distance before being trapped because of their higher mobility in the valence band. Thus, the trapped hole space charge is not very close to the injecting contact, so the efficiency of suppressing further hole injection is not as good as when the point electrode is negatively biased. Furthermore, any electrons appearing near the positive point or near the path of hole movement will be accelerated in a convergent field. The action of both holes and electrons will promote the energy released due to their recombination and trapping, causing structural degradation more effectively.

When the point electrode is negatively biased, the electrons will be decelerated in a divergent field away from the point electrode, demoting impact ionization. This is why tree initiation voltage is lower and tree penetration length is longer when the point electrode is positively biased than when it is negatively biased. Figure 8-25 shows the patterns of electrical trees. The tree is highly ramified when the point electrode is negatively biased, and it is highly extensive when the point electrode is positively biased.⁷⁰ It should be noted that the probability of the formation of a low-density region near

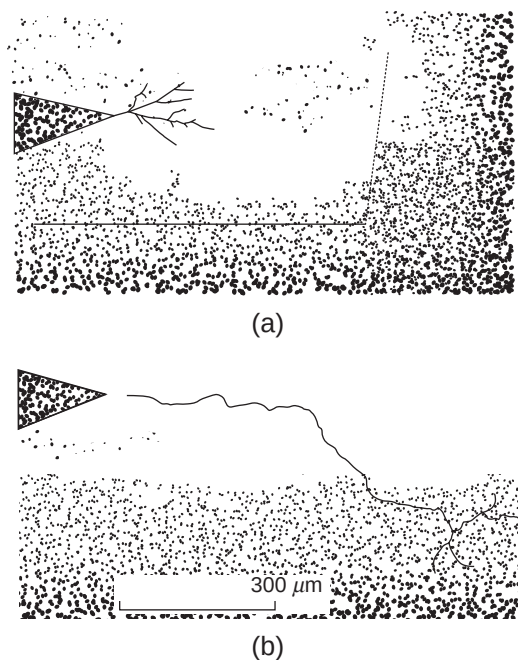


Figure 8-25 Electrical treeing in polyethylene with a point-plane electrode configuration (a) when the point is negative in polarity and (b) when the point is positive in polarity.

the point electrode tip is highest because chemical bonding is weakened at high fields.^{71,72}

For AC voltages, holes injected from the point electrode and captured by traps during the positive half-cycle will be recombined by the electrons injected from the point electrode during the negative half-cycle. Thus, the process of forming a low-density region in the vicinity of the point electrode is much more efficient under an AC field than under a DC field. This is why V_i and the breakdown voltage V_b under AC fields are much lower than under DC fields.

Inhibition of Electrical Treeing

The initiation of electrical treeing is associated mainly with nonuniform field distribution, which leads to carrier injection from sharp contact points and edges, and subsequently to energy evolved due to trapping and recombination. This results in structural degradation and the creation of low-density regions. To inhibit

electrical treeing, it is necessary to suppress this degradation process. Based on the causes of electrical treeing, the following methods may be used to inhibit its development:

- The use of emission shields to suppress the carrier injection: It has been found that with a point-plane electrode configuration, the initiation voltage for electrical treeing in polyethylene is much higher for the iron-point electrode tip that is oxidized (i.e., covered with a thin oxide layer) than not, because the iron oxide acts as an emission shield to suppress carrier injection.^{26,27}
- The use of suitable additives mixed with the polymer to improve nonuniform field distribution: Acetophenone has been used to dope polyethylene with the aim of increasing the effective electrical conductivity to improve field distribution. The major effect of additives in polyethylene is that electrical conductivity becomes strongly field dependent: the higher the field, the higher the conductivity. This, in turn, makes the field more uniform and increases the initiation voltage for electrical treeing.⁷³ However, this method has one important shortcoming: It creates electrical losses due to the increase in overall conductivity of the insulating polymers.
- The use of radical scavengers to reduce structural degradation: Since electrical treeing is an electrical aging process, the remedies for electrical aging can be used to inhibit electrical treeing (see Section 8.1.3).

There are other approaches, most using additives. Some additives are used to capture energetic electrons chemically; some are used to slow down the growth of discharge path electrically. Numerous compounds have been reported and patented to inhibit electrical treeing. However, their theory is complicated and still obscure, so we will not deal with them here. For more information, see references.^{61,67}

Electrical Treeing Initiated by Water Trees

Since the penetration of moisture into organic materials increases in the presence of an electric field, water trees usually grow from points

of high electrical stress, following the direction of the electric field into the surrounding insulation. It is well known that organic polymers are permeable to penetration by gas and moisture vapor. Water trees usually appear diffuse and often can be made to disappear by heating. Obviously, the presence of water trees can lead to the development of electrical trees, which may cause ultimate failure of the insulation system.

8.2.3 Surface Discharges and Corona Discharges

Both the surface and the corona discharges involve external discharges, implying that such discharges are associated mainly with gas discharges. These discharge phenomena lie outside the scope of this book, and therefore we shall not deal with this subject, but it is worth mentioning some of their features.

Surface discharges may occur when there exists a component of electric field parallel to the dielectric surface high enough to cause discharges. Usually surface discharges occur at the edges of electrical contacts. The behavior of surface discharges is similar to, though usually more complex than, that of internal discharges. There are many sites on the surface that may discharge at about the same voltage. Similar to internal discharges, surface discharges are intermittent. It can be imagined that external discharges can give rise both to surface breakdown and to penetration of the dielectric specimen. Surface breakdown means the development of a conducting channel between two electrical contacts on the dielectric surface. Surface breakdown may be divided into flashover and tracking. Flashover essentially involves a gas discharge but does not necessarily impair permanently the insulating properties of the surface. Tracking essentially requires a conducting channel on the dielectric surface, which would suffer damage in the tracking process.

Corona discharges occur around sharp points or edges at high voltages in air or other gases. They occur more readily at the negatively biased point electrode than at the positive one. For AC voltages, they occur often during the

negative half-cycle of the sinusoidal wave. For a point-plane electrode configuration in gas, corona discharges occur near the point, resulting in the formation of a positive ion space charge in the vicinity of the point. The positive ions may impinge the point electrode, releasing more electrons. According to the Townsend gas breakdown mechanism, this would produce a cloud of positive ions near the point and negative electrons away from the point, gradually forming a cloud of negative ions due to the attachment of the electrons to oxygen molecules. During the discharge process, radiation due to recombination takes place, producing photoionization toward the point and extending the ionized region laterally until the cathode spot is formed, from which the corona discharge emanates. This process will go on until enough negative space charge is accumulated to reduce the field near the point to a value too low to produce further ionization. Then, the discharge extinguishes. After extinction, the negative space charge moves to the positive plane electrode and becomes neutralized there. Then, the electric field rises, and the next discharge starts.

If the plane electrode is metallic, there are no deep traps on the electrode surface. Charges on it can move freely, so a discharge that takes place in an air gap between two metallic electrodes will release all the charges on the metallic electrode surfaces. Therefore, only one spark channel can be formed for each discharge. On the contrary, a dielectrode surface has a high surface resistance. The charges stored on the surface are sufficient to produce a high field along the surface. With an air gap between metallic and dielectric electrodes, surface charges cannot be released by a single discharge. Therefore, several discharges may occur simultaneously on different sites of the dielectric surface.

With an air gap between metallic plane and dielectric plane electrodes, Friedlander and Reed⁷⁴ have studied corona discharge phenomena using a metallic plane electrode covered with synthetic-resin bonded papers as the dielectric electrode. When the dielectric electrode is at the negative polarity, the positive charges left on the dielectric surface after a

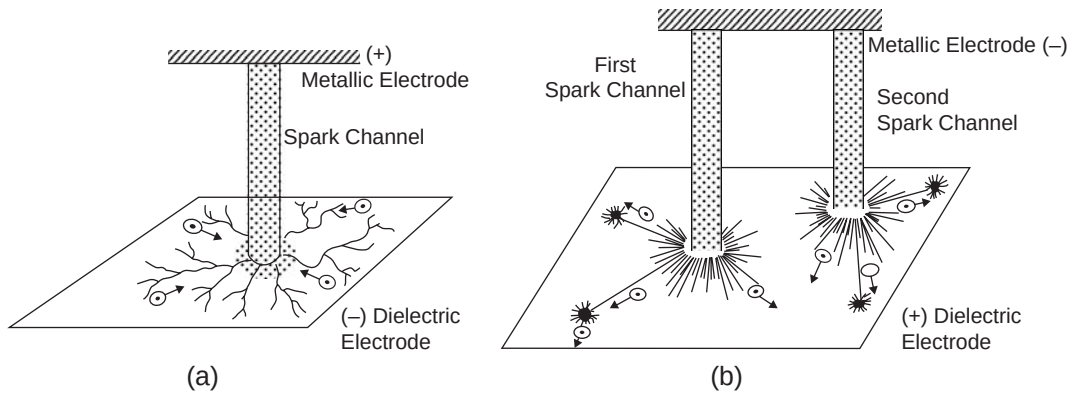


Figure 8-26 Schematic diagrams illustrating Lichtenberg figures of surface discharges on a dielectric electrode surface: (a) the dielectric electrode at negative polarity and (b) the dielectric electrode at positive polarity.

spark discharge create a tangential field, causing electron avalanches over the dielectric electrode surface, propagating toward the main spark discharge channel end on the surface. When the dielectric electrode is at the positive polarity, the electrons in a spark discharge channel reaching the dielectric surface will become trapped and form a negative surface charge, which creates a tangential field on the surface, tending to drive electrons away from the main spark channel end. These phenomena are shown schematically in Figure 8-26. However, the surface field created by trapped charges on the dielectric electrode surface tends to distort the applied field. This enhances the possibility that more discharges will occur simultaneously at other sites on the surface.

When corona discharges do not reach the surface of the dielectric material, the discharge may attack the material indirectly by forming ozone or other aggressive products. When the discharges reach the surface of the material, however, they may produce tangential fields on the material surface, resulting in surface discharges. In this case, the effect is similar to that of surface discharges mentioned earlier.

8.2.4 Detection of Partial Discharges

Electrical discharges are always accompanied by the production of one or more of the following phenomena: electric impulses, light,

heat, excess dielectric losses, gas pressure, noise, and chemical transformation. All techniques for discharge detection are based on the detection of one of these phenomena, so the sensitivity of a detection technique depends on the type of discharge. For example, if the discharge produces strong light and the light signal can easily be detected externally, then the light-detection technique should be adopted. On the other hand, if the discharges produce large electric current impulses, then a technique based on electric-impulse measurements should be used. Since the sensitivity required for detecting electrical signals is much higher than for detecting nonelectrical signals, the most commonly used detection techniques are based on the detection of electric current impulses or charges, unless the location of the high-voltage apparatus to be examined is inaccessible by electrical contacts, such as discharges in power transmission line insulators. In this case, infrared detectors are generally used. Partial discharge characteristics are important in determining voltage ratings and in choosing suitable insulating materials for particular apparatus and systems under a certain environmental condition. Many techniques can be used to detect partial discharges; this subject lies outside the scope of this book. Here, we will describe as an example the basic principle of one of the most commonly used techniques for detecting discharges under AC voltage conditions.

An internal partial discharge causes an increase in capacitance of the dielectric specimen. Since discharges are associated with only a small magnitude of charge, it is necessary to filter out the normal components of current and

voltage so that the discontinuous phenomena can be detected. The basic circuit of the discharge detector developed by Mole and Robinson⁷⁵ is shown in Figure 8-27(a), in which C_1 is the blocking capacitor, C_2 is the

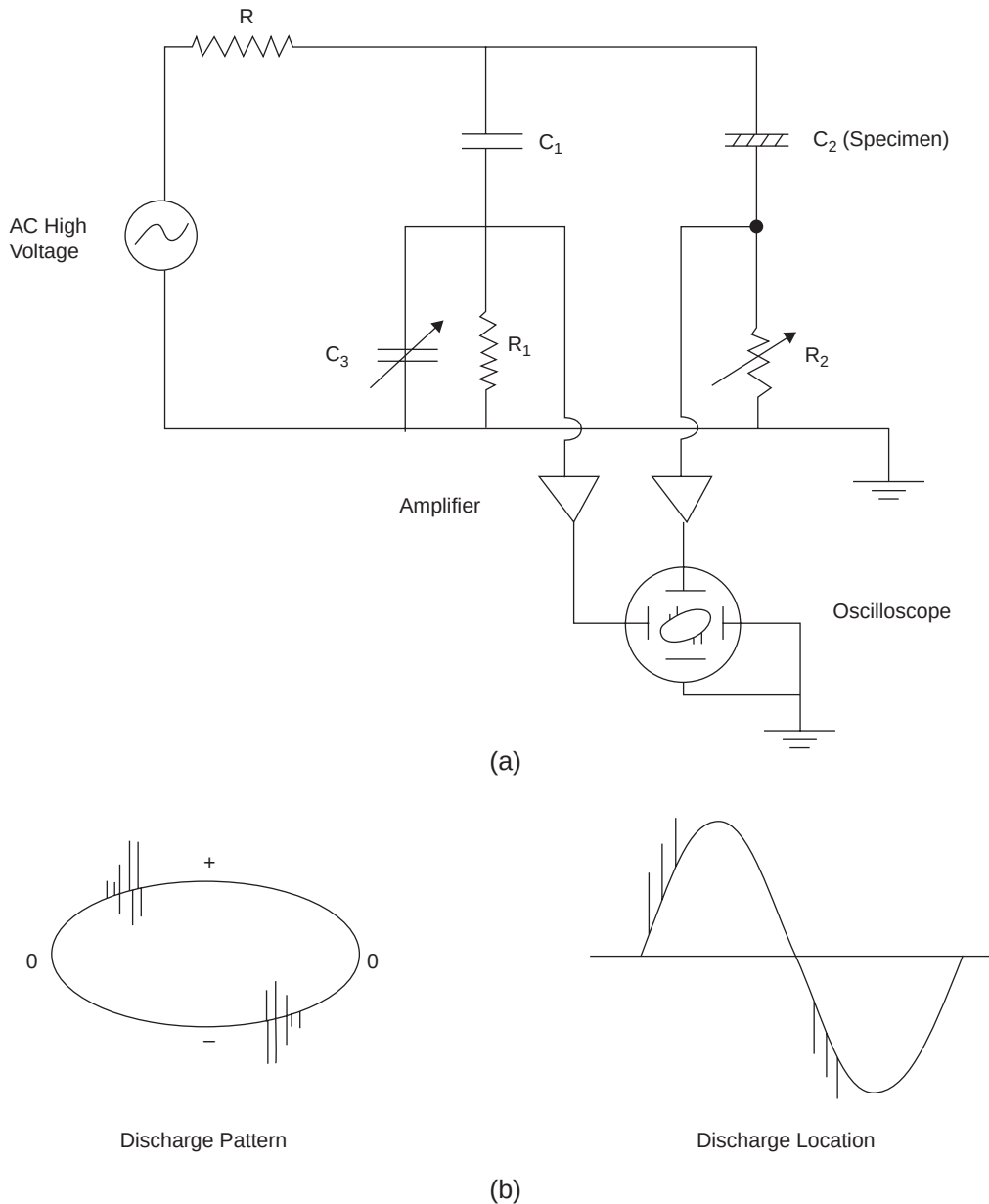


Figure 8-27 (a) The basic discharge detector circuit and (b) the typical discharge pattern and discharge location for internal discharges in cavities embedded in dielectric solids.

specimen being tested, R_1 is a fixed resistor, and R_2 and C_3 are, respectively, the adjustable resistor and capacitor.

The origin of the discharges may be determined on the basis of the discharge patterns observed. The phase angle and the polarity of an AC voltage at which a discharge occurs are the basic parameters for forming various discharge patterns. Two sinusoidal waves of the same frequency produce a Lissajous figure, which may be a straight line, an ellipse, or a circle, depending on the phase and the magnitude of the two waves. The basic principle of the detector is based on this concept by superposing the discharge pulses on a 60 Hz AC elliptic time base and displaying them oscillographically. The discharge magnitude is determined by injecting a known calibration pulse into the tuned circuit and adjusting the pulse to give the same response as the discharges. The location and characteristics of the discharge patterns can be used to distinguish the origins of the discharges.

For example, internal discharges in cavities embedded in a dielectric solid specimen give the discharge pattern and location, as shown in Figure 8-27(b). In this case, the discharge pattern is symmetrical and stationary. Discharges appear suddenly when the stressing voltage reaches a threshold value. The discharge magnitude increases rapidly, but this increase gradually diminishes as the voltage is increased.

When a discharge occurs, a fall of voltage suddenly occurs of a value q_c/C_2 , where q_c is the charge neutralized by the discharge. This voltage decays according to the relaxation time τ , which is given by

$$\tau = \frac{C_1 C_2 (R_1 + R_2)}{C_1 + C_2} = R_1 C_1 = R_2 C_2 \quad (8-39)$$

Only a small portion of the total voltage can usually be permitted across the low voltage arm R_1 , so $R_1 C_1$ is usually small. Therefore, τ is of the order of, say, 10^{-5} sec. If an amplifier and oscilloscope are used, which respond to frequencies from zero to well beyond $1/\tau$, then even a steep pulse can be recorded. For details about discharge detectors, see references.^{32,76}

8.3 Electrical Breakdown

Before going into this subject, it is worth reviewing briefly the mechanisms of electrical conduction and breakdown in gases, because there are similarities in the breakdown processes of gases and solids. Generally, for experimental studies of electrical breakdown phenomena, we measure the current-voltage (I - V) characteristics of a specimen up to the final destructive breakdown, whether the dielectric material is in gas, liquid, or solid phase. We shall start with gases.

8.3.1 Electrical Breakdown in Gases

The typical I - V characteristics of gases under normal temperature and pressure conditions are shown in Figure 8-28. In this figure, OA represents the current produced by the charge carriers already existing in the gas gap due to photoionization by photons from cosmic rays or other natural sources, and also to photoemission of electrons from the cathode. The electrons may be partly attached to gas molecules, forming negative ions. The current may be space charge-limited if the amount of the emitted electrons from the cathode is large. AB represents a current saturation region in which the current reaches a plateau. In this region, all charge carriers supplied to the gap from all sources are collected at the cathode and the anode. The external x-ray irradiation on the cathode produces additional current, so $I_{o2} - I_{o1}$ is the portion of current contributed by the x-ray irradiation. BC represents the current region, in which the increase in current is due to the carrier multiplication process associated with electron impact ionization.

When a sufficiently high field is applied to the electrodes, there are two types of discharges: non-self sustaining and self-sustaining. In the former, an initial source of electrons for ionization is required; in the latter, such an initial source of electrons is not necessary because electrons can be produced by the processes associated with intense ionization. Breakdown occurs when a non-self sustaining discharge makes a transition to a self-

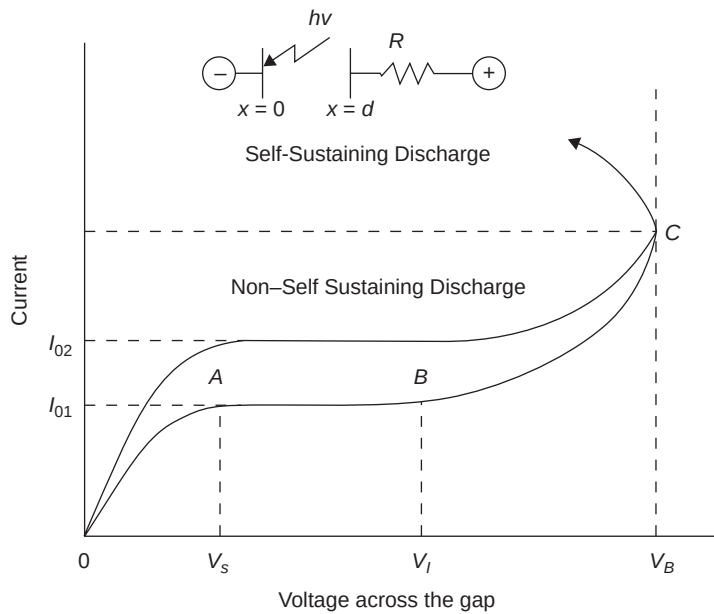


Figure 8-28 Typical current-voltage (I - V) characteristics of gases between two parallel plane metallic electrodes under normal temperature and pressure conditions with and without x-ray irradiation on the cathode.

sustaining discharge. The transition is very rapid and generally accompanied by a spark or light emission.

Once a self-sustaining breakdown occurs, the voltage across the gas gap drops (if the gas gap is connected in series with a resistance), as shown in Figure 8-28. This indicates that the gas becomes a high-conductivity material, which is a kind of plasma involving collisions of electrons, ions, and photons with gas molecules and electrodes. The most important process required to cause destructive breakdown is the feedback process leading to the formation of a self-sustaining discharge.

Impact ionization by electrons is necessary prior to destructive breakdown, but this process alone is not sufficient to cause breakdown. In fact, a non-self sustaining impact ionization avalanche is a nondestructive breakdown process; avalanche breakdown in p-n junction semiconductor devices and impact ionization in p-i-n semiconductor devices are, in fact, based on this nondestructive carrier multiplication process.

In the BC region in Figure 8-28, the rate of current increase from B onward increases rapidly with increasing applied voltage. This is the process of field-enhanced impact ionization by electrons, generally referred to as the α process; α is known as *Townsend's first ionization coefficient*, defined as the average number of ionizing collisions by one electron per unit length in the direction of the field. Obviously, α is a function of the cross-section for ionization, which depends on the density of gas molecules (or the gas pressure p), and of the average electron energy, which depends on the applied field F (or the energy gain, which is $qF\lambda$, where λ is the mean free path). Since α depends on p and F , the simple and convenient way to express the relation between α and F is

$$\begin{aligned}\alpha/p &= f(F/p) \\ &= A \exp(-B/F/p)\end{aligned}\quad (8-40)$$

where A is a constant characterizing a saturation value for α/p at large values of F/p and B is a parameter proportional to an effective ionization potential for the ionization process,

including excitation losses, etc.^{33,77,78} Equation 8-40 is valid only for a limited range of F/p .

Typical experimental results of α/p as a function of F/p are shown in Figure 8-29. The experimental maxima of α/p are of the order of 10 ion pairs/cm, mmHg, which correspond to the maxima of the ionization efficiency that occurs at the energy range of the primary electrons between 80 and 120 eV.⁷⁸ This implies that the concept that ionization efficiency is larger for gases with lower ionization potential is not generally true, because the ionization efficiency depends on F/p .

Several factors cause the occurrence of the maxima of α/p , including the following:

- The disturbance of the electric field by space charges causes α to vary from point to point in the discharge path.
- The probability of ionization by an electron with energy larger than the ionization energy is not unity. In fact, only a fraction of all collisions with the ionization energy produces ionization, because some inelastic collisions

produce only excited atoms or molecules, which lowers the value of α .

- The number of collisions made by an electron is much larger than λ^{-1} per unit length of its moving in the direction of the field, because many collisions are elastic.

As mentioned earlier, a feedback process is required to produce a self-sustaining discharge leading to destructive breakdown. There are two major self-sustaining discharge mechanisms: the Townsend mechanism and the streamer mechanism. These mechanisms will be discussed separately.

Townsend Discharges

Avalanches by one or more primary electrons starting from the cathode will not cause self-sustaining discharge. They produce only a high current, according to the following relation:

$$I = I_0 \exp(\alpha d) \quad (8-41)$$

where I_0 depends on the number of initial primary electrons available at the cathode. The Townsend discharge is one type of self-sustaining breakdown, which requires a secondary process to boost the current to make the ionization channel a highly conducting plasma. This secondary process must be created by the first ionization process. This means that the process is able to produce a secondary electron from the cathode; this secondary electron produces a second electron avalanche, and so on, without the need for a supply of electrons from external sources. Thus, the discharge can sustain itself.

This secondary electron can be produced by the following mechanisms:

Bombardment of positive ions on the cathode surface:

With this mechanism, the impinging positive ion must knock out two electrons. One is used to neutralize the ion itself and the other is ejected, producing an electron avalanche. Thus, for this secondary mechanism, the positive ion must have a total energy (kinetic and potential) equal to about twice the work function for electron emission.

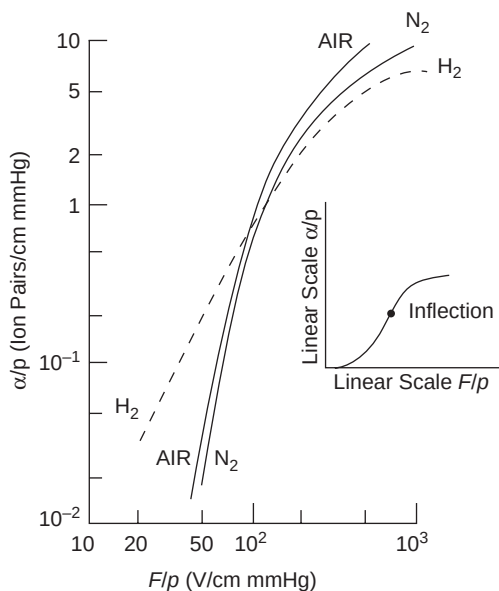


Figure 8-29 First Townsend ionization coefficient in α/p as a function of applied electric field in F/p for air, N_2 , and H_2 .

Bombardment of metastable atoms or molecules on the cathode surface: These atoms or molecules are excited by elastic collisions by energetic electrons during the α process.

Impact on the cathode surface of a photon with energy $h\nu$ larger than ϕ_m (work function): The photons may be produced by radiative transition of excited atoms from their excited state to their ground state or by radiative recombination.

The current involving the secondary process is given by

$$I = I_0 \frac{\exp(\alpha d)}{1 - \gamma[\exp(\alpha d) - 1]} \quad (8-42)$$

where γ is called *Townsend's second ionization coefficient*, defined as the average number of secondary electrons produced at the cathode per one ionizing collision (per each electron generated) in the gap.³³

The Townsend discharge criterion is that when breakdown occurs, I approaches infinity. For gases that are not electronegative, the condition for breakdown is

$$\begin{aligned} \gamma[\exp(\alpha d) - 1] &= 1 \\ \gamma \exp(\alpha d) &= 1 + \gamma \approx 1 \end{aligned} \quad (8-43)$$

since γ is normally very much less than unity. Equation 8-43 implies that the Townsend discharge requires one secondary electron to be produced in each electron avalanche, so that this secondary electron can initiate another

avalanche, and so on. Townsend discharge processes are shown in Figure 8-30.

There are two time lags in the Townsend discharge: the statistical time lag t_s , defined as the time elapsed between the application of a voltage and the appearance of a suitable primary electron, and the formative time lag t_f , defined as the time elapsed between the appearance of such a primary electron and the final spark breakdown.

Townsend discharge is dominant in the gas gap between two parallel plane electrodes or between two spherical electrodes with the radius of the spheres larger than $10d$ (gap length), with the pd product up to about 8000 cm mmHg, which is equivalent to a gap length $d = 8000/760 = 10$ cm at one atmospheric pressure.⁸⁰ However, the Townsend mechanism is no longer sufficient to explain the following facts:

- At very high overvoltages, the formative time lag t_f may be reduced to extremely low values, possibly even smaller than the transit time for an electron avalanche (e.g., 0.6×10^{-9} sec for $d = 1$ mm).
- The breakdown voltage of air at a pressure of several atmospheres does not vary with the nature of the cathode.
- The discharge is no longer diffuse in the long gap breakdown, as in the case of the Townsend discharge, but is concentrated in a narrow canal with branches and abrupt changes in direction.

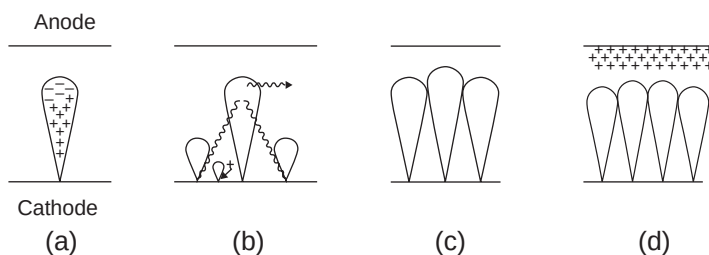


Figure 8-30 Townsend discharge processes. (a) Initial electron avalanche with a cone shape due to radial diffusion; (b) three possible mechanisms for generating a secondary electron from the cathode: positive ion bombardment, excited atom or molecular bombardment, and photon impact; (c) many electron avalanches created by secondary electrons; and (d) superposition of many electron avalanches to form highly conducting plasma.

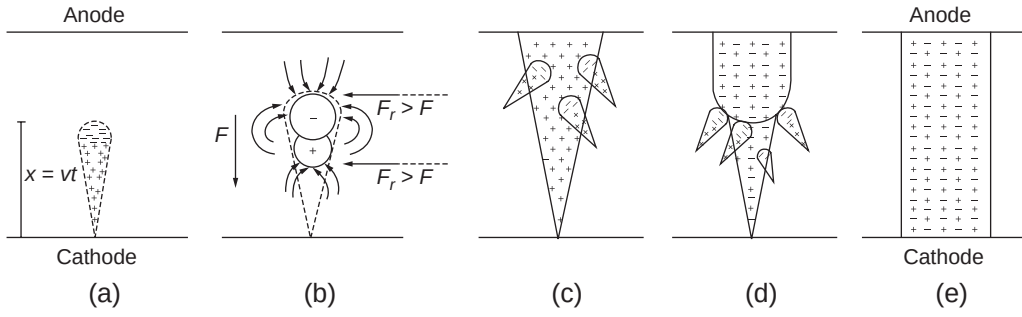


Figure 8-31 Development of a cathode-directed positive streamer: (a) single electron avalanche, (b) the radial field created by space charges ahead of and behind the avalanche, (c) auxiliary avalanches in the main avalanche, (d) a positive streamer propagates from the anode toward the cathode, and (e) the streamer bridges the gap.

- For a long gap, the time lag is much shorter than the time required for a positive ion to traverse the gap, since the mobility of the positive ions is of the order of 10^{-2} to 10^{-4} . The time of the mobility of the electrons. This observation is not consistent with the Townsend mechanism.
- In the path of the discharge canal, the charge density will strongly perturb the initial field. In the derivation of Equations 8-41 through 8-43, the field in the gap is assumed to be uniform, ignoring the effect of space charges, which is important in long-gap breakdown processes.

Streamer Discharges

The secondary process for creating a secondary electron for Townsend discharges is not likely to happen in nonuniform field positive point-plane gaps of large gap lengths or in lightning discharges between a cloud and ground. In these cases, a completely new mechanism is required to produce a self-sustaining discharge. This new mechanism is the streamer mechanism, developed by Meek and Loeb in 1940 for positive streamers and independently by Raether for negative streamers.^{79,80} This mechanism involves a single avalanche in which the space charge field developed ahead of and behind the avalanche itself creates

branch avalanches, which transform the main avalanche into a plasma.

For positive streamers, the criterion for the transition of an electron avalanche to a streamer is that the radial field created by the positive space charge reaches a value close to the average applied field. This radial field tends to enhance the field ahead of and behind the avalanche but to reduce the field in the middle, as shown in Figure 8-31(b). By assuming that the space charge takes the form of a sphere of radius r , the radial field is given by

$$F_r = \frac{Q}{4\pi\epsilon_0 r^2} = \frac{(4\pi r^3 N q / 3)}{4\pi\epsilon_0 r^2} = \frac{r N q}{3\epsilon_0} \quad (8-44)$$

where Q is the total positive charge in the sphere and N is the density of positive ions. In an elementary distance, dx at the end of an avalanche path x , the number of ions produced is $\alpha \exp(\alpha x) dx$. Thus, N can be written as

$$N = \frac{\alpha \exp(\alpha x) dx}{\pi r^2 dx} = \frac{\alpha \exp(\alpha x)}{\pi r^2} \quad (8-45)$$

The radius r is that of the avalanche after it has traveled a distance x from the cathode. On the basis of the diffusion equation, r is given by

$$r = (2Dt)^{1/2} = (2Dx/v)^{1/2} \quad (8-46)$$

where D and v are, respectively, the diffusion coefficient and the drift velocity of the ions. Substitution of Equation 8-46 into Equation 8-44 gives

$$F_r = \frac{\alpha q \exp(\alpha x)}{3\pi\epsilon_o(2Dx/v)^{1/2}} \quad (8-47)$$

The criterion for the onset of a streamer is that F_r is equal to the average applied electric field. Under this condition, the radial field is sufficiently high to suppress the expansion of the main avalanche and to initiate many secondary avalanches by electrons produced by photoionization of the gas molecules, due to the photons emitted by radiative recombination and transition from the dense ionized gas. In fact, the whole gas and the cathode are subjected to a shower of photons of various energies from the dense ionized gas region. The small, secondary avalanches are not directly associated with the breakdown process, but the electrons photo-generated from such avalanches near the channel of positive ions, and especially near the anode, will be drawn by the enhanced field (applied field plus the space charge field) into the positive ion core, making it a highly conducting plasma (positive ions and electrons), which starts at the anode.

The positive ions left behind produce secondary avalanches. In the same way, this process extends the plasma toward the cathode, as shown in Figure 8-31(c) and (d). This process is generally referred to as a *self-propagating positive space charge streamer*. The velocity of the propagation of the streamer depends on photoionization in the gas ($\sim 10^{-8}$ sec), photon propagation ($\sim$ light velocity), and short-distance travel of the electrons at high fields near the space charge region. It is more rapid than the velocity of the initial electron avalanche, which is about 10^7 cm/sec, and it is of the order of 10^8 cm/sec.⁷⁹

A similar criterion for the formation of negative streamers has been proposed by Raether⁸⁰: a negative streamer will start when the initial avalanche produces a sufficient number of electrons, $\exp(\alpha x)$, to give a field due to the space charge comparable to the applied field. This space charge field is given approximately by

$$F_r = \frac{q \exp(\alpha x)}{4\pi\epsilon_o r^2} \quad (8-48)$$

where r is the radius of the avalanche head, which is about the same as that given by Equation 8-46. Raether has proposed that the sufficient number of electrons is that the αd value is of the order of 20 corresponding to $\exp(\alpha d)$ to be of the order of 10^8 in one initial avalanche. Both the criterion and the formation process of a negative streamer are similar to those for a positive streamer. The secondary avalanches initiated by electrons created by photoionization in the gap and the enhanced field (applied field plus space charge field) are formed ahead of the negative streamer, which has already been formed, as shown in Figure 8-32.

If the gas gap is connected in series with a high resistance, then once a discharge occurs, the voltage across the gap will greatly decrease, as shown in Figure 8-28. If the series resistance is very large, then the voltage of the gas gap after breakdown may drop to a level at which the self-sustaining discharge cannot be maintained, and the discharge ceases. This situation is similar to a gas-filled cavity embedded in a high-resistivity dielectric material. The applied voltage to cause cavity breakdown is called the *inception voltage* for internal partial discharges, and the voltage at which the discharge extinguishes is called the *discharge extinction voltage*.

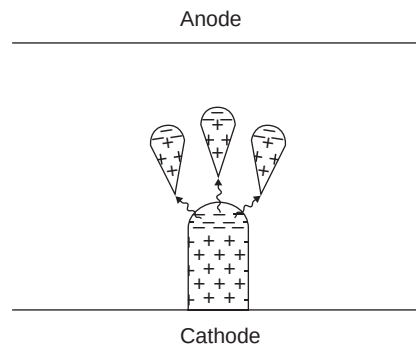
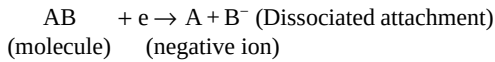
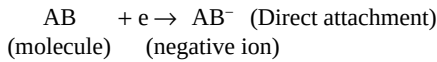


Figure 8-32 Anode-directed negative streamer.

Discharges in Electronegative Gases

If the gas is electronegative, or the gas medium is a mixture of nonelectronegative and electronegative gases, then the breakdown strength of this gas is higher, because the negative gas molecules have a high electron affinity to form low-mobility and large-mass negative ions. The electron attachment tends to remove electrons that may otherwise initiate avalanches. The electron attachment processes are



where A may be a carbon or sulfur atom and B is usually a halogen atom (electronegative). Sulfur hexafluoride (SF₆) is an electronegative gas that has been investigated extensively.

The expression for the criterion for the formation of Townsend or streamer discharges can be directly applicable to electronegative gases if the contribution of electron attachment is taken into account. By denoting β as the electron attachment coefficient, which is defined as the average number of electron attachments made by one electron traveling a unit distance along the field direction, Equation 8-42 can be modified to

$$I = I_0 \frac{\left[\frac{\alpha}{\alpha - \beta} \exp(\alpha - \beta)d - \frac{\beta}{\alpha - \beta} \right]}{1 - \gamma \left(\frac{\alpha}{\alpha - \beta} \right) [\exp(\alpha - \beta)d - 1]} \quad (8-49)$$

and Equation 8-43 to

$$\frac{\gamma\alpha}{\alpha - \beta} [\exp(\alpha - \beta)d - 1] = 1 \quad (8-50)$$

for the criterion for Townsend discharges.⁸¹ Of course, the condition for breakdown is $\alpha > \beta$. It can be imagined that α increases but β decreases with increasing F/p . If $\alpha = \beta$, there is a critical value of F/p , below which no discharge is possible.

In the same way, the criterion for the formation of a positive streamer can be modified to

$$F_r = \frac{(\alpha - \beta)q \exp(\alpha - \beta)x}{3\pi\epsilon_0(2Dx/v)^{1/2}} \approx F_{av} \quad (8-51)$$

and for the formation of a negative streamer to

$$F_r = \frac{q \exp[(\alpha - \beta)x]}{4\pi\epsilon_0(2Dx/v)} \approx F_{av} \quad (8-52)$$

Paschen's Law

Following Equation 8-40, we can also write

$$\gamma = \psi(F/p) \quad (8-53)$$

For uniform fields, the breakdown voltage $V_b = Fd$. Thus, the criterion for Townsend-type breakdown given by Equation 8-43 can be rewritten as

$$\psi(V_b/pd) \exp[pd\psi(V_b/pd)] = 1 \quad (8-54)$$

This equation implies that V_b depends only on the product pd , so we can write

$$V_b = f(pd) \quad (8-55)$$

which is known as Paschen's law.⁸²

Equation 8-55 gives no indication of the form of the function f . It is possible to obtain theoretical forms based on the expression for α such as that given by Equation 8-40, but these forms have limited validity because many factors may be associated with the α process. However, for values of pd within a certain range, V_b varies nearly linearly with pd for many gases consisting either of elemental atoms or molecules or of a mixture of several different gases. V_b has a unique minimum at $(pd)_{\min}$, which is generally called the *minimum sparking or breakdown potential*, as shown in Figure 8-33.

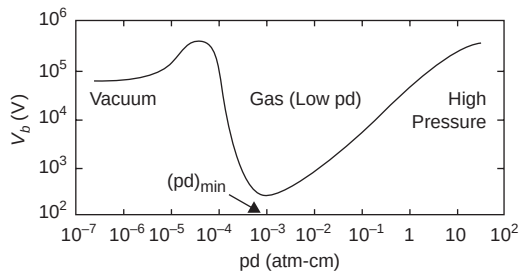


Figure 8-33 Paschen's law for breakdown voltage in nitrogen V_b as a function of the product of pressure and electrode separation pd .

Paschen's law is valid over a wide range of pd values, from about 5×10^{-5} to about 3 atmospheres for a 1 cm gas gap. The existence of a minimum spark potential can be explained by the fact that for $pd > (pd)_{\min}$, the number of collisions made by an electron is higher than at $(pd)_{\min}$, but the energy gained between collisions and the probability of ionization at a collision are lower because of a shorter mean free path (unless the applied voltage is increased). When $pd < (pd)_{\min}$, the collision frequency is low, so sufficient ionization can be maintained only by increasing the probability of ionization at each collision. Consequently, the electron velocity and hence the applied voltage must be increased.

At values of pd higher than this range, however, the breakdown voltage is somewhat higher for smaller gap length d for the same value of pd . This departure is probably associated with the transition from the Townsend mechanism to the streamer mechanism.

It is now desirable to have a clear definition of electrical breakdown. Electrical breakdown is an unstable, irreversible, and transient phenomenon. Its discontinuous occurrence indicates the transition from one more or less stable state of electrical conduction between two electrodes to another. The breakdown in gases is accompanied by a spark, which usually occurs suddenly. Once this occurs, the voltage across the gas gap drops by a process that produces a high conductivity channel between electrodes. In fact, electrical breakdown in any material, whether in the gas, liquid, or solid phase, must be caused by a unified process producing an abnormally high conductivity channel between electrodes.

How is such a high conductivity channel built up? Conductivity is proportional to the product of the concentration and the mobility of all charge carriers (electrons and ions). Therefore, it is necessary to find a mechanism that can lead to a continuous increase in carrier multiplication. The Paschen's curve shown in Figure 8-33 can be divided into three regions. We will discuss briefly the possible mechanisms in these three regions to produce continuously increasing carrier multiplication.

Gas Region (Low pd Values)

In this region, the primary process is impact ionization by electrons in gases. Electrons may originate from the cathode by photoemission directly by photons from cosmic rays or indirectly by ultraviolet light generated by photoionization of neutral molecules or atoms by photons from cosmic rays. The primary process produces electron avalanches, starting from the cathode, due to electron impact ionization of gas molecules in the gas gap when the electrons can gain sufficient energy within the mean free path λ from field F to cause ionization. For example, the ionization energies for nitrogen (N_2) and oxygen (O_2) are, respectively, 15.5 and 12.2 eV, and the mean free paths for N_2 and O_2 are, respectively, 6.28 and 6.79 μm . Thus, at normal pressure and temperature conditions a voltage of about 10^5 V is required to cause electrical breakdown for a 1 cm gas gap.

However, electron avalanches alone do not produce continuously increasing carrier multiplication. They can produce only non-self sustaining discharge, which may not lead to final destructive breakdown. To produce continuously increasing carrier multiplication, it is necessary to have a feedback process that leads to self-sustaining discharge. For Townsend breakdown, the feedback process produces additional electrons at the cathode from the primary avalanches. For streamer breakdown, the feedback process is a streamer process, which was discussed in earlier in Section 8.3.1. It can be concluded that electrical breakdown in gases involves mainly impact ionization of neutral molecules by electrons. This occurs whenever the $F\lambda$ product reaches a value higher than the ionization potential of the neutral molecules.

In the two extreme regions, vacuum and high pressure, Paschen's law does not hold. Although Paschen's law was derived on the basis of the Townsend breakdown mechanism, it still holds for the streamer breakdown mechanism, although the curve for higher pd products departs from the portion for lower pd products.

Vacuum Region

In a vacuum, the mean free path of the molecules is larger than the gap length between two

electrodes, so the concentration of residual gas molecules is so small that these may not play a major role in the creation of a high conductivity channel for an electrical spark.

In that case, how can a breakdown occur in a vacuum? At fields higher than 0.1 MV cm^{-1} , electron emission from the cathode will take place. Electrons emitted from small asperities on the cathode will be accelerated in the field. By the time they bombard the anode surface, two effects may arise: one is that positive ions are knocked out from the anode and then are accelerated traveling toward the cathode; the other is that when these positive ions impinge on the cathode, they will cause secondary emission of electrons from the cathode and also local heating of the cathode by bombardment, which produces clumps of metal vapor. When such clumps (some may be attached to electrons, forming negative ions) bombard the anode, they will, in turn, cause more positive ions to be knocked out from the anode and also local heating of the anode. Such an interchange of particles (electrons, positive ions, and negative ions) due to a kind of a feedback process will lead to the creation of a high conductivity channel.^{83,84}

In short, electrical breakdown in a vacuum involves electron injection from the cathode as the primary process. Subsequently, positive ions are knocked out from the anode and local heating of the electrodes occurs, producing clumps of metal vapor in which impact ionization may take place. The combination of these interactions continuously enhances carrier multiplication, until the final self-sustaining discharge occurs and breakdown condition is reached.

High Pressure Region

When the values of the pd product are larger 10 atm-cm , or when the pressures are higher than 10 atmospheres for a 1-cm gap, the breakdown voltage does not increase with increasing pd value at the same rate as at low pd values, as shown in Figure 8-33. This indicates that the variation of V_b with pd does not obey Paschen's law. For $pd > 5 \text{ atm-cm}$, the breakdown voltage

has no precisely defined value. Usually it has a spread of 5–10% in measured values and a marked dependence of the breakdown voltage on the cathode material.⁸⁵ At higher pd values, the breakdown voltage is of the order of 10^5 V (i.e., for a 1-cm gap, the breakdown field is about 100 kV cm^{-1}). For $pd > 10 \text{ atm-cm}$, the breakdown field may reach a value higher than 500 kV cm^{-1} for a 1 cm gap. For the same value of pd of 10 atm-cm , if the gap is reduced to 0.1 mm (10^{-2} cm), the p would be 1000 atmospheres. It can be imagined that at such a high pressure, the gas may be in liquid or solid phase, and the breakdown strength may reach several MV cm^{-1} . This brings up the question: What are the likely mechanisms leading to self-sustaining discharges and final destructive breakdown?

For electrical breakdown to occur, two criteria must be satisfied:

1. There must be suitable initiatory electrons.
2. There must be a mechanism for producing ionization to amplify the multiplication of electrons and ions, offsetting the loss by diffusion and drift in the interelectrode space.

At pressures around 1000 atmospheres, the gases would become condensed liquids or solids, and the mean free path would be reduced to about 5–20 Å. Under these conditions, electrons cannot get enough energy to cause ionization, even at an applied field of 10^7 V cm^{-1} . At fields higher than 100 kV cm^{-1} , electron emission will occur at the cathode, so the first criterion can always be satisfied. To satisfy the second criterion, it is necessary to find a mechanism such that the mean free path can be enlarged to the level at which impact ionization by electrons can take place at a field of the order of 10^6 V cm^{-1} . Let us first discuss such a mechanism for electrical breakdown in liquids.

8.3.2 Electrical Breakdown in Liquids

Since the mean free path of an electron in a dielectric liquid is of the order of 10 Å, an electron cannot gain sufficient energy from the field to ionize a molecule even at a field of

10^7 V cm^{-1} . It is also unlikely that molecules of dielectric liquids, such as hydrocarbons, can be dissociated under a high field of the same order. The presence of impurities in the liquid definitely affects electronic processes, but for pure liquids, impurities alone should not be the major cause of all high-field phenomena.^{86,87}

However, at fields higher than 10^5 V cm^{-1} , electron emission from the cathode will occur. It is most likely that electrical conduction due to injected carriers from the cathode is filamentary, because the electrode surface is not microscopically identical in asperity and surface condition from domain to domain (see Filamentary Charge-Carrier Injection in Solids in Chapter 7). Thus, there may be one or more microregions in which the potential barrier has a profile more favorable for carrier injection than in other regions. Furthermore, the liquid itself is never microscopically homogeneous, particularly under high fields after the onset of electrohydrodynamic motion.⁸⁸ Because of these unavoidable nonuniformities, for a given applied voltage across two parallel plane electrodes (with the edge effect ignored), the field F is nonuniform longitudinally due to the space charge effect, and the current density J is non-uniform radially due to filamentary current flow. The filamentary conduction is evidenced by the fact that the electrical breakdown channel is always filamentary, whether it is in a gas, a liquid, or a solid. Also, a breakdown usually produces only a tiny spot of damage on each electrode surface.

Electron emission from the cathode satisfies one of the criteria for the occurrence of breakdown. Now we must find a mechanism to satisfy the second criterion, that is, amplifying carrier multiplication via impact ionization of molecules by electrons. The most significant discovery for this mechanism due to Kao⁸⁹⁻⁹² is that the breakdown strengths of simple hydrocarbon liquids and transformer oil are dependent on applied hydrostatic pressure under all conditions, including extreme cleanliness and applied voltage pulses with the pulse width of only $1 \mu\text{s}$. Breakdown strength is also dependent on temperature. Typical results for n-hexane are shown in Figure 8-34. It can be seen that

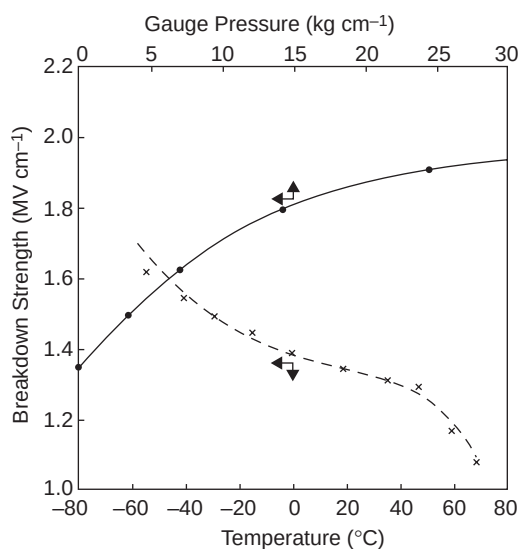


Figure 8-34 Breakdown strength of n-hexane as a function of hydrostatic pressure and temperature for a rectangular pulse duration of $4.5 \mu\text{s}$ and a gap length of 0.02 cm .

breakdown strength increases with increasing pressure and decreases with increasing temperature, indicating clearly that electron ionization depends on both pressure and temperature. This experimental finding has led to the development of bubble theory for electrical breakdown in liquids.

The quantity of electrons injected from the asperity on the cathode surface increases with increasing applied field. When it reaches a certain level, these electrons will repel each other and produce transiently large cavities in which the mean free path may suddenly or transiently increase to values at which electron impact ionization can take place. This process will be enhanced by the heat generated due to the mutual repulsion of electrons and the recombination of electrons and ions resulting from the transient ionization, leading to the formation of low-density bubbles. The formation of low-density bubbles has been observed experimentally by several investigators,^{48-51,93-95} and light emission at the cathode region by Kao et al.^{54,55} The formation of bubbles can be considered important experimental evidence that electrical breakdown in

condensed dielectric materials requires the creation of low-density domains or regions in the bulk by carrier injection from the electrical contacts as a necessary prelude, based on Kao's model.⁶ Obviously, the formation of bubbles will lead to the amplification of carrier multiplication and to the final self-sustaining breakdown.

8.3.3 Electrical Breakdown in Solids

In this section we shall describe several breakdown processes in solids.

Thermal Breakdown

Electrical breakdown in all materials, whether in gas, liquid, or solid phase, is ultimately due to thermal instability, leading to the destruction of the material. Thermal breakdown generally refers to the breakdown caused by joule heating continuously generated within the dielectric specimen, due mainly to electrical conduction and polarization, which cannot be extracted fast enough by thermal conduction or convection. The general equation governing the balance of the heat generation rate and the heat loss rate is given by

$$C_v \frac{dT}{dt} - \text{div}(K_t \text{grad} T) = \sigma F^2 \quad (8-56)$$

where C_v is the specific heat per unit volume, K_t is the thermal conductivity, and σ is the electrical conductivity. In the case of DC fields, σ is the DC electrical conductivity; in the case of AC fields, σ should include the conductivity due to dielectric polarization loss ($\omega \epsilon_0 \epsilon_r'$) in addition to the normal conductivity, in which ω is the frequency and ϵ_r' is the imaginary component of the complex relative permittivity of the material.

The first term on the left of Equation 8-56 represents the heat absorbed by the material, and the second term represents the heat lost to the surroundings. Qualitatively, if heat loss by cooling is linearly related to the temperature rise above the ambient temperature T_0 and the heat generated increases exponentially with temperature, then the heat generation rate and

the heat loss rate can be depicted schematically, as shown in Figure 8-35. With applied field F_1 , the rate of heat generation always exceeds the rate of heat loss, and thermal instability will occur at any temperature. But with applied field F_2 , equilibrium will be attained, provided that the temperature of the specimen does not exceed T_B . Equation 8-56 cannot be solved analytically for the general case since C_v , K_t , and σ all may be functions of temperature, and σ may be dependent on the applied field. Here, we will consider only two simple and extreme cases for the solution of Equation 8-56.

Case 1: Impulse Thermal Breakdown—In this case, we assume that the buildup of the electrical field is so rapid that the heat lost to the surroundings may be neglected, so Equation 8-56 may be reduced to

$$C_v \frac{dT}{dt} = \sigma F^2 \quad (8-57)$$

Supposing that we use a linear ramp field with the ramp rate $r_g = dF/dt$, then F can be written as

$$F(t) = r_g t \quad (8-58)$$

Assuming that σ is a function of temperature following the relation

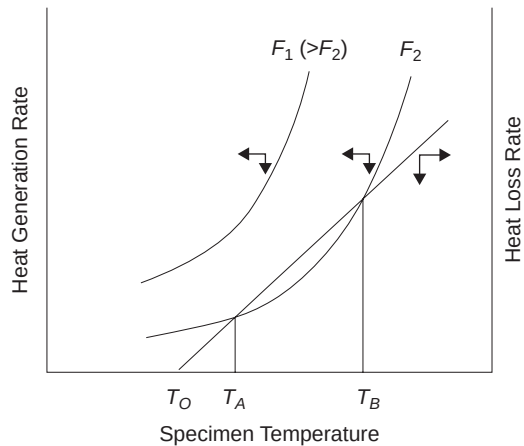


Figure 8-35 Thermal instability occurring in a dielectric specimen for $T > T_B$, the ambient temperature being T_0 .

$$\sigma = \sigma_o \exp(-E_\sigma/kT) \quad (8-59)$$

where σ_o is the preexponential factor and E_σ is the activation energy, substitution of Equations 8-58 and 8-59 into Equation 8-57 yields

$$C_v \frac{dT}{dt} = \sigma_o (r_g t)^2 \exp(-E_\sigma kT) \quad (8-60)$$

Separation of the variables and integration for T from T_o to T and for F from 0 to F give⁹⁶

$$\begin{aligned} (k/E_\sigma)[T_o^2 \exp(E_\sigma/kT_o) - T^2 \exp(E_\sigma/kT)] \\ \simeq \sigma_o F^2 / 3C_v r_g \end{aligned} \quad (8-61)$$

This equation is valid for most cases, provided that C_v can be assumed to be independent of temperature and σ independent of electric field, and $E_\sigma \gg kT$. From Equation 8-61, we can find T as a function of applied field F . Almost all the impulse time is used in raising the temperature of the specimen, and the temperature rise occurs very rapidly within a very short time. So, as soon as the specimen's temperature exceeds T and the applied field reaches a value at the time $t = F/r_g$, then the specimen will undergo thermal breakdown. Several investigators have reported that electrical breakdown in NaCl and KCl under the impulse condition is thermal, following the theoretical prediction of Equation 8-61.⁹⁶⁻⁹⁸

Case 2: Maximum Thermal Voltage—In this case, we assume that there is a finite maximum thermal voltage for a very thick specimen and that this voltage depends only on the physical properties of the dielectric material and the initial temperature. For simplicity, both parallel plane electrodes are assumed to be so large that they can pass all escaping heat, and the hottest point or the highest temperature is at the center of the specimen. We also assume that the temperature at the electrode surfaces is T_1 , the highest temperature at the center is T_m , and the temperature of the ambient medium is T_o . All heat generated in the specimen is assumed to be carried away to its surroundings, so the term $C_v (dT/dt)$ may be ignored. By neglecting the term $C_v (dT/dt)$, Equation 8-56, for the one-dimensional case, can be reduced to

$$-\frac{d}{dz} \left(K_t \frac{dT}{dz} \right) = \sigma \left(\frac{dV}{dz} \right)^2 \quad (8-62)$$

Substituting Equation 8-59 for σ into Equation 8-62 and integrating from the center of the specimen to the electrode surface with the specimen thickness being d , we obtain

$$V^2 = 8 \int_{T_1}^{T_m} \left[\frac{K_t}{\sigma_o \exp(-E_\sigma/kT)} \right] dT \quad (8-63)$$

When K_t is constant and $T_1 = T_o$ (most effective edge cooling), the critical thermal breakdown voltage V_c at which the temperature at the center reaches T_m , corresponding to the temperature at which the specimen material starts to decompose, is

$$V_c \simeq \left(\frac{8K_t k T_o^2}{\sigma_o E_\sigma} \right)^{1/2} \exp(E_\sigma/2kT) \quad (8-64)$$

For thick specimens, thermal breakdown voltage is independent of specimen thickness, but for thin specimens, thermal breakdown voltage is proportional to the square root of the specimen thickness. Furthermore, the dielectric losses that generate heat in the specimen are much greater under AC fields than under DC fields. Consequently, thermal breakdown strength is generally lower for AC fields, and it decreases with increasing frequency of the AC field.³⁷

Thermal breakdown is a well established mechanism. This is why, for the applications of dielectric materials under AC fields, it is important to keep the $\tan \delta$ loss as low as possible. This loss is an essential parameter in determining the thermal breakdown voltage.

Electrical Breakdown

Before presenting the relatively new theory about the mechanisms for the creation of low-density domains that enable electrons to carry out impact ionization in solids, it is worth reviewing briefly some early theories of electrical breakdown in solids.

The basic difference between thermal breakdown and electrical breakdown is that for the former, carrier multiplication is due mainly to mutual feedback between the temperature rise

caused by joule heating and thermal excitation, while for the latter it is due to electronic processes other than thermal excitation. Even for purely electrical breakdown, there are several early theories, and these are briefly described in the following sections.

Intrinsic Breakdown

So-called *intrinsic breakdown* means that the intrinsic breakdown strength is an intrinsic property of the dielectric material, depending only on temperature and independent of specimen size and electrode material. As intrinsic breakdown takes place in a very short time (10^{-8} – 10^{-7} second), it does not depend on the waveform and the duration of the applied electrical voltage pulse, so the breakdown is purely electronic in nature.

There are always electron traps in the material due to the presence of impurities, structural defects, or dislocations. These traps have a ground state and several excited states located below the bottom of the conduction band, as shown in Figure 8-36. Free electrons in the conduction band at an applied field F will gain energy at a rate P_e , which can be expressed as

$$P_e = \sigma F^2 = \frac{dE}{dt} \quad (8-65)$$

where σ is the electrical conductivity and E is the energy gained by electrons. This energy will be consumed in electron–electron scattering in the conduction band, electron–trapped electron scattering, and electron–phonon scattering by virtue of the lattice vibration with thermal energy. For high-purity crystals with low free

electron concentration, the first two types of scatterings are negligible; the main consuming mechanism is electron–phonon scattering.

Taking into account only electron–phonon scattering, P_e is a function of F , electron energy E , and lattice temperature T , while the energy loss P_r is a function of both E and T . Thus, we can write the equation of energy balance as

$$\begin{aligned} P_e &= A(F, E, T) \\ P_r &= B(E, T) \\ A(F, E, T) &= B(E, T) \end{aligned} \quad (8-66)$$

In this case, thermal conductivity is not present in the material since this breakdown phenomenon occurs very rapidly. So the loss by thermal conductance to the surroundings is negligible. An instability appears as soon as $P_e > P_r$, and the material will undergo thermal breakdown.

High-temperature (or High-energy) Criterion—

This criterion is due mainly to Frohlich.^{96,99} The electron temperature T_e is always higher than the lattice temperature T . The energy accumulated by electrons in the field is not simultaneously transferred to the lattice. It takes some time after the application of a steady field for a thermodynamic equilibrium to be reached. In crystals with traps, scattering of electrons with trapped electrons may occur, which would release the trapped electrons and hence, increase the number of electrons in the conduction band. This in turn increases electrical conductivity, implying that the electron temperature will be increased, leading to high-temperature breakdown.

Figure 8-37 shows that at electron temperature T_e , corresponding to electron energy E_1 for applied field F_1 , breakdown occurs. When breakdown occurs at electron temperature T_e larger than the critical temperature T_c , the breakdown is considered high-temperature breakdown. The critical temperature T_c corresponding to the critical energy E_c is defined, according to Frohlich⁹⁹ and Whitehead,³⁷ as that at E_c ; there is a critical field F_c at which the rate of energy gain of the electrons equals the rate of energy loss to the lattice at $E = E_c$, that is, $A(F_c, E_c T) = B(E_c, T)$. F_c is given by

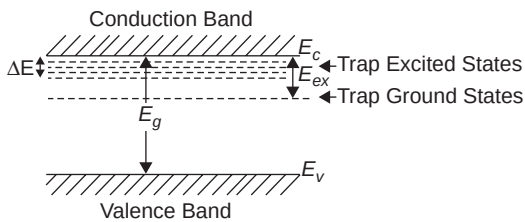


Figure 8-36 Schematic diagram illustrating the energy band structure of a dielectric material and the excited states of traps.

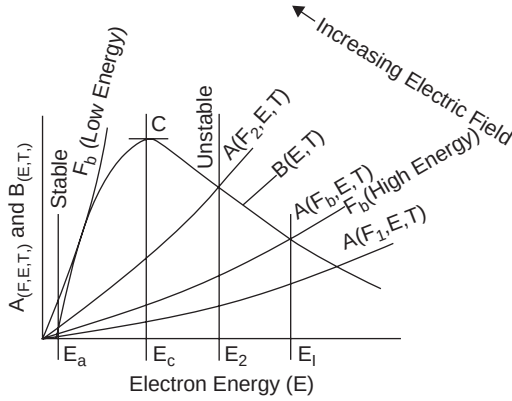


Figure 8-37 The average rate of energy gain of electrons $A(F, E, T)$ from the applied field F and the average rate of energy loss to the lattice $B(F, E, T)$ as functions of electron energy E . E_i is the ionization energy approximately equal to the energy band gap; F_b is the breakdown strength.

$$F_c = C \exp\left(\frac{\Delta E}{2kT}\right) \quad (8-67)$$

where C is a constant, and ΔE is the mean energy gap between the excited states of traps and the bottom of the conduction band. The electron temperature at the critical field F_c is T_c , which can be evaluated from³⁷

$$\frac{1}{T} - \frac{1}{T_c} = \frac{k}{\Delta E} \quad (8-68)$$

From Figure 8-37 it can be seen that instability occurs for $E > E_i$ corresponding to the applied field F_1 at which the electrical conductivity will increase rapidly, leading to final breakdown. In the high-temperature regime, the breakdown strength is expected to decrease with increasing temperature, as shown in Figure 8-38 for NaCl and NaCl + AgCl crystals.

Low-temperature (or low-energy) criterion— This criterion is due mainly to von Hippel.^{100–102} In pure crystals, electron–phonon interaction prevails and the energy transfer rate is proportional to the reciprocal of the relaxation time $\tau(E)$ for electron energy $E < E_c$. $\tau(E)$ is proportional to $E^{1/2}$. Thus, P_r can be written as

$$P_r = C_r E^{1/2} \quad (8-69)$$

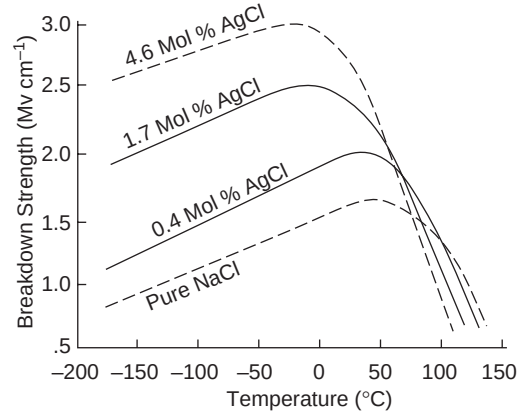


Figure 8-38 Variation of breakdown strength with temperature for NaCl pure crystals and the mixture of NaCl and AgCl, showing the effect of impurity (AgCl) in NaCl crystals.

where C_r is a constant. For a high relaxation time, the rate of energy loss is small and the rate of energy gain for electrons increases. For conduction electrons with mean energy E , each electron will gain energy from the field within an average time and then lose it in exciting lattice vibrations. If the average time between collision is τ , the average velocity of an electron at applied field F will be $qF\tau/m^*$, which constitutes an electron current $q^2F\tau/m^*$. Thus, the rate of the energy gain can be written as

$$P_e = \frac{q^2 F^2 \tau}{m^*} \quad (8-70)$$

where m^* is the effective mass of the electron.

For $P_e > P_r$, the electrons will gain energy continuously from the field, tending to reach a state in which the high-energy electrons may be able to ionize the atoms by direct impact. Once impact ionization occurs, more free electrons are generated, causing an explosive increase in current by producing an avalanche.

The low-temperature criterion is that breakdown will not occur until the critical energy E_c is less than or equal to ionization energy E_i . The minimum field F_c at which breakdown will occur is given by the condition $E_c = E_i$. Making $P_e = P_r$, from Equations 8-69 and 8-70, we obtain

$$F_c = \frac{1}{q} \left(\frac{C_r m^*}{\tau E_i^{1/2}} \right)^{1/2} \quad (8-71)$$

F_c depends on T through τ , which is proportional to $T^{-3/2}$. In this case, the field causes thermal instability at electron energy E larger than E_a , corresponding to the applied field F_a , as shown in Figure 8-37. The temperature dependence of τ for electron-phonon scattering leads to the conclusion that breakdown strength should increase with increasing temperature, but not very rapidly. Typical results for NaCl and NaCl with additive AgCl are shown in Figure 8-38. Since the critical field is dependent on τ , the increase in the concentration of impurities (or other defects) in the crystal would obviously increase the probability of scattering. This may be why F_c increases and shifts toward a lower temperature as the concentration of impurities is increased.

However, there has been no direct experimental evidence to show whether the concept of intrinsic breakdown does exist in solids. Some facts cannot easily be accommodated by the theories described above. Some of them are listed here:

- Prebreakdown light emission from the cathode has been observed in alkali halide crystals and other solids.

- Breakdown strength of alkali halide crystals and other solids is dependent on specimen thickness.
- Carrier injection from electrical contacts has been confirmed to be one of the important processes for high-field conduction and breakdown.
- The effects of space charges are also very important in affecting the breakdown strength.

Zener Breakdown due to Band-Band Transition

The first theory of intrinsic breakdown was put forward by Zener.¹⁰³ The mechanism can be described simply as free electrons generated by tunneling from the occupied valence band to the conduction band in the presence of a strong field, as shown in Figure 8-39. When an electron undergoes a Bragg reflection at a Brillouin zone boundary, there is a small, finite probability of making a transition by tunneling instead of reflection to the next zone, and so to a band of higher energy (conduction band). At high field F , the electron will be accelerated through the band and its wave vector k will be increased at a steady rate until $k = \pi/a$, where a is the lattice constant. Usually, the wave vector will revert to $-\pi/a$, corresponding to a Bragg reflection.

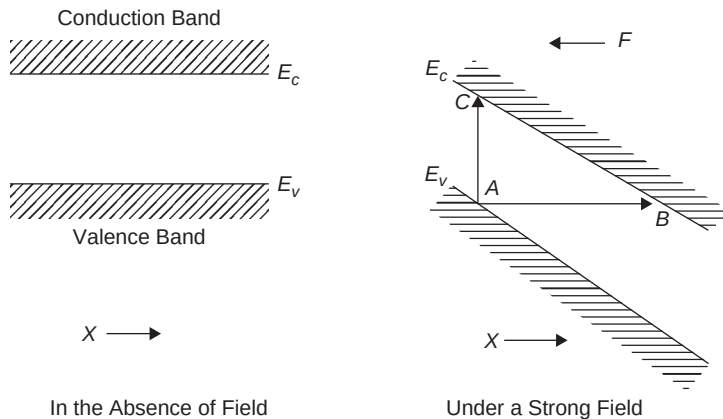


Figure 8-39 Zener tunneling from position A at $k = \pi/a$ in the valence band to position B of equal energy and wave vector in the conduction band.

tion at the zone boundary. However, depending on the tunneling distance (see AB in Figure 8-39), the wave vector may retain the value π/a and make a transition to a higher energy band.¹⁰⁴ Calculating the probability for such a transition even at a steady field is quite formidable mathematically, and it is beyond the scope of this book.

Zener originally intended to use this mechanism to explain breakdown phenomena in dielectric solids, but in most cases, the breakdown is due to other causes. It is most likely that this mechanism may prevail in narrow-bandgap semiconductors, such as Zener p-n junction diodes. In dielectric materials with a large band gap, the Zener effect may be neglected.

Avalanche Breakdown

Seitz¹⁰⁵ has proposed that breakdown in solids is similar to that in gases, due mainly to electron avalanches. There is a major question regarding this mechanism: How can electrons gain sufficient energy from the electric field for impact ionization with such a small mean free path in solids?

O'Dwyer¹⁰⁶ has pointed out that a theory that considers impact ionization an important process and at the same time retains the assumption of a uniform field is not realistic when there are many ionizing collisions. By assuming that the average mobility of one type of carrier (i.e., electrons) is much higher than that of the other type (i.e., holes) and that electron injection from the cathode plays a decisive role in the breakdown processes, then it is also assumed that the positive hole space charges will enhance the field toward the cathode and hence electron injection.

Impact ionization and electron injection tend to promote each other, resulting in a large distortion of the field. The field near the cathode may reach many times higher than the average field. This mutual feedback process leads to final breakdown. This model may explain the thickness dependence of breakdown strength, but it fails to explain many other breakdown phenomena. Several investigators have

attempted to further modify avalanche breakdown theory by including the effects of electron and hole traps, but the modifications do not improve the theory's weakness.^{107,108}

Electromechanical Breakdown

In Electromechanical Effects in Chapter 2, we mentioned that the attractive force between two parallel metallic electrodes with opposite electric charges in contact with a solid dielectric specimen of permittivity ϵ tends to compress the material, reducing its dimension in the direction of the applied field. The compression may reach several kNm^{-2} under a field of 10^6 V cm^{-1} . This compression force is restrained by the elastic force, but the stress-strain relationship is usually not simple for polymers. Stark and Garton¹⁰⁹ assumed a logarithmic relation for polymers. Thus, in equilibrium, the change in specimen thickness at the applied field F can be estimated by the following equation

$$\frac{1}{2} \epsilon_r \epsilon_o F^2 = Y \ell n \frac{d_o}{d} \quad (8-72)$$

or

$$\frac{1}{2} \epsilon_r \epsilon_o \left(\frac{V}{d} \right)^2 = Y \ell n \frac{d_o}{d}$$

where Y is Young's modulus, V is the applied voltage, d_o and d , are respectively, the specimen thickness before and after the application of the field. By differentiating Equation 8-72 with respect to d , we find that V has a maximum when $d/d_o = \exp(-1/2) = 0.6$. When the applied voltage is higher than this value, so that $d/d_o < 0.6$, the specimen becomes unstable and will undergo collapse. Thus, the highest apparent breakdown strength is given by

$$F_b = \frac{V}{d_o} = 0.6 \left(\frac{Y}{\epsilon_r \epsilon_o} \right)^{1/2} \quad (8-73)$$

It has been reported that electromechanical breakdown occurs in polyethylene and other polymers in the high-temperature region.¹¹⁰ However, electromechanical breakdown can be prevented by encapsulation.¹¹¹⁻¹¹³

Electrical Breakdown due to Hot Carriers in High-Mobility States

Because of the disorder in polymers, a mobility edge is expected to exist in both the conduction band and the valence band. According to the electronic processes in noncrystalline materials,¹¹⁴ high carrier band mobilities can only be expected at fields higher than a critical field F_c in order to keep the charge carriers hot above the mobility edges. Thus, at fields higher than F_c most electrons (or holes) may be considered hot electrons (or holes), moving in extended states with a relatively high mobility.

Zeller et al.¹¹⁵ have estimated that the chemical structure of saturated polymers such as polyethylene may provide band mobilities of the order of $10 \text{ cm}^2/\text{V}\cdot\text{s}$. If the applied field is sufficiently high, electrons injected from the cathode can become hot and may subsequently collide with molecules of the material, causing either impact ionization or chemical damage to the material. Zeller et al. have also considered that the mechanical stress created at high field at the tip of the point-plane electrode configuration may be responsible for the initiation and the growth of electrical trees. At the voltage for the onset of tree initiation, the mechanical stress developed at the tip may be larger than the bonding strength of the material. However, structural degradation in materials starts at a much lower electric field but based on a different mechanism, which was discussed in Section 8.1 and will be discussed further later in this section.

Electrical Breakdown due to the Formation of the Gas Phase in Solids

Budenstein^{71,116} has proposed that electrical breakdown consists of four stages: a formative stage, a tree initiation stage, a tree growth stage, and a return streamer stage. During the formative stage, the energy from the applied field will be stored in the dielectric specimen by local rearrangement of the charge distribution in the solid specimen via polarization, impact ionization, trapping, and atomic displacement. The result is alteration of the charge balance locally so that molecular bonds are broken. The transition from the solid phase to the gas phase is

hypothesized to occur at this stage if local charge density increases to the point where nonbonding orbitals are formed.

The formation of the gas phase will lead to the tree initiation stage. Tree growth is due to the energy transferred from the field to the material in the gas phase in the form of electrical discharges, which tend to erode the material in a manner similar to the formative stage. The return streamer stage occurs when the tree extends from one electrode to the other, resulting in a highly conducting streamer and leading to the formation of the final breakdown channel. This model is very similar to gaseous discharge processes, but the mechanism leading to the formation of the gas phase from the solid phase is not very convincing.

Of the six early theories, each has its shortcomings. To develop an electrical breakdown model, it is important to take into account of all observed prebreakdown and breakdown phenomena, and the model must be consistent with all available experimental facts. In the following, we shall present a more compromising model.

A solid dielectric material has about 10^{22} – $10^{23} \text{ atoms cm}^{-3}$, which is more than 10^3 times larger than its gas counterpart, which has about 10^{19} atoms or molecules per cm^3 at normal atmospheric pressure and room temperature. If the electron multiplication process is due to impact ionization of the molecules by electrons in the materials, in the same manner as in gases, the breakdown strength of solids would be about 10^8 V cm^{-1} . But in practice, the breakdown strength of most solid dielectric materials is lower than 10^7 V cm^{-1} , indicating that the electron multiplication process in solid materials is quite different from that in gases.

In general, solid dielectric materials have the following features:

- The energy band gap is large ($E_g > 4 \text{ eV}$).
- Carrier effective mass is large and carrier mobility low ($\mu < 10^{-1} \text{ cm}^2/\text{V}\cdot\text{s}$).
- Carrier mean free path is small ($\ell \approx 5 - 20 \text{ \AA}$).
- The concentration of localized gap states is much larger than that of thermal equilibrium carriers ($N_t \gg n$).

- Dielectric relaxation time is much larger than carrier lifetime ($\tau_d \gg \tau$)
- Electron–phonon scattering cannot be treated as a perturbation but should be considered a major interaction. Electrical transport is thus more likely by phonon-assisted hopping.

With these features, it is almost impossible for carriers to gain energy from the applied field to become sufficiently hot to cause structural change of the material or to gain energy comparable to the energy band gap to cause impact ionization in the material, even at a field of 10^7 V cm^{-1} . Zener-type internal tunneling emission is not possible for large energy bandgap materials.

Before presenting the possible mechanism enabling electron impact ionization in solid dielectric materials, it is worth summarizing some significant experimental observations about prebreakdown and breakdown phenomena, because these led to the development of this mechanism.

- The threshold field for the occurrence of partial discharges (electrical treeing) and the breakdown strength of condensed materials, such as hydrocarbon liquids and polymers, are dependent on the applied hydrostatic pressure.^{90,117,118}
- Electrical discharges and breakdown always occur in filaments, as in hydrocarbon liquids,⁹² in NaCl and polyethylene,¹¹⁹ and in SiO_2 films.^{120,121}
- Breakdown strength is strongly dependent on specimen thickness, as for NaCl and polyethylene¹¹⁹ and for SiO_2 films.^{120–122}
- Partial discharge and breakdown are always preceded by light emission.^{56–60}
- The threshold voltage for the initiation of electrical treeing and the breakdown voltage in polymers under a DC stressing field is much higher than under an AC stressing field in the point–plane electrode configuration. Tree penetration length is much longer and tree channel size much smaller for the former than for the latter.^{69,73}
- Prebreakdown disturbance has been observed in high-viscosity epoxy fluids⁵¹ and in

polyethylene⁵² using a Schlieren technique. Such disturbance at the point of a point–plane electrode configuration occurs intermittently, which is consistent with the intermittent current pulses observed. The refractive index of the disturbance region is smaller than that of the surrounding medium, indicating that the disturbance region may be the low-density region or domain.

These typical and significant prebreakdown and breakdown phenomena occurring in condensed dielectric materials are similar to those that occur in insulating gases. This similarity suggests that there may be a unified theory for electrical breakdown in all dielectric materials, irrespective of whether they are in solid, liquid, or gas phase. For the occurrence of electrical breakdown, two criteria must be satisfied: the primary electrons required for initiating the breakdown processes, and a mechanism to initiate impact ionization required for producing carrier multiplication. As the breakdown strength in dielectric solids is generally higher than 10^4 V cm^{-1} , electron injection from electrical contacts has already started, so the first criterion for breakdown is readily satisfied. Therefore, impact ionization is the key process that leads effectively to carrier multiplication.

If the injected electrons (or holes) can be excited under a high field to the extended state above the mobility edge, then these electrons may become hot enough to cause impact ionization and structural damage due to their bombardment of the molecules.¹¹⁵ But most dielectric materials consist of a large quantity of various traps and the mean free path of electrons is usually very small, so it is unlikely that injected electrons have a chance to be excited to a high mobility state before being trapped. However, to produce impact ionization, it is necessary to create low-density domains or regions at high fields near the carrier-injecting contacts, similar to the formation of bubbles in dielectric liquids before breakdown.^{89,90} According to the model developed by Kao,⁶ electrical breakdown in solids involves the creation of low-density domains or regions in the bulk by carriers injected from the electrical

contacts and, subsequently, dissociative trapping and recombination as a necessary prelude. Breakdown is initiated by impact ionization within such low-density domains or regions in which the mean free path is large, as in gas phase, then followed by a continuous increase in carrier multiplication and extension of the regions to form high conducting channels, leading to final destructive breakdown of the material inside them.

Electrical aging is a gradual degradation process leading to final destructive breakdown. In Section 8.1.1, we discussed in detail the theory of electrical aging, which is, in fact, also the theory for electrical breakdown. So, we will not repeat this here. In short, structural degradation will produce microvoids. When the degree of this degradation reaches a certain critical level, the number and the size of microvoids increases, creating a chance for low-density domains or regions to be formed. The formation of low-density domains or regions converts locally the solid phase to a gas phase in the material, particularly near the carrier-injecting contacts. Impact ionization will occur because of large mean free paths inside the low-density domains or regions. The lifetime of an electrically stressed dielectric material can be considered the time required for the concentration of the new traps created by structural degradation to reach a certain critical value at a given field F_b (breakdown strength), which leads to final breakdown. The critical concentration of the stress-created traps $N'_{t(\text{crit})}$, given by Equation 8-8, is the criterion for electrical breakdown. The applied field required to cause electrical breakdown (i.e., breakdown strength F_b) decreases with increasing stressing time (i.e., lifetime t), but the concentration of the stress-created traps N' remains practically unchanged, as shown in Figure 8-10.

This model can explain qualitatively most prebreakdown and breakdown phenomena in solids. For example, the dependence of the breakdown strength on the ramp rate of the applied linear ramp voltage and on specimen thickness can be explained on the basis of this model. The effective field at the carrier-injecting contact F' is equal to $F - F_i$, where F is the

applied average field and F_i is the internal field created by trapped space charges. Obviously, the slower the ramp rate is, the higher the internal field, implying that more stress-created new traps are produced at a given field because more carriers are trapped in the specimen near the injecting contact. This is why the slower the ramp rate is, the smaller the breakdown strength, as shown in Figure 8-40. Similarly, the fact that breakdown strength decreases with increasing specimen thickness can also be attributed to the effects of trapped space charges.

8.3.4 Similarity in Breakdown Mechanisms for Gas, Liquid, and Solid Dielectrics

Electrical breakdown strengths of gases, liquids, or solids are dependent on the cathode material or, in more general terms, on the material of the carrier-injecting contacts. This is because the first criterion for breakdown requires a suitable source for primary electrons (or holes) to start the breakdown processes. In liquids and solids, carrier injection from electrical contacts is always present and serves as the major source of primary electrons. In gases at low pressure (i.e., in a so-called vacuum with the electron mean free path larger than the electrode separation or $p < 10^{-4}$ torr), electron injection is present because the field required to cause breakdown is higher than 10^4 V cm^{-1} . In this case, electrons injected from the cathode will bombard the anode surface, knocking out positive ions and producing metallic vapor from the anode surface. When they reach the cathode, these positive ions will cause secondary emission of electrons and produce metallic vapor from the cathode surface. This is why the breakdown strength depends strongly on the electrode material.

In gases at high pressure (>10 atms), the efficiency of impact ionization decreases because of the smaller mean free path. Electron injection becomes very important, because the carrier multiplication process depends on photoionization by photons due to radiative recombination. Even in gases at normal pressure and

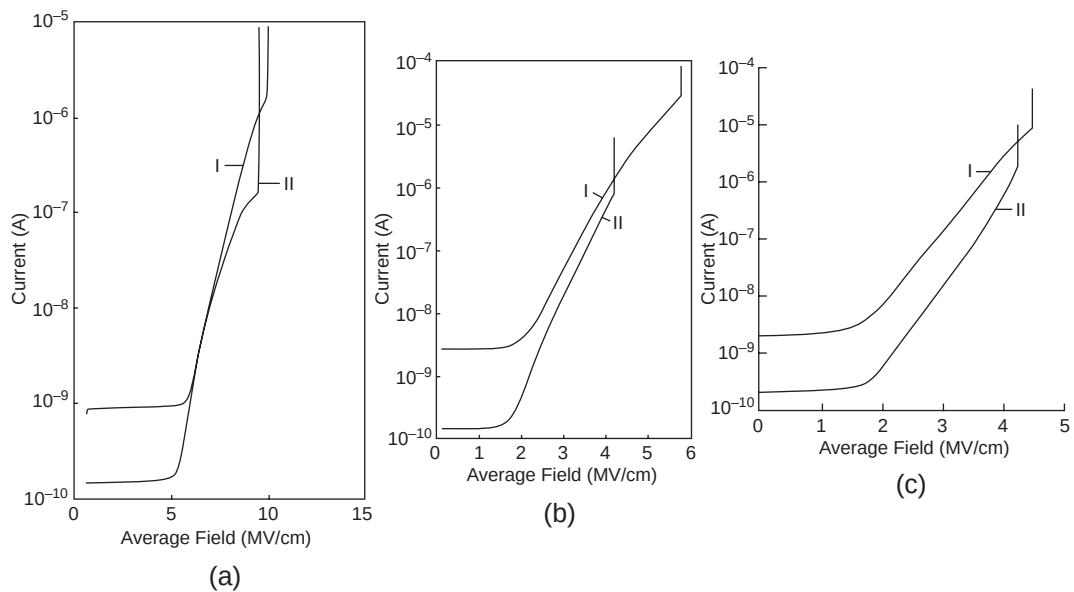


Figure 8-40 The effects of the ramp rate on electrical breakdown strength of (a) SiO₂ films: (I) ramp rate = 0.8 MV/cm/s, (II) ramp rate = 0.1 MV/cm/s; (b) polyethylene films: (I) ramp rate = 0.6 MV/cm/s, (II) ramp rate = 0.06 MV/cm/s; and (c) polyimide films: (I) ramp rate = 0.26 MV/cm/s, (II) ramp rate = 0.026 MV/cm/s.

temperature conditions, electron avalanche requires primary electrons starting from the cathode, which are usually produced by photoemission from the cathode by photons due to cosmic rays.

Regarding the second criterion for breakdown (i.e., impact ionization), it can be imagined that to cause impact ionization, electrons must have a mean free path large enough for them to gain sufficient energy from the applied field. This can be achieved only in low-density regions. In these regions, the density of constituent molecules is much smaller than in solids, implying that the regions must be in gas phase. In other words, to satisfy the second criterion for breakdown, we need a mechanism to create low-density regions or domains. In gases at low pressures ($<10^{-4}$ torr), the metallic vapor produced by the bombardment of charge particles on the electrode surfaces can be considered the low-density regions in which impact ionization can take place. The continuous bombardment by more particles continuously generated in the process can be considered the

feedback process to boost carrier multiplication until breakdown occurs.

In gases at high pressures (>10 atms), impact ionization can take place, although the efficiency may be lower because of smaller mean free paths, but continuous carrier injection from the injecting contacts can serve as a good feedback to enhance carrier multiplication. In gases at normal pressure and temperature conditions, the whole medium is a low-density medium. Efficient impact ionization prevails when the applied field is high enough for the electrons to gain energy equal to or higher than the ionization energy of the molecules. The feedback is the secondary emission of electrons from the cathode by the bombardment of positive ions.

In liquids, the bubbles formed near the cathode before the occurrence of breakdown can be considered the low-density regions or domains in which impact ionization can take place. Continuous carrier injection and radiative recombination serve as a feedback to make the bubbles to grow until final breakdown occurs.

In solids, the low-density regions or domains are created by carriers injected from electrical contacts and, subsequently, dissociative trapping and recombination. The gradual degradation of the material by this process results in electrical aging and final breakdown.

Electrical breakdown in dielectric materials, whether in gas, liquid, solid phase, involves the satisfaction of two major criteria: primary electrons (or holes) and subsequent impact ionization. In addition to these, a feedback process is required to ensure that carrier multiplication can continuously increase. Based on the similarity in breakdown mechanisms among gases, liquids, and solids, there may be a unified theory for electrical breakdown for all dielectric materials of any phase. The development of such a theory will be a big challenge to scientists in the future.

References

1. D. Liu and K. C. Kao, *J. Appl. Phys.*, **69**, 2489 (1991).
2. C. M. Osburn and E. J. Weitzman, *J. Electrochem. Soc.*, **119**, 603 (1972).
3. E. Criseld, *Chemical Bonding and Structure*, (Raytheon Education, San Francisco, 1985).
4. B. Ranby and J. F. Rabek, *Photodegradation, Photexcitation and Photostabilization*, (Wiley, New York, 1975).
5. J. S. Blakemore, *Semiconductor Statistics*, (Pergamon, Oxford, 1962).
6. K. C. Kao, *J. Appl. Phys.*, **55**, 752 (1984).
7. D. J. DiMaria, *Appl. Phys. Lett.*, **51**, 658 (1987).
8. D. Liufu, X. S. Wang, D. M. Tu, and K. C. Kao, *J. Appl. Phys.*, **83**, 2209 (1998).
9. A. Badihi and B. Eitan, *Appl. Phys. Lett.*, **40**, 396 (1982).
10. Y. Nissan-Cohen, J. Shappier, and D. Frohman-Bentchkowsky, *J. Appl. Phys.*, **60**, 2024 (1956).
11. E. Avni and J. Shappier, *Appl. Phys. Lett.*, **51**, 1857 (1987).
12. K. Tihira and K. C. Kao, *J. Phys. D: Appl. Phys.*, **18**, 2247 (1985).
13. J. Lindmayer, *J. Appl. Phys.*, **36**, 196 (1965).
14. R. H. Walden, *J. Appl. Phys.*, **43**, 1178 (1972).
15. H. J. Wintle, *J. Appl. Phys.*, **44**, 3514 (1973).
16. H. J. Wintle, *J. Non. Cryst. Solids*, **15**, 471 (1974).
17. H. J. Wintle, *IEEE Trans. Electr. Insul.*, **EI-12**, 97, 424 (1977).
18. D. M. Tu, X. S. Wang, and K. C. Kao, *Annual Report of 1993 IEEE Conference on Electrical Insulation and Dielectric Phenomena (CRIDP)*, (IEEE Dielectrics and Electrical Insulation Society, New York, 1993), p. 550.
19. R. G. Ranby and H. Yoshida, *J. Polym. Sci., Pt. C-12*, 263 (1966).
20. E. Harari, *J. Appl. Phys.*, **49**, 2478 (1978).
21. G. M. Sessler and R. G. Mulhaupt, *IEEE Trans. Electr. Insul.*, **EI-19**, 190 (1984).
22. R. Zbinden, *Infrared Spectroscopy of High Polymers*, (Academic Press, New York, 1964).
23. L. E. Alexander, *X-ray Diffraction Methods in Polymer Science*, (Wiley, New York, 1969).
24. A. Guinier and G. Fournet, *Small Angle Scattering of X-rays*, (Wiley, New York, 1955).
25. D. Tu, Y. Yin, X. Wang, F. Yong, J. Wang, and Q. Lei, *Proc. Of 6th International Conference on Properties and Applications of Dielectric Materials (ICPADM)* (IEEE Dielectrics and Electrical Insulation Society, New York, 2000), p. 46.
26. D. Liu, D. M. Tu, and K. C. Kao, *Proc. of International Conference on Properties and Applications of Dielectric Materials (ICPADM)* (IEEE Dielectrics and Electrical Insulation Society, New York, 1985), p. 197.
27. D. M. Tu, W. B. Liu, G. P. Zhuang, Z. Y. Liu, and K. C. Kao, *IEEE Trans. Electr. Insul.*, **EI-24**, 587 (1989).
28. D. Liu, L. Kan, and K. C. Kao, *J. Appl. Phys.*, **70**, 919 (1991).
29. W. Jin and D. M. Tu, *IEEE Electrical Insulation Magazine*, **9**, 52 (1993).
30. T. Mizutani, Y. Susuoki, M. Mikita, K. M. Yoo, and M. Ieda, *Annual Report of 1986 Conf. on Electrical Insulation and Dielectric Phenomena* (IEEE Dielectrics and Electrical Insulation Society, New York, 1986), p. 37.
31. M. Ikeda and T. Ohki, *Proc. 2nd Intern. Conf. on Conduction and Breakdown in Solid Dielectrics* (IEEE Dielectrics and Electrical Insulation Society, New York, 1986), p. 71.
32. F. H. Kreuger, *Discharge Detection in High Voltage Equipment*, (American Elsevier, New York, 1965).

33. J. M. Meek and J. D. Craggs, *Electrical Breakdown of Gases*, (Clarendon, Oxford, 1953).
34. H. C. Hall and M. Russek, *Proc. Inst. Electr. Engineers, Pt. II-101*, 47 (1954).
35. J. H. Mason, *Proc. IEE, Pt. I-98*, 44 (1951), and *Pt. IIA-100*, 149 (1953).
36. J. H. Mason, *IEE Monograph 127M, 102C*, 254 (1955).
37. S. Whitehead, *Dielectric Breakdown of Solids*, (Clarendon, Oxford, 1961).
38. J. H. Mason, "Dielectric Breakdown in Solid Insulation," in *Progress in Dielectrics*, vol. 1, edited by J. B. Birks and J. H. Schulman (Heywood, London, 1959), pp. 1-58.
39. R. J. Densley, *IEEE Trans. Electr. Insul., IE-14*, 148 (1979).
40. R. J. Densley, "Partial Discharges under Direct Voltage Conditions" and "Partial Discharges under Impulse Voltage Conditions," in *Engineering Dielectrics*, vol. 1, edited by R. Bartnikas and E. J. McMahon (ASTM Publications, New York, 1979), p. 409 and p. 468.
41. R. Bartnikas and G. L. d'Ombraiu, *IEEE Trans. Power Apparatus and Systems, PAS-84*, 770 (1965).
42. R. Bartnikas, *IEEE Trans. Electr. Insul., EI-6*, 63 (1971).
43. G. M. Sessler, "Physical Principles of Electrets," in *Electrets*, edited by G. M. Sessler (Springer-Verlag, New York, 1987), p. 13.
44. R. Bartnikas, "Corona Discharge Processes in Solids," in *Engineering Dielectrics*, vol. 1, edited by R. Bartnikas and E. J. McMahon, (ASTM Publications, New York, 1979), p. 22.
45. E. C. Rogers, *J. Inst. Electr. Engrs.*, 4, 621 (1958).
46. T. Tanaka and Y. Ikeda, *IEEE Trans. Power Apparatus and Systems, PAS-90*, 2692 (1971).
47. D. W. Holder and R. J. North, "Schlieren Methods," National Physical Laboratory, Department of Scientific and Industrial Research Note on Applied Science 31 (Her Majesty's Stationery Office, London, 1963).
48. S. S. Hakim and J. B. Higham, *Nature*, 189, 996 (1961).
49. F. Heiman, P. Sibillot, and R. Coelho, *J. Phys. D: Appl. Phys.*, 9, 45 (1976).
50. A. H. Sharbaugh, J. C. Devins, and S. J. Rzed, *IEEE Trans. Electr. Insul, EI-13*, 249 (1976).
51. H. Sueda and K. C. Kao, *IEEE Trans. Electr. Insul, EI-17*, 221 (1982).
52. H. K. Xie and K. C. Kao, *IEEE Trans. Electr. Insul., EI-20*, 293 (1985).
53. P. Sibillot and R. Coelho, *J. Phys. Paris*, 35, 141 (1974).
54. C. W. Smith, K. C. Kao, J. H. Calderwood, and J. D. McGee, *Nature*, 210, 192 (1966).
55. K. C. Kao and M. M. Rashwan, *Proc. IEEE*, 62, 856 (1974).
56. R. Cooper and C. T. Elliott, *J. Phys. D: Appl. Phys.*, 17, 481 (1966).
57. D. W. Aucland, R. Cooper, and J. Sanghera, *IEE Proc.*, 128A, 209 (1980).
58. N. Shimizu, H. Kasaki, M. Miyachi, M. Kosaki, and K. Norii, *IEEE Trans. Electr. Insul., EI-14*, 256 (1974).
59. S. S. Bemji, A. T. Bulinski, R. J. Densley, and N. Shimizu, *Annual Report of 1982 Conf. Electr. Insul. and Dielectric Phenomena* (CRIDP) (IEEE Electr. Insul. Soc., New York, 1982), p. 592.
60. S. S. Bemji, A. T. Bulinski, and R. J. Densley, *IEEE Trans. Electr. Insul., EI-24*, 91 (1989).
61. R. M. Eichhorn, "Treeing in Solid Organic Dielectric Materials," in *Engineering Dielectrics*, vol. IIA, edited by R. Bartnikas and R. M. Eichhorn (ASTM Publications, New York, 1983), p. 355.
62. G. Bahder, T. W. Dakin, and J. H. Lawson, *Proc. Conference Internationale des Grands Reseaux Electriques* (CIGRE), Paris, Paper No. 15-05 (1974).
63. J. H. Lawson and W. Vahlstrom, *IEEE Trans. Power Apparatus and Systems, PAS-92*, 824 (1973).
64. E. J. McMahon, *IEEE Trans. Electr. Insul., EI-13*, 277 (1978).
65. F. Noto and N. Yoshimura, *Annual Report of 1974 Conf. Electr. Insul. and Dielectric Phenomena* (IEEE Dielectrics and Electrical Insulation Soc., New York, 1974), p. 207.
66. J. C. Fothergill, L. A. Dissado, and P. J. J. Sweeney, *IEEE Trans. Dielectrics and Electr. Insul., DEI-1*, 474 (1994).
67. R. M. Eichhorn, *IEEE Trans. Electr. Insul., EI-12*, 2 (1976).
68. C. Laurent and C. Mayoux, *IEEE Trans. Electr. Insul., EI-15*, 33 (1980).
69. K. C. Kao and D. M. Tu, *Proc. 1982 IEEE Intern. Symposium on Electr. Insul.* (IEEE Electr. Insul. Soc., New York, 1982), p. 300.
70. R. McLeod, D. Liu, W. Pries, K. C. Kao, and H. C. Card, *Solid State Commun.*, 56, 197 (1985).

71. P. O. Budenstein, IEEE Trans. Electr. Insul., *EI-15*, 225 (1980).
72. C. B. Duke and T. J. Fabish, J. Appl. Phys., *49*, 315 (1978).
73. D. M. Tu, L. H. Wu, X. Z. Wu, C. K. Cheng, and K. C. Kao, IEEE Trans. Electr. Insul., *EI-17*, 539 (1982).
74. E. Friedlander and J. R. Reed, Proc. IEE, *Part IIA-100*, 121 (1953).
75. G. Mole and F. C. Robinson, *Annual Report of 1962 Conf. Electrical Insulation* (National Academy of Science, National Research Council, Washington, D.C. Publication, 1080, 1966), pp. 54–56.
76. E. B. Curdts, "Fundamentals of Partial Discharge Detection: System, Sensitivity and Calibration," in *Engineering Dielectrics*, edited by R. Bartnikas and E. J. McMahon (ASTM Publications, New York, 1979), pp. 68–100.
77. J. S. Townsend, *Electricity in Gases*, (Oxford University Press, Oxford, 1914).
78. A. von Engel, *Ionized Gases*, (Clarendon, Oxford, 1965).
79. L. B. Loeb and J. M. Meek, *The Mechanism of Electric Spark*, (Stanford University Press, Stanford, 1940).
80. H. Raether, *Electron Avalanches and Breakdown in Gases*, (Butterworths, London, 1964).
81. E. Kuffel and W. S. Zaengl, *High Voltage Engineering*, (Pergamon Press, Oxford, 1984).
82. F. Paschen, Weid. Ann., *37*, 69 (1889).
83. R. Hawley and A. A. Zaky, in *Progress in Dielectrics*, edited by J. B. Birks and J. H. Schulman (Haywood, London, 1967) Vol. 7, p. 115.
84. F. Llwellyn-Jone, Vacuum (London), *3*, 116 (1954).
85. F. Llwellyn-Jone, *Ionization and Breakdown in Gases*, (Methuen & Co., London, 1957).
86. T. J. Gallagher, *Simple Dielectric Liquids*, (Clarendon, Oxford, 1975).
87. I. Adamczewski, *Ionization, Conductivity and Breakdown in Dielectric Liquids*, (Taylor and Francis, London, 1969).
88. N. J. Felici, Direct Current, *2*, 147 (1972).
89. K. C. Kao, AIEE Conference Paper no. CP 60–84, Amer. Inst. Electr. Engrs. (1960).
90. K. C. Kao and J. B. Higham, J. Electrochem. Soc., *108*, 522 (1961).
91. K. C. Kao and J. P. C. McMath, IEEE Trans. Electr. Insul., *EI-5*, 64 (1970).
92. K. C. Kao, IEEE Trans Electr. Insul., *EI-11*, 121 (1976).
93. R. Kattan, A. Denat, and O. Lessaint, J. Appl. Phys., *66*, 4062 (1989).
94. R. Kattan, A. Denat, and N. Bonfati, IEEE Trans. Electr. Insul., *EI-26*, 656 (1991).
95. P. K. Watson, M. I. Qureshi, and W. G. Chadband, *Annual Report of 1998 IEEE Conf. on Electrical Insulation and Dielectric Phenomena*, (IEEE Dielectrics and Electr. Insul. Soc., New York, 1998), vol I, p. 35.
96. J. J. O'Dwyer, *The Theory of Electrical Conduction and Breakdown in Solid Dielectrics*, (Clarendon, Oxford, 1973).
97. J. R. Hanscome, Aust. J. Phys., *15*, 504 (1962).
98. J. R. Hanscome, J. Phys. D: Applied Phys., *2*, 1327 (1969).
99. H. Frohlich, Proc. Royal Soc., *A-188*, 521 (1947).
100. A. R. von Hippel, Ergebn. exakt. Naturw., *14*, 79 (1935); also Z. Phys., *88*, 352 (1939).
101. A. R. von Hippel and G. M. Lee, Phys. Rev., *59*, 824 (1941).
102. H. B. Callen, Phys. Rev., *76*, 1394 (1949).
103. C. Zener, Proc. Royal Soc., *A145*, 523 (1934).
104. R. A. Smith, *Wave Mechanics of Crystalline Solids*, (Chapman and Hall, London, 1963), p. 275.
105. F. Seitz, Phys. Rev., *73*, 1375 (1949).
106. J. J. O'Dwyer, J. Appl. Phys., *40*, 3887 (1969).
107. T. H. DiStefano and M. Shatzkes, J. Vac. Sci. Technol., *12*, 37 (1975).
108. N. Klein, J. Appl. Phys., *53*, 5829 (1982).
109. K. H. Stark and C. G. Garton, Nature, *176*, 1225 (1955).
110. R. A. Fava, Proc. IEE, *112*, 819 (1965).
111. J. J. McKeown, Proc. IEE, *112*, 824 (1965).
112. J. Blok and D. G. LeGrand, J. Appl. Phys., *40*, 288 (1969).
113. W. G. Lawson, Proc. IEEE, *113*, 197 (1966).
114. N. F. Mott and E. A. Davis, *Electronic Processes in Non-Crystalline Materials*, (Clarendon, Oxford, 1979).
115. H. R. Zeller, P. Pfluger, and J. Bernasconi, IEEE Trans. Electr. Insul., *EI-19*, 200 (1984).
116. J. Knaur and P. P. Budenstein, IEEE Trans. Electr. Insul., *EI-15*, 313 (1980).
117. K.C. Kao, H.K. Xie, and D.M. Tu, J. Electrostatics, *16*, 115 (1984).
118. D. M. Tu and K. C. Kao, *Annual Report of 1983 Conference on Electrical Insulation and Dielectric Phenomena (CEIDP)* (IEEE

- Electrical Insulation Soc., New York 1983), p. 307.
119. D. B. Watson, W. Heyes, K. C. Kao, and J. H. Calderwood, IEEE Trans. Electr. Insul. *EI-I*, 30 (1965).
120. N. Klein, Thin Solid Films, *50*, 223 (1978).
121. P. Solomon, J. Vac. Sci. Technol., *14*, 1122 (1977).
122. T. T. Chau, S. R. Mejia, and K. C. Kao, J. Vac. Sci. Technol., *B9*, 50 (1991).

Index

A

Absorption, 92
 absorption coefficient, 154–156
Accumulation region, 340
Acousto-optic effects, 143–144
 elasto-optic effect, 143
 photoelastic effect, 143
Activation energy, 381
Activators, 170
Actuators, 267
Airy or diffraction disk, 122
Alloys of PZT type ferroelectric ceramics, 221, 230
Amoore's law, 3
Angular momentum, 11
 orbital angular momentum, 11
 spin angular momentum, 12
Anomalous behavior or anomalies, 216–217
Anomalous dispersion, 156
Anomalous photovoltaic effects, 206
Antiferroelectric transition, 241
Antiferromagnetic materials, 24
Antireflection coating, 120
Argand diagram, 93
Atomic polarization, 59
 ionic polarization, 59
 vibrational polarization, 59
Auger recombination processes, 150
Autoionization, 162, 164, 166
Avogadro's number, 80

B

Band conduction, 389
Barium titanate type ferroelectrics, 221
Barkhausen effect, 22, 226
Basic crystal systems, 214
 cubic structure, 214, 222, 231
 hexagonal structure, 214
 monoclinic structure, 214, 229
 orthorhombic structure, 214, 222, 231
 rhombohedral structure, 214, 222, 231
 tetragonal structure, 214, 222, 231
 triclinic structure, 214

Basic effects of electrets, 213
Bimolecular recombination, 168
Bimorph cantilever, 268
Birefringence, double refractions, 128
 extraordinary rays (E-rays), 128
 ordinary rays (O-rays), 128
Black body, 144–145
Blocking contacts, 336
Boltzmann's statistics, 69
Bound surface charges, 55
Bragg diffraction, 144
Brewster angle, 124, 126
Bulk-limited electrical conduction, 406

C

Capacitance transient spectroscopy, 478
Carrier capture cross section, 425
 electron capture cross section, 425
 hole capture cross section, 425
Carrier capture rate, capture coefficient, 425
Carrier drift mobilities, 397
Carrier lifetime, 432
Carrier thermal velocity, 397
Carrier transit time, 433
Cathodoluminescence, 164
Centrosymmetric group of crystals, 215
Ceric capacitors, 249
 EIA codes, 251
Characteristic luminescence, 170
Characteristic times, 432
Charge carrier injection from electrical contacts, 345
 field emission, 354
 potential barrier lowering, Schottky effect, 345–350
 thermionic emission, 350
 thermionic-field emission, 363
Charge carrier generation, 381
 intrinsic from material itself, 381
 extrinsic from impurities, 381
 injection from electrical contacts, 381
Charge carrier species determination, 472–476

- Charge storage, 297–312
 - for dipolar charges, 297
 - for real charges, 303
- Charge transfer at metal-polymer interface, 376
- Charge-transfer excitons, 160–161
- Classification of crystals, 215
 - centrosymmetric groups, 215
 - noncentrosymmetric groups, 215
- Clausius–Mossotti equation, 80
 - Clausius–Mossotti catastrophe, 80
- Co-activators, 170
- Coercive field, 218
- Cole-Cole plot, 95
- Compensated semiconductors, 396
- Complex conductivity, 100
- Complex permittivity, 86
 - field dependence of complex permittivity, 100
 - temperature dependence of complex permittivity, 97
- Complex refractive index, 94
 - refractive index, 94
 - extinction coefficient or absorption index, 94
- Contact potential, 327, 330
- Contact potential photovoltaic effects, 194
 - MIS solar cells, 196
 - p–n junction photovoltaic behavior, 196
 - p–i–n structure photovoltaic devices, 201
 - Schottky barrier photovoltages, 194
- Corona discharges, 546
 - surface discharges, 546
- Corpuscular theory, 115
- Coulombic traps, 430–432
 - attractive traps, 430
 - neutral traps, 430
 - repulsive traps, 430
- Crystal symmetry point groups, 215
- Curie temperature, 217
 - Curie constant, 217
- Curie–Weiss relation, 217
- Current transient phenomena, 463
 - space charge free (SCF) transient, 466
 - space charge limited (SCL) transient, 468
 - space charge perturbed (SCP) transient, 470
- D**
- Dangling bonds, 342
- Davidson-Cole equation, 97
- Davydov splitting, 159
- Debye equations, 82, 92
- Decay function, 89
- Deep traps, 414
- Defect-controlled conduction, 398
- Delay lines, 267
 - microwave delay lines, 267
- Demarcation levels, 427–429
- Dember effect, 192
 - Dember photovoltage, 192
- Density of quantum states, 393
- Depletion region, 336
- Depolarization current, 58
- Derived units, 35
- Detection of partial discharges, 547
- Diamagnetism, 13
- Dielectric constant, 56, 86
 - relative permittivity, 56
- Dielectric losses, 57
- Dielectric materials, 78
- Dielectric relaxation phenomena, 86, 105
- Dielectric relaxation time, 58, 86, 433
- Dielectric relaxation and chemical structure, 110
- Dielectric saturation, 104
- Diffraction, 121
- Diffusion potential, 332
- Dimensions and units, 34
- Dipolar materials, 78
- Dipole holes, 382
- Direct band gap materials, 152
- Discharge extinction voltage, 529
- Discharge inception voltage, 529
- Dispersion, 92, 127
 - anomalous dispersion, 156
 - normal dispersion, 92, 127
- Dispersive transport, 505
- Distribution of relaxation times, 95, 108
 - distribution function of relaxation times, 108
- Distinction between stored dipolar and real charges, 312
- Domains, 20, 218, 242
 - ferroelectric domains, formation and dynamics, 242, 245
 - ferromagnetic domains, formation and dynamics, 20
- Domain walls, 23, 245
 - for ferroelectric materials, 245
 - for ferromagnetic materials, 23
- Double hysteresis loop, 219

- Drift mobilities of carriers, 397
 - time of flight measurements, 466–468
- Dynamic polarization, 92
- E**
 - Effective carrier mobility, 397
 - Effective mass, 393
 - Effects of ionic conduction, 385
 - Einstein's coefficients, 148
 - Einstein's relation, 401
 - Elastic and inelastic tunneling, 367
 - Electrets, 283
 - charges, electric fields and currents in electrets, 290–294
 - Electrets, applications of, 321
 - electret microphones, 321
 - electromechanical transducers, 322
 - pyroelectric detectors, 322
 - other applications, 323
 - Electrets, materials for, 316
 - Electric charge carriers and their motion, 41
 - Electric field dependence of Curie temperature, 241
 - Electric polarization and relaxation, 52, 56
 - under static electric fields, 52–54
 - under time-varying electric fields, 86
 - Electric polarization mechanisms, 59
 - atomic or ionic polarization, 65
 - electronic polarization, 59
 - orientational polarization, 66
 - space charge polarization, 75
 - hopping polarization, 75
 - interfacial polarization, 77
 - spontaneous polarization, 74
 - Electric susceptibility, 56
 - Electrical aging theory, 515–520
 - Electrical breakdown in gases, 549
 - discharges in electronegative gases, 555
 - streamer breakdown processes, 553
 - Townsend breakdown processes, 551
 - Electrical breakdown in liquids, 557
 - Electrical breakdown in solids, 559
 - electrical breakdown due to
 - avalanche breakdown, 564
 - electromechanical breakdown, 564
 - formation of gas phase in solids, 565
 - hot carriers in high mobility states, 565
 - intrinsic breakdown, 561
 - Zener breakdown, 563
 - thermal breakdown, 559
 - Electrical conduction, 381
 - extrinsic conductivity, 381
 - injection controlled conductivity, 381
 - intrinsic conductivity, 381
 - Electrical contacts, 334
 - blocking contacts, 336
 - neutral contacts, 334
 - ohmic contacts, 336
 - surface states, 341
 - Electrical contacts and potential barriers, 327
 - contact potential, 330
 - electron affinity, 331
 - vacuum level, 328
 - work function, 328
 - Electrical discharges, 528
 - discharge current impulses and recurrence, 528
 - discharge magnitude, 530
 - discharge power losses and energy, 530
 - internal discharges, 528
 - Electrical transport, 389
 - band conduction, 389
 - defect-controlled conduction, 398
 - hopping transport, 401
 - polaron conduction, 402
 - tunneling transport, 399
 - Electrical treeing, 540
 - mechanisms and characteristics, 542
 - Electroluminescence, 171
 - classical electroluminescence, 171
 - (intrinsic electroluminescence)
 - injection electroluminescence, 173
 - (extrinsic electroluminescence)
 - luminescence in anthracene crystals, 174–178
 - Electromagnetic waves and fields, 32
 - Electromechanical effects, 44–51
 - dielectrophoretic force, 49
 - electrostriction force, 49
 - force acting on boundary, 47
 - force acting on conductor surfaces, 47
 - force causing elongation of bubbles or globules, 48
 - torque orienting a solid body, 50
 - Electron affinity, 331
 - Electron doping, 495, 503
 - Electron paramagnetic resonance, 28
 - Electron spin resonance, 30
 - Electron traps, 407

- Electron velocity-field dependent relation, 451–453
- Electronic conduction, 387
- Electronic polarization, 59
 - classical approach, 59
 - electronic polarizability, 60–63
 - quantum mechanical approach, 64
- Electro-optic effects, 131
- Electro-optic processes, 115
- Electrostatic induction, 51
- Electrostriction, 258
- Energy band gap, 389
 - direct band gap materials, 392
 - indirect band gap materials, 392
 - temperature dependence, 496–497
- Energy band structure, 390–393
- Energy distribution of trapped real charges, 308
- Energy transfer processes, 163
 - internal conversion, 166
 - intersystem crossing, 166
 - nonradiative transfer, 166
 - radiative transfer, 166
- Exciton interactions, 162–164
- Excitons, formation and behaviour, 158
- Exciton transport, 161
 - coherent transport, 161
 - incoherent transport, 162
- Experimental methodology and characterization, 472
- Extrinsic conduction, 395
- Extrinsic photoconduction, 490
- F**
- Faraday's law, 5
 - Lenz's law, 5
- Faraday's pail method, 296
- Fermi–Dirac distribution function, 394
 - Fermi levels, 327, 394
- Ferroelectric structure, 242–243
- Ferrimagnetic materials, 24
- Ferroelectric materials, 79, 246
- Ferroelectric phenomena, 216
 - Curie temperature, 216
 - Curie–Weiss relation, 217
 - ferroelectric hysteresis loops, 217
 - first order transition, 220
 - inversible spontaneous polarization, 217
 - second order transition, 220
- Ferroelectric piezoelectricity, 258
- Ferroelectric polar axis, 222
- Ferroelectric polarization, 74
- Ferroelectrics, 213
- Ferromagnetic materials, 24
- Ferromagnetism, 18
 - ferromagnetization, 18
- Field-dependent carrier mobility, 451–452
- Field dependent complex permittivity, 100
 - ferroelectric materials, 101
 - insulating materials, 102
 - semiconducting materials, 100
- Field dependent detrapping processes, 447–451
- Field emission, 354
- Filamentary charge carrier injection, 443
- Filamentary conduction, 444–447
- Fission, 162
- Flat band condition, 334, 464
- Fleming's left-hand rule, 8
- Fluorescence, 165
 - delayed fluorescence, 165, 169
 - prompt fluorescence, 165
- Fluorophors, fluors, 166
- Formation of electrets, 284
 - corona discharge method, 288
 - electromagnetic radiation method, 290
 - (photo- or radio-electrets)
 - electron beam method, 288
 - liquid contact method, 287
 - thermo-electrical method, 284
- Fowler theory, 182
 - Fowler plot, 182
- Fowler–Nordheim equation, 354, 363
 - Fowler–Nordheim tunneling, 354, 516
- Franck–Condon principle, 146
 - Franck–Condon shift, 146
 - Stokes shift, 146
- Franz–Keldysh effect, 157–158
- Free surface charges, 55
- Frequency domain approach, 86
- Frenkel defects, 383
 - Schottky defects, 383
- Frenkel excitons, 158–159
- Frohlich equation, 85
- Fundamental absorption, 200
- Fuoss–Kirkwood equation, 97
- G**
- Garnet ferromagnetic materials, 24
- Gas igniters, 266

- Gauss's law, 9
- Generation of radiation, 144
- Goos–Hanchen effect, 127
- H**
- Hamon approximation, 105
- Havriliak–Negami equation, 97
- High energy electrical pulse generators, 252–255
- High field effects, 443
- Homocharges, 284
 - heterocharges, 284
- Homogeneous photoconduction, 497–499
 - non-homogeneous photoconduction, 499–500
- Homo-pn junctions, 180
 - hetero-pn junctions, 181
- Hopping conduction, 401
- Hot electrons and hot holes, 517–518
- Hysteresis loops, 23, 217
 - for ferroelectric materials, 217–219
 - for ferromagnetic materials, 23
- I**
- Ideality factor, 354
- Impurity conduction, 374
- Indirect band gap materials, 152–154
- Infrared absorption spectroscopy, 522
- Infrared quenching, 504
- Inhibition of electrical treeing, 545
- Injection electroluminescence, 173
 - carrier injection from electrical contacts, 174
 - carrier injection through p–n junctions, 178
 - through p–n homojunctions, 178
 - through p–n heterojunctions, 181
- Inseparable magnetic poles, 8
- Interaction between radiation and matter, 144–164
- Interfacial polarization, 77
- Interference, 118
 - division of amplitude, 119
 - division of wavefront, 119
- Interferometry, 118
 - Fizeau interferometer, 121
 - Michelson interferometer, 121
- Intermolecular phenomenon, 59, 74
 - intermolecular transfer, 398
- Internal discharges in a cavity, 529
- Internal fields, 79
 - local fields for non-dipolar materials, 79
 - reaction fields for dipolar materials, 81
- Intramolecular phenomenon, 59, 74
 - intramolecular transfer, 398
- Intrinsic breakdown, 561
- Intrinsic conduction, 394
- Intrinsic photoconduction, 490
- Ionic conduction, 382
 - extrinsic ionic conduction, 385
 - intrinsic ionic conduction, 383
- Ionic polarization, 59
 - atomic or vibrational polarization, 59
- Ions, negative and positive, 65–66
- Isothermal polarization decay processes, 302
- Isothermal real charge decay processes, 310
- J**
- J–V characteristics, 410, 412, 416, 418–420, 422
- K**
- Kao's model of electrical discharges and breakdown, 566
- Kerr effect, 133
 - quadratic electro-optic effect, 133
- Kirkwood equation, 84
- Kramers–Kronig relations, 91
- L**
- Lampert triangle, 415
- Lande factor, 16
- Langevin function, 17, 69, 70
- Larmor precession, 13
 - Larmor frequency, 13
- Ledge in J–F characteristics, 361–363
- Lifetime electrical conduction, 403
- Lifetime of electrically stressed materials, 524
- Lifetime of non-equilibrium charge carriers, 486
 - linear recombination, 486–487
 - quadratic recombination, 488
 - instantaneous lifetimes, 489
- Light emission, 540
- Light intensity dependence of photoconductivity, 494
 - superlinear photoconductivity, 495
- Light velocity, 2
- Light wavelength dependence of photoconductivity, 495
- Linear piezoelectricity, 258
- Local field, 56
- Lorentz field, 79, 80

Lorentz–Lorenz equation, 80
 Loss factor, 86
 Loss tangent, loss angle, 87
 Luminescence, 164

- bioluminescence, 165
- cathodoluminescence, 164
- chemiluminescence, 165
- electroluminescence, 164
- iono- or radio-luminescence, 165
- photoluminescence, 164
- thermoluminescence, 164
- triboluminescence, 164

 LUMOCEN (luminescence from molecular centers), 173

M

Magnetic domains, 20
 Magnetic materials, 23–26

- hard magnetic materials
- soft magnetic materials

 Magnetic memories, 30
 Magnetic mementos, 20
 Magnetic resonance, 26–30
 Magnetic susceptibility, 10
 Magnetization, 9
 Magneto-optic effects, 142

- Faraday effect, 142
- Voigt effect, 142

 Magnetostriction, 23
 Mass action laws, 396
 Maxwell's equations, 2
 Maxwell–Wagner two layer system, 285
 Measurements of electrical aging, 520–524
 Memory states, 255

- bistable polarization, 255

 Metal-insulator-semiconductor systems, 196, 472
 Minimum sparking potential, 555
 Modulation of light, 128
 Molar polarization, 80
 Monomolecular or unimolecular recombination, 168
 Mott–Gurney equation, 410
 Multiple phonon transition, 150
 Multiplicities, 166
 Multi-quantum processes, 189

N

Nature of light, 115
 Negative differential resistance region, 442, 454

current controlled S-shape NDR, 442
 voltage controlled N-shape NDR, 454
 Negative temperature coefficient materials, 273
 Neutral contacts, 334–336
 Non-centrosymmetric groups, 215
 Noncharacteristic luminescence, 170
 Nonequilibrium charge carriers, 482

- energy distribution, 484
- generation, 482–484
- lifetime, 486
- spatial distribution, 484

 Nonferroelectric materials, 78

- dipolar materials, 78
- nonpolar materials, 78
- paraelectric materials, 78
- polar materials, 78

 Non-homogeneous photoconduction, 499

- npn or pnp phototransistors, 499
- p–n and p–i–n junction diodes
- Schottky barrier photodiodes, 499

 Non-radiative transition processes, 146–154

- Auger recombination processes, 150
- multi-phonon transition, 150
- non-radiative transition due to defects, 152
- non-radiative transition due to indirect band gap structure, 152

 Non-reflecting surface, 120

- antireflection coating, 120

 Nuclear magnetic resonance (NMR), 27

O

One carrier (single) injection SCL electrical conduction, 408

- effects of carrier diffusion, 422
- effects of various trap distributions, 412–420
- scaling rule, 420–422

 Onsager detrapping model, 449
 Onsager equation, 83
 Optical activity, 131
 Optical and electro-optic processes, 115
 Optical constants, 93
 Optical polarization, 59

- electronic polarization, 59

 Optical or infrared quenching, 504
 Orientational polarization, 59, 67–74

P

Paraelectric materials, 78

- dipolar materials, 78
- nonpolar materials, 78

- polar materials, 78
- Paraelectric polarization, 74
- Paramagnetism, 15
- Paschen's law, 555–557
- Permanent magnets, 30
- Permeability, 3
- Permittivity, 3, 55, 123
- Phase diagram of PZT, 231
- Phenomenological approach to piezoelectric effects, 257–262
- Phenomenological approach to pyroelectric effects, 270–271
- Phenomenological properties and mechanisms, 221
 - alloys of PZT type ferroelectric ceramics, 230–234
 - barium titanate type ferroelectrics, 221–227
 - polyvinylidene fluoride type ferroelectrics, 234–236
 - potassium dihydrogen phosphate type ferroelectrics, 227–228
 - Rochelle salt type ferroelectrics, 228–229
 - triglycine sulphate type ferroelectrics, 229–230
- Phosphorescence and phosphors, 169
- Photoconduction, 381, 480, 490
 - extrinsic photoconduction, 490
 - intrinsic photoconduction, 490
- Photoconductivity measurements, 484
 - longitudinal photoconductivity, 485
 - transverse photoconductivity, 485
- Photocurrent-voltage characteristics, 491
- Photo-electric-magnetic (PEM) effect, 193
- Photo-emission, 181, 230
 - from crystalline solids, 185
 - from metallic contacts, 181
- Photoluminescence, 165
 - fluorescence, 165
 - phosphorescence and phosphors, 165
- Photometry, 146
- Photons, 115
- Photorefractive effects, 138–142
- Photoresponse times, 500–502
- Photosensitization, 503
- Photosynthesis photovoltaic effects, 203
- Photovoltaic effects, 191
 - anomalous photovoltaic effects, 206
 - bulk photovoltaic effects, 191
 - contact potential photovoltaic effects, 194–203
 - photosynthesis photovoltaic effects, 203
- Piezoelectric materials, 264
- Piezoelectric parameters and measurements, 262
- Piezoelectric phenomena, 257
- Piezoelectrics, 243
- Piezoelectrics, applications of, 266
 - delay lines, 267
 - gas igniters, 266
 - piezoelectric positioners and actuators, 267
 - piezoelectric transformers, 268
- PIN structure for amorphous Si photovoltaic devices, 201
- Planck's constant, 11
- Plane polarization, 123
 - circularly polarized waves, 123
 - elliptically polarized waves, 123
 - linearly polarized waves, 123
- PN junctions, 178
 - pn heterojunction devices, 181
 - pn homojunction devices, 178
- Pockels effect, 131
 - linear electro-optic effect, 131
- Polarizability, 55
- Polarization, 55, 123
- Polarizers, 124–127
 - light polarized by reflection, 124–126
 - light polarized by transmission, 126–127
- Polarons, 402–403
- Poling processes, 297
- Polyvinylidene fluoride type ferroelectrics, 221, 234
- Poole–Frenkel detrapping model, 447
- Positive temperature coefficient materials, 273
- Potassium dihydrogen phosphate type ferroelectrics, 221, 227
- Potassium sodium tartrate tetrahydrate type ferroelectrics, 221, 228
- Potential barrier height and Schottky effect, 345
- Prebreakdown disturbance, 540
 - light emission, 540
- Principle of superposition, 117
- Protons, 115
- Pulsed electro-acoustic (PEA) method, 305–308
- Pyroelectric and thermally sensitive materials, 272
 - NTC materials, 273
 - PTC materials, 274
- Pyroelectric coefficients, 270
- Pyroelectric parameters and measurements, 271
- Pyroelectric phenomena, 269

Pyroelectrics, applications of, 275
 pyroelectric burglar alarm system, 278
 pyroelectric energy conversion, 279
 pyroelectric radiation detectors, 275, 322
 pyroelectric thermometry, 278

Q

Quantum numbers, 11
 Quantum efficiency for photoconduction, 480
 Quantum yield, 167, 480
 Quarter wave plates (QWP), 130
 Quenching or poisoning centers, 504

R

Radiation induced conductivity (RIC), 289
 Radiative transition processes, 146–149
 Radicals, 521
 radical scavengers, 526–527
 Radiometry, 146
 Raman–Nath diffraction, 144
 Random walk model, 505
 Rayleigh criterion, 122
 Recombination cross section, 426
 Recombination processes, 433
 kinetics of recombination processes, 433
 band-to-band recombination, 434
 with a single set of recombination centers, 435
 Recombination rate, 426
 Rectifying contact, 336
 Redox reaction, 204
 Reduced mass, 67
 Reflection, 123
 reflectance, 126
 Refraction, 123
 Relative permittivity, 55
 Relaxation electrical conduction, 404–406
 Relaxation processes, 59, 87
 Relaxation regime, 93, 157
 Relaxation times, 88, 296
 distribution of relaxation times, 108
 Remanent polarization, 217
 Remedy for electrical aging, 525
 emission shields, 525
 radical scavengers, 526
 Resonance process, 87
 Resonance regime, 93, 157
 Resonance function, 88
 Response time, 433

Richardson line, 461
 Rydberg energy, 160

S

Sawyer and Tower method, 217
 Scaling rule, 420
 Scher and Montroll model, 505
 Schlieren electro-optical system, 541
 Schottky barrier photovoltages, 194
 Schottky barriers, 334
 Schottky effects, 345
 Shallow traps, 413
 Shockley states, 342
 Similarity in breakdown mechanisms for
 gas, liquid, and solid dielectrics, 567
 Small angle scattering of X-ray spectroscopy, 523
 Space charge limited electrical conduction, 406
 Space charge polarization, 59, 75
 hopping polarization, 75–76
 interfacial polarization, 77–78
 Spatial distribution of dipolar polarization, 302
 Spatial distribution of trapped real charges, 303, 305
 electron beam sampling method, 304
 pulsed electro-acoustic method, 305–308
 sectioning method, 303
 other methods, 308
 Spin magnetic moments, 24
 antiferromagnetic, 24
 ferrimagnetic, 24
 ferromagnetic, 24
 garnet ferromagnetic, 24
 Spin-orbit interaction, 15
 Spontaneous polarization, 74, 217
 Static polarization, 52
 Statistical-mechanical approaches, 84
 Strokes's theorem, 3
 Superposition, principle of, 117
 Surface charge, 55
 Surface charge density measurements, 294
 compensation method, 294
 capacitive probe method, 295
 Surface discharges and corona discharges, 546
 Surface potential measurements, 477, 523
 Surface states, 341
 Susceptibility, 14, 56

electric susceptibility, 56
magnetic susceptibility, 14

T

Tamm states, 342
Temperature dependence of complex permittivity, 98
Thermal breakdown, 559
Thermal hysteresis, 240
Thermal pulse method, 296
Thermal quenching, 503
Thermal radiation, 144–145
Thermal velocity of carriers, 397
Thermally stimulated discharge current, 298–299
Thermionic emission, 330, 350
Thermionic-field emission, 368
Thermo-autostabilization nonlinear dielectric elements, 252
TANDEL, 252
Thermodynamic theory, 236
Helmholtz free energy function, 236
Gibbs free energy function, 237
Thermoluminescence, 164
Threshold voltages, 455–463
Time-dependent electric polarization, 87
Time domain approach, 86
Time of flight measurements, 467–468
Total charges, measurements of, 296
Faraday pail method, 296
thermal pulse method, 296
Total internal reflection, 127
Transient current, 463
space charge free transient, 466
space charge limited transient, 468
space charge perturbed transient, 470
Transient photoconduction, 505
Transitions between electrical conduction processes, 458
basic transition processes, 455
from bulk limited to electrode limited processes, 460
from electrode limited to bulk limited processes, 461
from single injection to double injection processes, 462
Transitions between crystal structures, 220
first order transition, 220

second order transition, 220

Trapping processes, 423
capture cross section, 425
capture rates, 426
Triboluminescence, 164
Triglycine sulphate, 221, 229
Triplet excitons, 166
Tunneling through thin dielectric film in MIM or MIS systems, 364–367, 371
Two carrier (double) injection SLC electrical conduction, 437–442

U

Uniaxial crystals, 130
negative uniaxial crystals, 130
positive uniaxial crystals, 130
Unit of Debye, 38
Unit of mole, 38
Units of light, 145
photometry, lumens, 146
radiometry, 145
Universality, 505

V

Vacuum level, 134
Van der Waals bonds, 398
Vibronic states, 146–148, 167
Virtual electrodes, 304, 341

W

Wannier excitons, 158–160
Wave theory, 116
Wentzel-Kramers-Brillouin (WKB) approach, 154
Work functions, 328–330

X

X-rays, small angle scattering, 523

Y

Young's double-slit experiment on interference, 119

Z

Zeeman effect, 14
Zeeman splitting, 14
Zener effect, 157
Zener breakdown, 563

