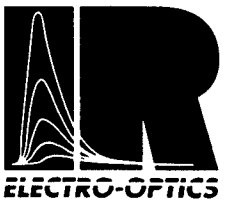
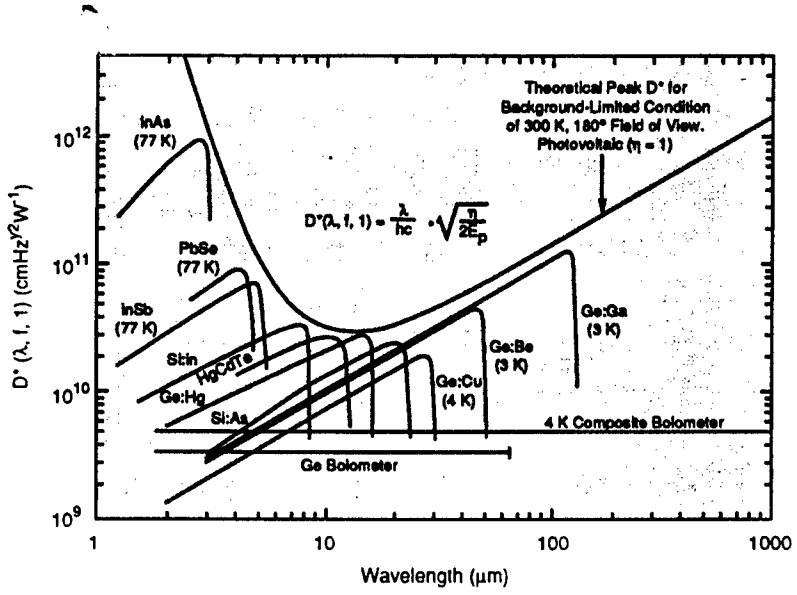


The Infrared &
Electro-Optical
Systems Handbook

VOLUME 3

Electro-Optical Components

William D. Rogatto, *Editor*



Electro-Optical Components

VOLUME 3

The Infrared and Electro-Optical Systems Handbook

The Infrared and Electro-Optical Systems Handbook

Joseph S. Accetta, David L. Shumaker, *Executive Editors*

Chapter 3. Optomechanical System Design, Daniel Vukobratovich
Chapter 4. Infrared Imaging System Testing, Gerald C. Holst
Chapter 5. Tracking and Control Systems, Robert E. Nasburg
Chapter 6. Signature Prediction and Modeling, John A. Conant, Malcolm A. LeCompte

■ **VOLUME 5.** **Passive Electro-Optical Systems,**
Stephen B. Campana, *Editor*

Chapter 1. Infrared Line Scanning Systems, William L. McCracken
Chapter 2. Forward-Looking Infrared Systems, George S. Hopper
Chapter 3. Staring-Sensor Systems, Michael J. Cantella
Chapter 4. Infrared Search and Track Systems, Joseph S. Accetta

■ **VOLUME 6.** **Active Electro-Optical Systems,** Clifton S. Fox, *Editor*

Chapter 1. Laser Radar, Gary W. Kamerman
Chapter 2. Laser Rangefinders, Robert W. Byren
Chapter 3. Millimeter-Wave Radar, Elmer L. Johansen
Chapter 4. Fiber Optic Systems, Norris E. Lewis, Michael B. Miller

■ **VOLUME 7.** **Countermeasure Systems,** David Pollock, *Editor*

Chapter 1. Warning Systems, Donald W. Wilmot, William R. Owens, Robert J. Shelton
Chapter 2. Camouflage, Suppression, and Screening Systems, David E. Schmieder, Grayson W. Walker
Chapter 3. Active Infrared Countermeasures, Charles J. Tranchita, Kazimieras Jakstas, Robert G. Palazzo, Joseph C. O'Connell
Chapter 4. Expendable Decoys, Neal Brune
Chapter 5. Optical and Sensor Protection, Michael C. Dudzik
Chapter 6. Obscuration Countermeasures, Donald W. Hoock, Jr., Robert A. Sutherland

■ **VOLUME 8.** **Emerging Systems and Technologies,**
Stanley R. Robinson, Editor

Chapter 1. Unconventional Imaging Systems, Carl C. Aleksoff, J. Christopher Dainty, James R. Fienup, Robert Q. Fugate, Jean-Marie Mariotti, Peter Nisenson, Francois Roddier
Chapter 2. Adaptive Optics, Robert K. Tyson, Peter B. Ulrich
Chapter 3. Sensor and Data Fusion, Alan N. Steinberg
Chapter 4. Automatic Target Recognition Systems, James W. Sherman, David N. Spector, C. W. "Ron" Swonger, Lloyd G. Clark, Edmund G. Zelnio, Terry L. Jones, Martin J. Lahart
Chapter 5. Directed Energy Systems, Gary Golnik
Chapter 6. Holography, Emmett N. Leith
Chapter 7. System Design Considerations for a Visually-Coupled System, Brian H. Tsou

Copublished by

Infrared Information Analysis Center
Environmental Research Institute of Michigan
Ann Arbor, Michigan USA

and

SPIE OPTICAL ENGINEERING PRESS
Bellingham, Washington USA

Sponsored by

Defense Technical Information Center, DTIC-DF
Cameron Station, Alexandria, Virginia 22304-6145

Electro-Optical Components

William D. Rogatto, *Editor*
Santa Barbara Research Center

VOLUME 3

The Infrared and Electro-Optical Systems Handbook

Joseph S. Accetta, David L. Shumaker, *Executive Editors*
Environmental Research Institute of Michigan

Library of Congress Cataloging-in-Publication Data

The Infrared and electro-optical systems handbook / Joseph S. Accetta, David L. Shumaker, executive editors.
p. cm.
Spine title: IR/EO systems handbook.
Cover title: The Infrared & electro-optical systems handbook.
Completely rev. ed. of: Infrared handbook. 1978
Includes bibliographical references and indexes.
Contents: v. 1. Sources of radiation / George J. Zissis, editor — v. 2. Atmospheric propagation of radiation / Fred G. Smith, editor — v. 3. Electro-optical components / William D. Rogatto, editor — v. 4. Electro-optical systems design, analysis, and testing / Michael C. Dudzik, editor — v. 5. Passive electro-optical systems / Stephen B. Campana, editor — v. 6. Active electro-optical systems / Clifton S. Fox, editor — v. 7. Countermeasure systems / David Pollock, editor — v. 8. Emerging systems and technologies / Stanley R. Robinson, editor.
ISBN 0-8194-1072-1
1. Infrared technology—Handbooks, manuals, etc. 2. Electrooptical devices—Handbooks, manuals, etc. I. Accetta, J. S. II. Shumaker, David L. III. Infrared handbook. IV. Title: IR/EO systems handbook. V. Title: Infrared & electro-optical systems handbook.
TA1570.I5 1993
621.36'2—dc20 92-38055
CIP

Copublished by

Infrared Information Analysis Center
Environmental Research Institute of Michigan
P.O. Box 134001
Ann Arbor, Michigan 48113-4001

and

SPIE Optical Engineering Press
P.O. Box 10
Bellingham, Washington 98227-0010

PRINTED IN THE UNITED STATES OF AMERICA

Preface

The Infrared and Electro-Optical Systems Handbook is a joint product of the Infrared Information Analysis Center (IRIA) and the International Society for Optical Engineering (SPIE). Sponsored by the Defense Technical Information Center (DTIC), this work is an outgrowth of its predecessor, *The Infrared Handbook,* published in 1978. The circulation of nearly 20,000 copies is adequate testimony to its wide acceptance in the electro-optics and infrared communities. *The Infrared Handbook* was itself preceded by *The Handbook of Military Infrared Technology.* Since its original inception, new topics and technologies have emerged for which little or no reference material exists. This work is intended to update and complement the current *Infrared Handbook* by revision, addition of new materials, and reformatting to increase its utility. Of necessity, some material from the current book was reproduced as is, having been adjudged as being current and adequate. The 45 chapters represent most subject areas of current activity in the military, aerospace, and civilian communities and contain material that has rarely appeared so extensively in the open literature.

Because the contents are in part derivatives of advanced military technology, it seemed reasonable to categorize those chapters dealing with systems in analogy to the specialty groups comprising the annual Infrared Information Symposia (IRIS), a Department of Defense (DoD) sponsored forum administered by the Infrared Information Analysis Center of the Environmental Research Institute of Michigan (ERIM); thus, the presence of chapters on active, passive, and countermeasure systems.

There appears to be no general agreement on what format constitutes a "handbook." The term has been applied to a number of reference works with markedly different presentation styles ranging from data compendiums to tutorials. In the process of organizing this book, we were obliged to embrace a style of our choosing that best seemed to satisfy the objectives of the book: to provide derivational material data, descriptions, equations, procedures, and examples that will enable an investigator with a basic engineering and science education, but not necessarily an extensive background in the specific technology, to solve the types of problems he or she will encounter in design and analysis of electro-optical systems. Usability was the prime consideration. In addition, we wanted each chapter to be largely self-contained to avoid time-consuming and tedious referrals to other chapters. Although best addressed by example, the essence of our handbook style embodies four essential ingredients: a brief but well-referenced tutorial, a practical formulary, pertinent data, and, finally, example problems illustrating the use of the formulary and data.

The final product represents varying degrees of success in achieving this structure, with some chapters being quite successful in meeting our objectives and others following a somewhat different organization. Suffice it to say that the practical exigencies of organizing and producing a compendium of this magnitude necessitated some compromises and latitude. Its ultimate success will be judged by the community that it serves. Although largely oriented toward system applications, a good measure of this book concentrates on topics endemic and fundamental to systems performance. It is organized into eight volumes:

Volume 1, edited by George Zissis of ERIM, treats sources of radiation, including both artificial and natural sources, the latter of which in most military applications is generally regarded as background radiation.

Volume 2, edited by Fred Smith of OptiMetrics, Inc., treats the propagation of radiation. It features significant amounts of new material and data on absorption, scattering, and turbulence, including nonlinear propagation relevant to high-energy laser systems and propagation through aerodynamically induced flow relevant to systems mounted on high-performance aircraft.

Volume 3, edited by William Rogatto of Santa Barbara Research Center, treats traditional system components and devices and includes recent material on focal plane array read-out electronics.

Volume 4, edited by Michael Dudzik of ERIM, treats system design, analysis, and testing, including adjunct technology and methods such as trackers, mechanical design considerations, and signature modeling.

Volume 5, edited by Stephen Campana of the Naval Air Warfare Center, treats contemporary infrared passive systems such as FLIRs, IRSTs, IR line scanners, and staring array configurations.

Volume 6, edited by Clifton Fox of the Night Vision and Electronic Sensors Directorate, treats active systems and includes mostly new material on laser radar, laser rangefinders, millimeter-wave systems, and fiber optic systems.

Volume 7, edited by David Pollock, consultant, treats a number of countermeasure topics rarely appearing in the open literature.

Volume 8, edited by Stanley Robinson of ERIM, treats emerging technologies such as unconventional imaging, synthetic arrays, sensor and data fusion, adaptive optics, and automatic target recognition.

Acknowledgments

It is extremely difficult to give credit to all the people and organizations that contributed to this project in diverse ways. A significant amount of material in this book was generated by the sheer dedication and professionalism of many esteemed members of the IR and EO community who unselfishly contributed extensive amounts of precious personal time to this effort and to whom the modest honorarium extended was scarcely an inducement. Their contributions speak elegantly of their skills.

Directly involved were some 85 authors and editors from numerous organizations, as well as scores of technical reviewers, copyeditors, graphic artists, and photographers whose skill contributed immeasurably to the final product.

We acknowledge the extensive material and moral support given to this project by various members of the managements of all the sponsoring and supporting organizations. In many cases, organizations donated staff time and internal resources to the preparation of this book. Specifically, we would like to acknowledge J. MacCallum of DoD, W. Brown and J. Walker of ERIM, and J. Yaver of SPIE, who had the foresight and confidence to invest significant resources in the preparation of this book. We also extend our appreciation to P. Klinefelter, B. McCabe, and F. Frank of DTIC for their administrative support during the course of this program.

Supporting ERIM staff included Ivan Clemons, Jenni Cook, Tim Kellman, Lisa Lyons, Judy Steeh, Barbara Wood, and the members of their respective organizations that contributed to this project.

We acknowledge Lorretta Palagi and the publications staff at SPIE for a professional approach to the truly monumental task of transforming the manuscripts into presentable copy and the patience required to interact effectively with the authors.

We would like to pay special tribute to Nancy Hall of the IRIA Center at ERIM who administrated this at times chaotic project with considerable interpersonal skill, marshaling the numerous manuscripts and coordinating the myriad details characteristic of a work of this magnitude.

We properly dedicate this book to the people who created it and trust it will stand as a monument to their skills, experience, and dedication. It is, in the final analysis, a product of the community it is intended to serve.

Joseph S. Accetta
David L. Shumaker
Ann Arbor, Michigan

January 1993

Notices and Disclaimer

This handbook was prepared by the Infrared Information Analysis Center (IRIA) in cooperation with the International Society for Optical Engineering (SPIE). The IRIA Center, Environmental Research Institute of Michigan, is a Defense Technical Information Center-sponsored activity under contract DLA-800-C-393 and administrated by the Defense Electronics Supply Center, Defense Logistics Agency.

This work relates to the aforementioned ERIM contract and is in part sponsored by the Department of Defense; however, the contents do not necessarily reflect the position or the policy of the Department of Defense or the United States government and no official endorsement should be inferred.

Contents

CHAPTER 1

Optical Materials

William L. Wolfe
Optical Sciences Center/University of Arizona
Tucson, Arizona

CONTENTS

1.1 INTRODUCTION

For the purposes of this chapter, optical materials are considered to be the transparent materials that are useful for windows, lenses, prisms and the like, mirror substrates, mirror coatings, and various blackening agents that may be used in baffles or for thermal detectors. Although of great importance, diamond-like materials are not included in this chapter. They are not "stand-alone" materials. Likewise, filters and films are special subjects that are treated elsewhere.

1.2 DESCRIPTION OF PROPERTIES

The properties of optical materials that are pertinent to their use as windows, lenses, mirrors, substrates, beamsplitters, and other elements are described in this section. Equations and some of the theoretical descriptions of material properties are also provided.

1.2.1 Reflection and Transmission of Nonabsorbing Materials

This section presents expressions for the reflection and transmission of nonabsorbing materials for polarized and unpolarized radiation. The expressions are forms of Snell's and Fresnel's laws for amplitudes and for power forms.

Snell's law states that the angle of reflection from a surface is equal to the angle of incidence when both are measured with respect to the surface normal. It also states that the two rays are in the plane of incidence that is defined by the incident ray and the surface normal. The refracted ray direction is given by the law of sines and is also in the plane of incidence:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \ .$$

The subscripts represent the two media. The two reflected-wave amplitudes are given by

$$\tilde{r}_s = |r_s| e^{j\delta} = \frac{E_s^{\text{rfl}}}{E_s^{\text{inc}}} = -\frac{\sin(\theta_1 - \theta_3)}{\sin(\theta_1 + \theta_3)} e^{j\delta} \ , \tag{1.1}$$

$$\tilde{r}_p = |r_p| e^{j\delta} = \frac{E_p^{\text{rfl}}}{E_p^{\text{inc}}} = \frac{\tan(\theta_1 - \theta_3)}{\tan(\theta_1 + \theta_3)} e^{j\delta} \ , \tag{1.2}$$

where

$$j = \sqrt{-1} \ ,$$

$$r = \tilde{E}_i / \tilde{E}_r \ ,$$

$$\delta = \frac{2\pi n d}{\lambda} \ .$$

The subscript 1 indicates the incident medium and incident angle; subscript 2 refers to the refracted ray; and subscript 3 refers to the reflected ray. The phase shift δ is given by the product of 2π and the optical path nd, divided by the wavelength λ. The optical path is the refractive index n times the true length d.

The ratios of the refracted (transmitted) wave to the incident one for the two polarizations are given by

$$\tilde{t}_s = |t_s|\, e^{j\delta} = \frac{2 \sin\theta_2 \cos\theta_1}{\sin(\theta_1 + \theta_2)}\, e^{j\delta} \ , \tag{1.3}$$

$$\tilde{t}_p = |t_p|\, e^{j\delta} = \frac{2 \sin\theta_2 \cos\theta_1}{\sin(\theta_1 + \theta_2)\cos(\theta_1 - \theta_2)}\, e^{j\delta} \ . \tag{1.4}$$

The four amplitude ratios can also be cast in a form that is dependent only on the refractive index and the angle of incidence:

$$\tilde{r}_s = \frac{(n_2^2 - n_1^2 \sin^2\theta_1)^{1/2} - n_1 \sin\theta_1}{(n_2^2 - n_1^2)^{1/2} - n_1 \sin\theta_1 \cos\theta_1} \ , \tag{1.5}$$

$$\tilde{t}_s = \frac{2n_1 \sin\theta_1}{(n_2^2 - n_1^2 \sin^2\theta_1)^{1/2} + n_1 \sin\theta_1} \ , \tag{1.6}$$

$$\tilde{r}_p = \frac{(n_2^2 - n_1^2 \sin^2\theta_1)^{1/2} \sin\theta_1 - n_1 \sin\cos\theta_1}{(n_2^2 - n_1^2 \sin^2\theta_1)^{1/2} - n_1 \sin\theta_1 \cos\theta_1} \ , \tag{1.7}$$

$$\tilde{t}_p = \frac{2n_1 n_2 \sin\theta_1}{n_1 \sin\theta_1 + (n_2^2 - n_1^2 \sin\theta_1)^{1/2}} \ . \tag{1.8}$$

1.2.2 Absorption

Absorption of radiation is described by Beer's law, which states that equal thicknesses of an absorbing material absorb equal fractions of the power incident on them. This may be written as

$$\frac{d\Phi}{\Phi} = -\alpha\,(\lambda)\, dx \ . \tag{1.9}$$

The fraction of power absorbed is on the left side. The negative sign indicates a loss mechanism. The constant of proportionality is α, and the differential thickness is dx. Absorption is a function of wavelength. For most purposes, α is found experimentally for a finite but small spectral bandwidth. The solution is the well-known exponential law of absorption:

$$\Phi = \Phi_0 e^{-\alpha x} \ . \tag{1.10}$$

Unfortunately, α is used in two different and closely related ways. In the first sense of the equation, it is the ratio of power that is lost to the material. This is called the *absorptivity* or *absorptance*. In the second use, in the exponent, it is the absorption coefficient. The pathlength in the material is usually designated by some Cartesian coordinate or r or s or d. The internal transmittance of a sample is defined as the ratio of flux at the second surface to that at the first. This is the ones-complement of the absorption:

$$\tau_i = 1 - e^{-\alpha x} \simeq \alpha x \ . \tag{1.11}$$

The absorption coefficient is normally expressed in units of reciprocal centimeters (cm^{-1}), but it can be expressed in other units if x is expressed in the reciprocal units. For instance, if x is in kilometers, α might be expressed as dB/km. If x is expressed in atm cm of a gas, then α is expressed as the reciprocal of that. The argument of the exponential must be dimensionless.

The absorption is sometimes expressed in terms of the extinction coefficient. This is the imaginary part of the complex refractive index:

$$\tilde{n} = n - j\kappa \ , \tag{1.12}$$

where κ is the extinction coefficient. It is related to the absorption coefficient by

$$\alpha = \frac{4\pi\kappa}{\lambda} \ . \tag{1.13}$$

There is some ambiguity in these definitions; for an explanation, see Section 1.2.4 on the refractive index.

1.2.3 Transmission, Reflection, Absorption, and Emission for an Absorbing Sample

A plane parallel plate that reflects, absorbs, and transmits has effective values of these quantities as a result of the multiple reflections within the sample. If ρ represents the single-surface Fresnel reflectivity and τ represents the internal transmittance, then it may be shown that the effective, or external, transmittance and the effective reflectance are given by the following equations:

$$\tau_\infty = \frac{(1 - \rho)^2\tau}{1 - \rho^2\tau^2} \ , \tag{1.14}$$

$$\rho_\infty = \rho + \frac{\rho\tau(1 - \rho)^2\tau}{1 - \rho^2\tau^2} \ . \tag{1.15}$$

The effective absorptivity, which is equal to the effective emissivity, is then given by

$$\varepsilon_\infty = \alpha_\infty = \frac{(1 - \rho)\,(1 - \tau)}{1 - \rho\tau} \ . \tag{1.16}$$

1.2.4 Refractive Index

The refractive index has been defined as the ratio of the velocity of light in a vacuum to that in a medium. This relates to the real part of the refractive index for materials and includes the special case of a mirror, which gives a value of -1. The refractive index has also been defined as a complex quantity, where the real part may be thought of as representing the phase delay, the retardation of the light, and the imaginary part, its attenuation.

The complex refractive index is normally defined as

$$\tilde{n} = n - j\kappa \ . \tag{1.17}$$

It may also be defined as

$$\tilde{n} = n\,(1 - j\kappa) \ . \tag{1.18}$$

The obvious difference relates to whether there is a refractive index in the extinction term.

The refractive index is a function of both wavelength and temperature, as well as other less important parameters. The variation with wavelength is called the *dispersion*. The two main types of dispersion equations are attributed to Herzberger and to Sellmeier. They are, respectively,

$$n = A + \frac{B}{\lambda^2 - 0.028} + \frac{C}{(\lambda^2 - 0.028)^2} + D\lambda^2 + E\lambda^4 \ , \tag{1.19}$$

$$n^2 - 1 = \sum \frac{K_i \lambda^2}{\lambda_2 - \lambda_i^2} \ , \tag{1.20}$$

where A, B, C, D, E, F, K_i, and λ_i are constants. Their values are theoretically related directly to the various absorption bands of the material, but are usually found empirically.

The change with temperature is sometimes called the thermorefractive coefficient and is given by dn/dT. It usually has values of about several millionths per degree, and is a function of both the wavelength and the temperature. The coefficient is not a constant.

1.2.5 Thermal Properties

Although almost every physical property of a material depends on temperature, some properties are more dependent on temperature than others. In this section the following properties are described and defined: melting and softening temperature, specific heat, heat capacity, heat capacitance, thermal conductivity, thermal conductance, and thermal expansion.

The *melting temperature* is the temperature at which a crystal melts to form a liquid from a solid. This is also known as the temperature of fusion or the fusing temperature. In very careful work, the solidification temperature is differentiated from the fusion temperature.

The *transition temperature* is the temperature at which a glass may be annealed. For glasses, DIN52324 defines a transition temperature in terms of the thermal expansion coefficient. At low temperatures, near absolute zero, the curve of relative expansion has a small and slightly curving slope. At higher temperatures, approaching room temperature, the curve is linear but increasing. At increasing temperatures, the curve gradually changes to a second linear region with a much higher slope. The intersection of the extensions of these two linear expansion regions defines transition temperature T_g. The transformation temperature differs from the annealing temperature by about 10° as defined by ASTM C336-1.

Since a glass may be thought of as a supercooled liquid, it does not have an abrupt transition from the liquid phase to the solid phase (or vice versa). The definition of *softening temperature* must therefore be somewhat arbitrary. The softening temperature is higher than the transition point and is the point at which plastic deformation begins to occur. It is defined in ASTM C338-73.

Specific heat is a measure of the degree to which a certain amount of heat increases the temperature of a body of a given mass. Specific heat is the ratio of the heat capacity of a material to that of water. The *heat capacity* is the amount of heat it takes to raise a unit mass of a material one degree. In the English system of units it is specified in Btu lb^{-1} $°F^{-1}$. In the SI system, it is J g^{-1} K^{-1}. An alternative is to specify the heat in calories, in which case differentiation between calories and Calories (kilocalories) must be made. Heat capacitance is the heat capacity times the mass of the material. It has the same relationship to heat capacity as electric capacitance has to electric capacity. As with most physical quantities, heat capacity changes with temperature. It may be a second-order effect for many applications.

Thermal expansion is a measure of the change in a material's dimensions as a result of a change in temperature. It is often described in terms of the linear thermal expansion. In this instance, *linear* describes the geometry—in one dimension. The variation of linear expansion with temperature is almost never linear, that is, a straight line. The volume expansion is just three times the linear expansion. The definition of linear expansion is the change in length divided by the length of a sample for a 1° change in temperature. Thus the units are reciprocal degrees, preferably kelvin or Celsius degrees. The table for these values is based on the simple equation

$$\alpha = \frac{1}{L}\frac{dL}{dT} = A + BT + \mathrm{C}T^2 \ . \tag{1.21}$$

Higher order terms could be used, but the constant, linear, and quadratic terms usually suffice. (Note that this raises the possibility of the expression *linear, linear expansion coefficient.*)

Thermal conductivity is a measure of the ease with which a material conducts heat. The rate of heat flow is proportional to the cross-sectional area and temperature difference, and it is inversely proportional to the length. The proportionality constant is the thermal conductivity, *k:*

$$q = \frac{kA}{l}\Delta T \ . \tag{1.22}$$

Thermal conductivity is a function of temperature, and a constant value is only an approximation.

The units are energy per time per cross-sectional area per degree times the length, and thus in SI terms are W m^{-1}. Thermal conductance is the thermal conductivity times the area of the sample divided by its length. It is analogous to electrical conductance compared to electrical conductivity.

1.2.6 Debye Temperature

The Debye temperature normalizes the properties of many important characteristics of infrared materials. It is defined by the following equation:

$$\theta_D = \frac{h\nu_D}{k} , \tag{1.23}$$

where k is the Boltzmann constant.

1.2.7 Hardness

Some materials are harder than others. Some are more difficult to polish and to scratch or abrade in use. There are several measures of hardness. The Moh scale is based strictly on which materials scratch others. The Moh scale ranges from 1 to 10 and does not have equal steps of hardness from one number to the next. The Moh scale is given in Table 1.1.

Most other methods of measuring hardness, which are more precise, are based on pressing an indenter of a specially prescribed shape into the material. The measure is a description of the area or length of the indentation, and the units are generally kilograms per square millimeter (kg mm^{-2}). Usually, and sometimes necessarily, the load on the indenter is specified.

Three types are the Vickers, Brinnell, and Knoop indenters. The Vickers hardness is the load in kilograms divided by the area of indentation made by a pyramidal indenter that has an angle of 136 deg between opposite faces and 146 deg between opposite edges. Brinnell hardness is the load in kilograms divided by the curved area made by a spherical indenter. Knoop values are obtained with an indenter almost identical to the Vickers indenter. The area

Table 1.1 Moh Hardness Scale

1	Talc	$Mg_3Si_4O_{10}(OH)_2$
2	Gypsum	$CaSO_4 \cdot 2H_2O$
3	Calcite	$CaCO_3$
4	Fluorite	CaF_2
5	Apatite	$Ca_5(PO_4)_3(F,Cl,OH)$
6	Orthoclase	$KAlSi_3O_8$
7	Quartz	SiO_2
8	Topaz	$Al_2SiO_4(OH,F)_2$
9	Corundum	Al_2O_3
10	Diamond	C

is usually stated in square millimeters. Some measurements vary with the applied load, especially if cracks, cold flow, or stress relief occur.

1.2.8 Solubility

Some infrared materials are highly soluble in water; others are hardly attacked at all. Almost all are attacked to some degree by acids and alkalies. The solubilities quoted here are in terms of the number of grams of water that are dissolved in 100 ml of water. Since water has a density of 1 g ml^{-1}, this is approximately a percentage of solution. The solubility is at the temperature specified or approximately room temperature if not stated. The solubility may be taken as a measure of weathering ability of the material, but surely not rain erosion resistance. For example, a window of salt, NaCl, left overnight in Michigan will deteriorate, if not dissolve, in the high humidity. The same window will survive on a normal day in Arizona!

Stain is often specified for glasses for use in the visible region. It has no meaning or usefulness for infrared applications.

1.2.9 Scattering

Mirrors, lenses, and windows all scatter to some degree. The degree of scatter can be described by the bidirectional reflectivity or transmissivity. This has come to be called by many the *bidirectional reflectance distribution function* (BRDF) or the *bidirectional transmittance distribution function* (BTDF). The term to describe both or either is the *bidirectional scattering distribution function* (BSDF), which is often plotted as a function of the sine of the scatter angle minus the sine of the angle of specular reflection, often referred to as $\beta - \beta_0$ or $\delta\beta$. This leads to a nice normalization, because most materials then have values that are largely independent of the incident angle.

For a smooth surface, one that has an rms roughness smaller than the wavelength of the incident light, the expressions for scatter are

$$\rho_b = k^3 F(\theta) \cos\theta_i \cos\theta_s W(p,q) \ , \tag{1.24}$$

where ρ_b is the bidirectional reflectivity, k is $2\pi/\lambda$, $F(\theta)$ is the geometry factor (which will be discussed later), and $W(p,q)$ is the power spectrum of the surface height distribution as a function of the radian spatial frequencies p and q. The optical factor is different for different polarizations. These are given here, where s means polarized perpendicular to the plane of incidence and p is parallel. The two together indicate the states of polarization of the source and receiver, respectively. The incident angle is θ_i, the scatter angle in the plane of incidence is θ_s, and the scatter angle out of the plane of incidence is ϕ:

$$F_{ss} = \cos^2\phi \ , \tag{1.25}$$

$$F_{pp} = \frac{\cos\phi|\cos\phi - \sin\theta_i \sin\theta_s|^2}{|\cos\theta_i \cos\theta_s|^2} \ , \tag{1.26}$$

$$F_{sp} = \frac{\sin^2\phi}{\cos^2\theta_s} \ , \tag{1.27}$$

$$F_{ps} = \frac{\sin^2\phi}{\cos^2\theta_i} \,. \tag{1.28}$$

The spectrum of the surface height is usually of the form

$$W(u) = \frac{\sigma^2\delta}{\pi(1 + u^2\delta^2)} \,, \tag{1.29}$$

where σ is the rms surface height, δ is the autocorrelation length, and u is the radial spatial frequency ($u^2 = p^2 + q^2$). The asymptotic log-log plot of this is a horizontal straight line to the point where $u = 1/\delta$, followed by a line of slope -2. It is *extremely* similar to the Bode plot of a single-time-constant circuit.

The maximum value of the bidirectional reflectivity is limited only by the spread of the scattered beam and may be many orders of magnitude larger than unity. The value for a perfectly isotropic reflector is the hemispherical reflectivity ρ_h divided by π.

1.2.10 Elastic Moduli

Most elastic moduli are defined in terms of stress and strain. A stress is a force per unit area, a pressure. If the force is normal to the body, the stress is dilational or compressional. If the force is parallel to a surface, the stress is a shear. A dilational strain can be the change in length divided by the mean or the original length. A shear is the difference in displacement of two parallel planes divided by the distance between them. The bulk modulus, sometimes indicated by k, is the compressional stress divided by the volume strain (relative change in volume). Young's modulus E is the tensile stress divided by the linear strain. The shear modulus or rigidity is the shear stress divided by the shear strain. Poisson's ratio is the lateral strain divided by the linear extensional strain, the ratio of the extension in one direction to the contraction in the other.

1.2.11 Density and Specific Gravity

The density of a material is its mass per unit volume. It may be expressed in either the English or metric system of units, or as a specific gravity. The specific gravity is the ratio of the density of a substance to the density of water at 4°C. The specific gravity of the material is usually accompanied by a reference temperature. Since the density of water in the metric system is approximately 1 g cm^{-3}, the density and the specific gravity have the same numerical values. The density of water in the English system is 62.4 lb ft^{-3}.

1.2.12 Engineering Moduli

Hooke's law states that for small deformations, the stress acting on a solid is proportional to the strain existing within it. The components of stress are linear functions of the components of strain; the proportionality constants are called the *stiffness constants* or *elastic coefficients*. They are usually designated

as c_{ij}, and they have values from 1 to 6. For cubic crystals, there are three independent c's, where $i = j = 1, 2, 4$. For tetragonal crystals, there are five independent coefficients, where $i = j = 11, 12, 13, 33$ and 44. For hexagonal crystals there are six independent coefficients, 11, 12, 13, 14, 33, and 44. The elastic compliance constants are the coefficients of the inverse ratios. They are denoted by s and have the same set of subscripts.

There are several engineering moduli of importance. They include Young's modulus, the shear modulus, the modulus of rigidity, the bulk modulus, and Poisson's ratio.

Young's modulus is defined as the ratio of stress to strain, where stress is the force per unit area perpendicular to the direction in which the force is applied and the strain is the fractional change in length in the direction of the force. The modulus of rigidity,

$$H_K = \frac{F/A}{dl/l} , \tag{1.30}$$

is the ratio of shearing stress to shearing strain, where the shearing stress is the force per unit area parallel to the direction in which the force is applied, and the shear strain is the angle of shear in radians:

$$H_R = \frac{F_{\parallel}/A}{\theta} . \tag{1.31}$$

Young's modulus may be calculated from the elastic constants by the following equations:

$$H_K = \frac{(c_{11} + 2c_{12})(c_{11} - c_{12})}{c_{11} + c_{12}} ,$$

$$H_R = c_{44} , \tag{1.32}$$

$$H_B = \frac{c_{11} + c_{12}}{3} .$$

For years, the values of Young's modulus have been given in English units of psi. Modern usage is metric, in which the values are given in dyne cm^{-2} or pascals (Pa). In fact, the units are usually GPa (gigapascals). To convert from psi to Pa, multiply the magnitude in psi by 6.9×10^4; to do the reverse, multiply by the inverse, 1.45×10^{-5}.

1.2.13 Permittivity (Dielectric Constant)

The permittivity is the coefficient that relates the displacement field to the electric field. It is a function of temperature and frequency and is related, in general, to the square of the refractive index of a material. Data given in this chapter are generally for the microwave region, since the interest is normally in how applicable a material is for both infrared and microwave use.

1.3 REFRACTIVE MATERIALS

Many important properties need to be considered in the application of infrared optical materials. Their relative importance depends on the application. For lenses and windows, there can be no doubt that the degree and spectral regions of transparency are uppermost. The refractive index is more important for lenses, and the various thermal and mechanical ones come next.

1.3.1 Transparency

Figure 1.1 is a bar graph of the regions over which different materials have transparency. Such a chart serves only as a guide in that transparency is a relative term and often depends on the thickness and the antireflection coatings. Samples of spectral curves of external transmittance are given in Figs. 1.2 through 1.23 (Ref. 1).

The transmission, unless otherwise specified, is *external transmittance,* the measured value of a sample, including reflection losses.

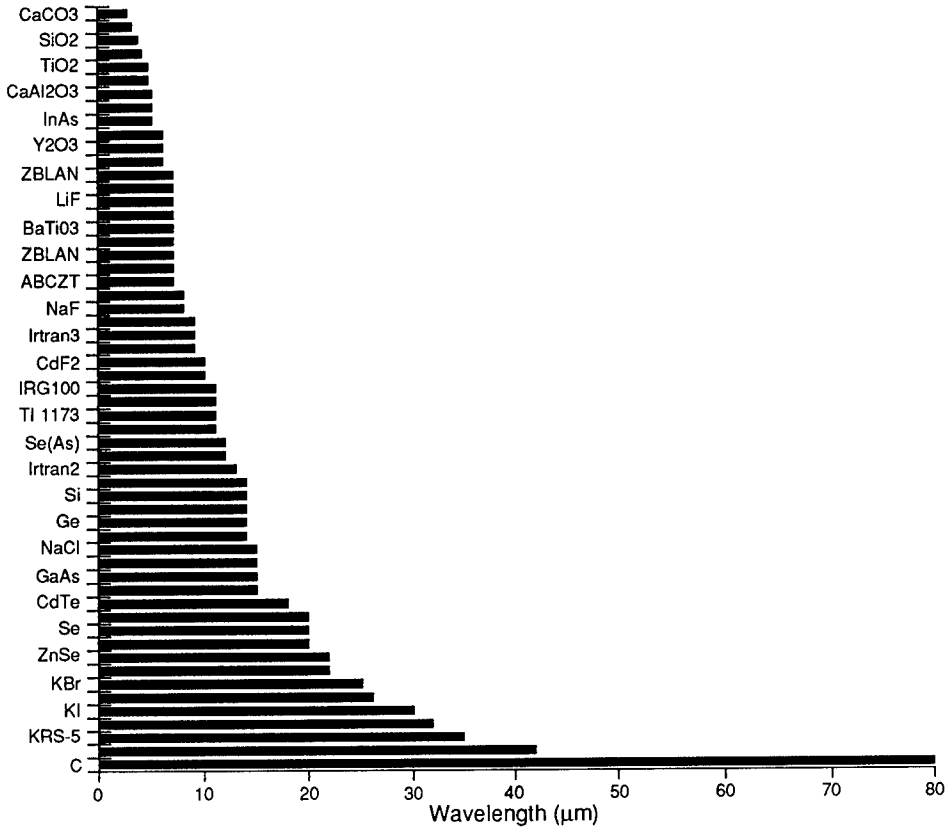


Fig. 1.1 Transparency regions of infrared optical materials.

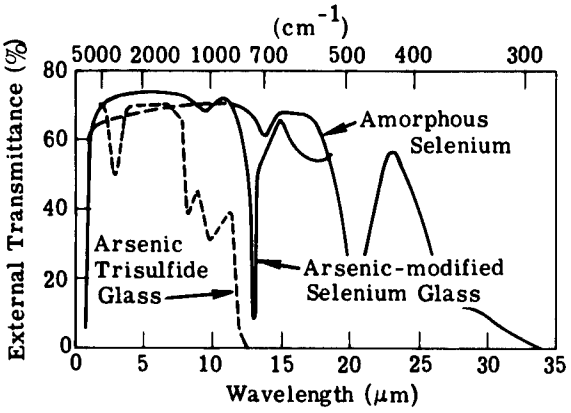


Fig. 1.2 The transmission of amorphous selenium, arsenic-modified selenium, and arsenic trisulfide. The properties of glass vary from batch to batch; these curves can be regarded as typical.

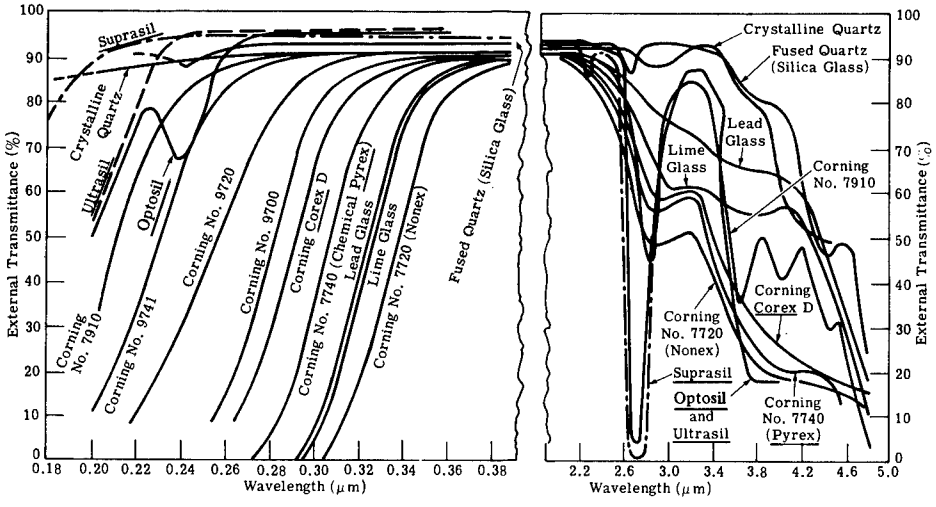


Fig. 1.3 The transmission of several samples of fused quartz and quartz glasses. Pyrex® and Corex® are registered trademarks of Corning Glass Works. Suprasil®, Optosil®, and Ultrasil® are registered trademarks of Heraeus-Amersil, Inc.

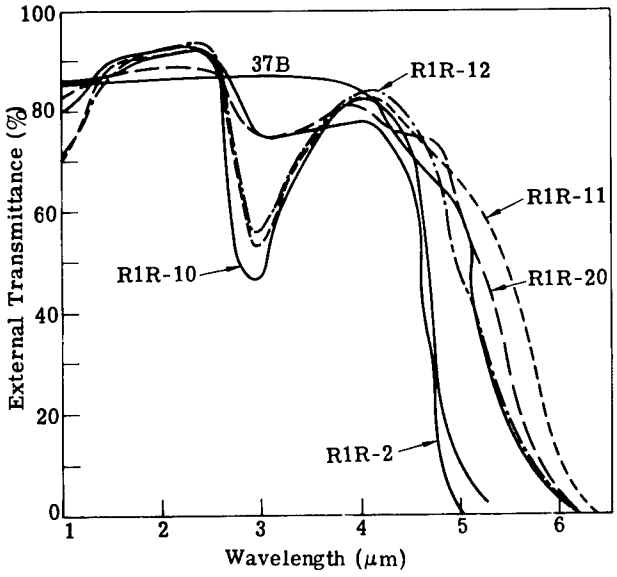


Fig. 1.4 The transmission of several calcium aluminate glasses (2.0 mm thick).

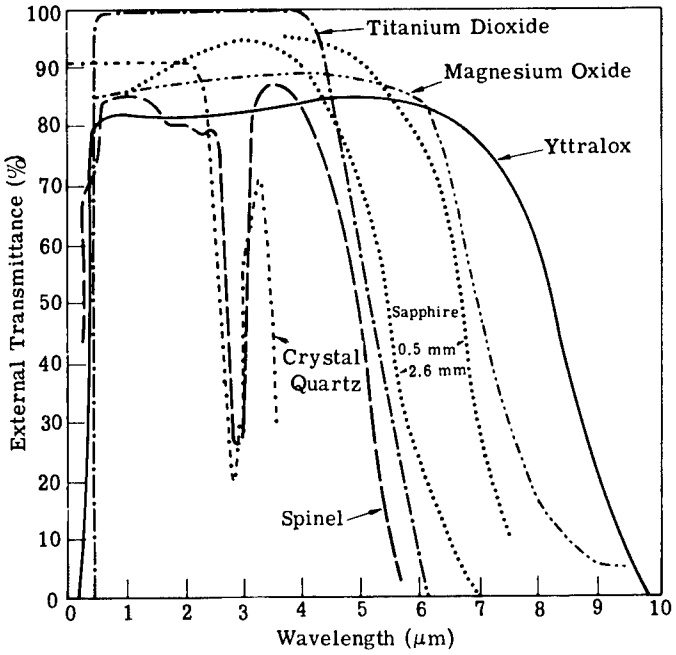


Fig. 1.5 The transmission of sapphire, spinel, titanium dioxide, crystal quartz (for the ordinary ray), Yttralox, and magnesium oxide.

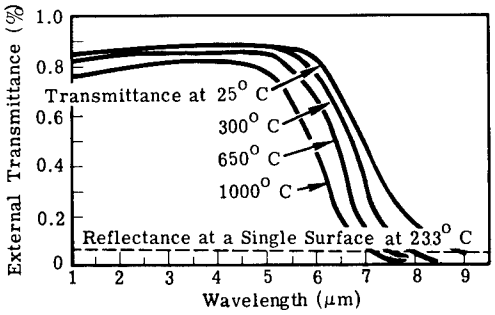


Fig. 1.6 The transmission of magnesium oxide at several temperatures (5.5 mm thick).

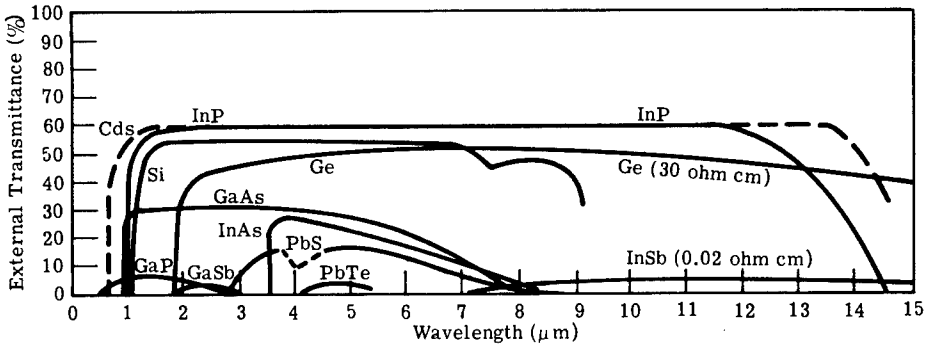


Fig. 1.7 The transmission of cadmium sulfide, indium phosphide, silicon, germanium, gallium arsenide, gallium phosphide, gallium antimonide, indium arsenide, indium antimonide, lead telluride, and lead sulfide.

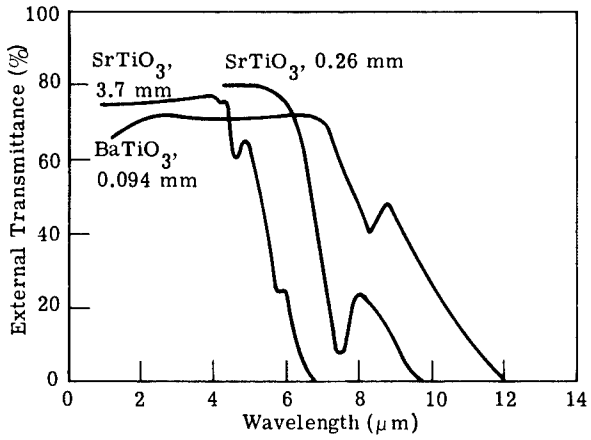


Fig. 1.8 The transmission of barium titanate and strontium titanate.

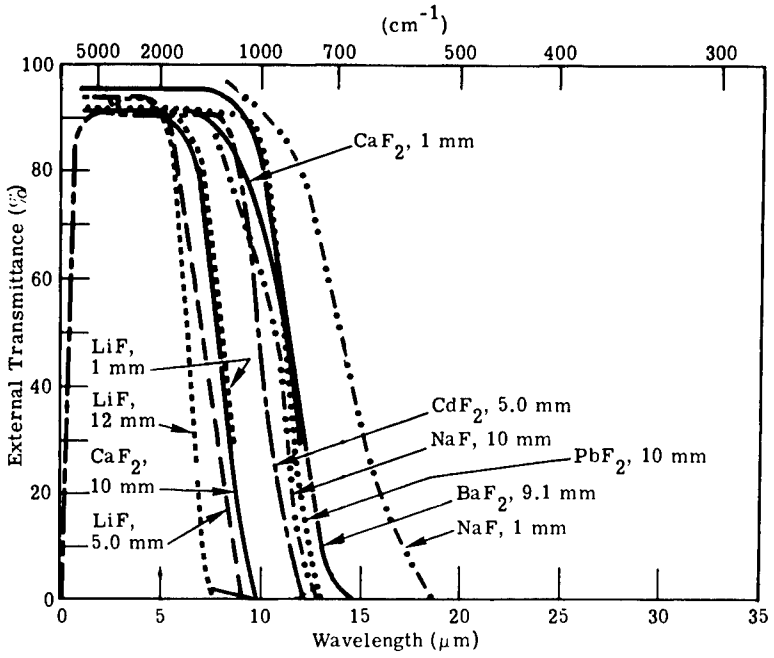


Fig. 1.9 The transmission of barium fluoride, cadmium fluoride, lithium fluoride, calcium fluoride, lead fluoride, and sodium fluoride.

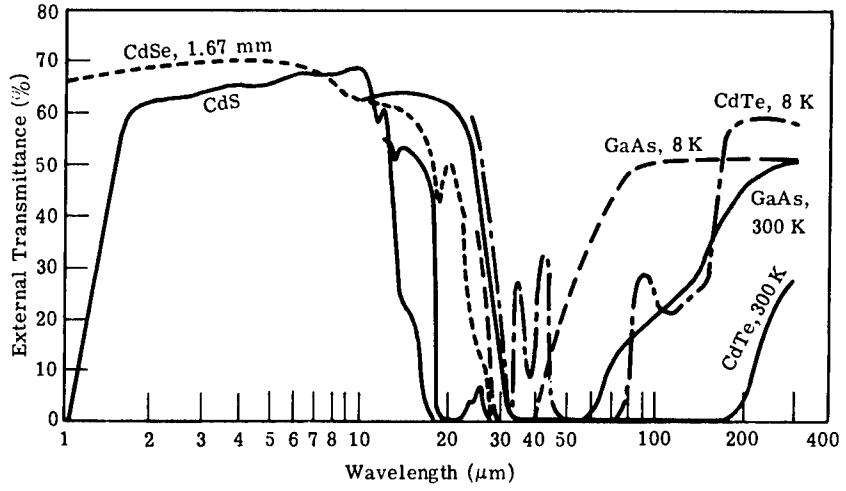


Fig. 1.10 The transmission of cadmium sulfide, cadmium telluride, gallium arsenide, and cadmium selenide.

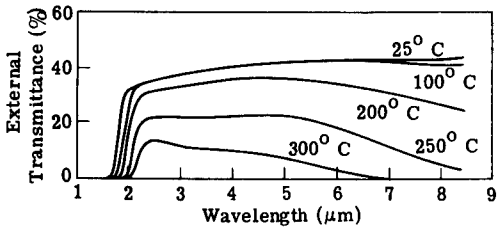


Fig. 1.11 The transmission of germanium for several temperatures (1.15 mm thick).

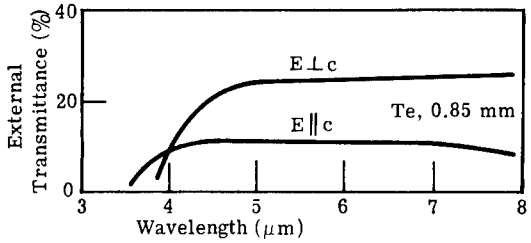


Fig. 1.12 The transmission of tellurium for two polarizations (0.85 mm thick).

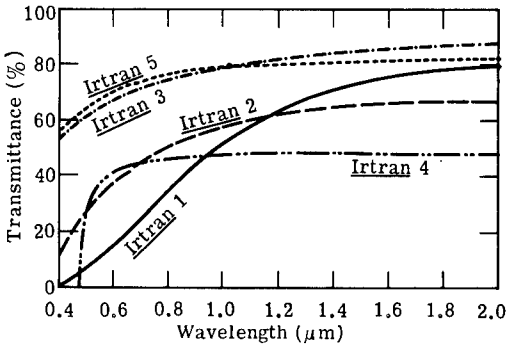


Fig. 1.13 The transmissions of Irtran 1 through 5 (2.0 mm thick). Irtran® is a registered trademark of the Eastman Kodak Company.

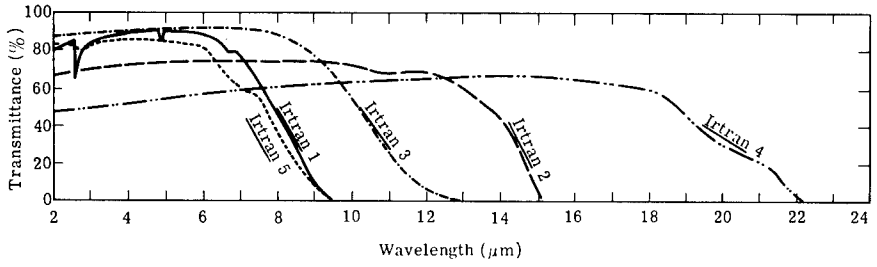


Fig. 1.14 The transmission of Irtran materials (2.0 mm thick). Irtran® is a registered trademark of the Eastman Kodak Company.

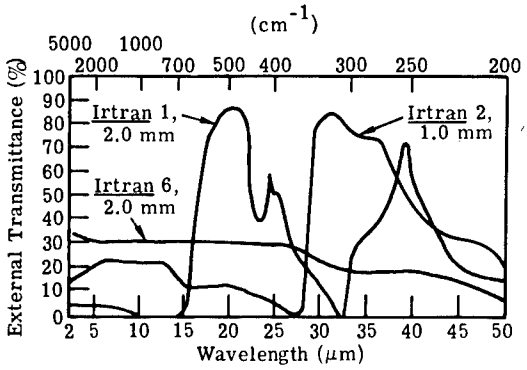


Fig. 1.15 The reflection of Irtran 1, Irtran 2, and Irtran 6. Irtran® is a registered trademark of Eastman Kodak Company.

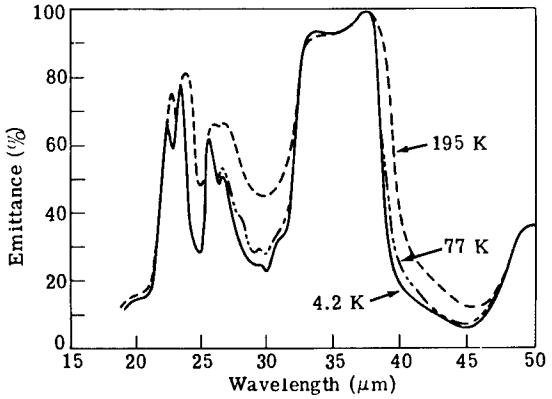


Fig. 1.16 The spectral emission of Irtran 4 at 4.2, 77, and 195 K (0.37 mm thick). Irtran® is a registered trademark of the Eastman Kodak Company.

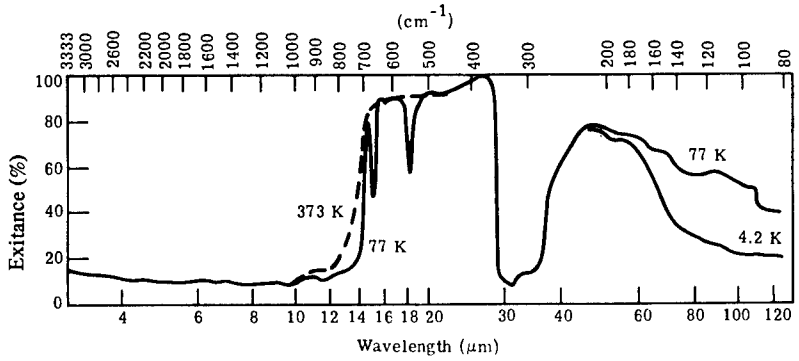


Fig. 1.17 The spectral emission of Irtran 2 at 4.2, 77, and 373 K (2.0 mm thick). Irtran® is a registered trademark of the Eastman Kodak Company.

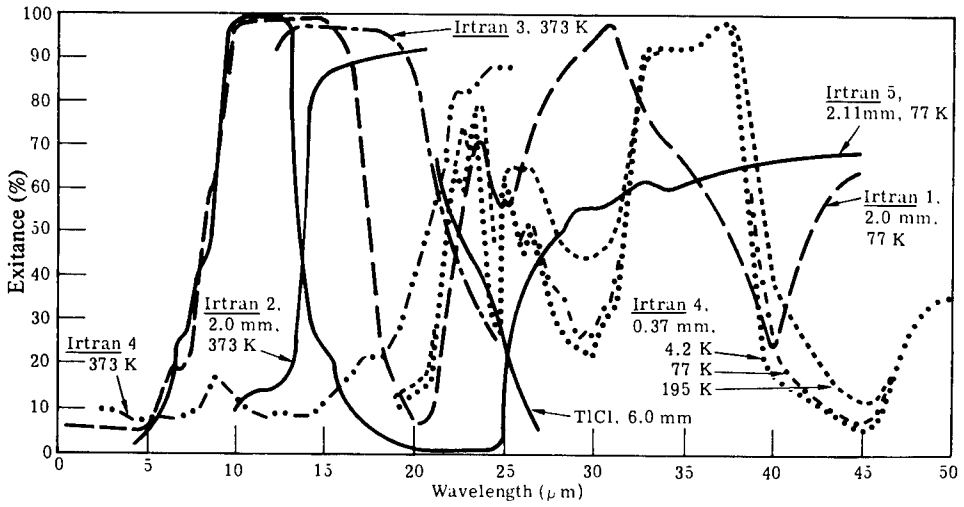


Fig. 1.18 The spectral emission of Irtran 1 at 77 K; Irtran 2 at 373 K; Irtran 3 at 373 K; Irtran 4 at 4.2, 77, 195, and 373 K; Irtran 5 at 77 K; and thallium chloride. Irtran® is a registered trademark of the Eastman Kodak Company.

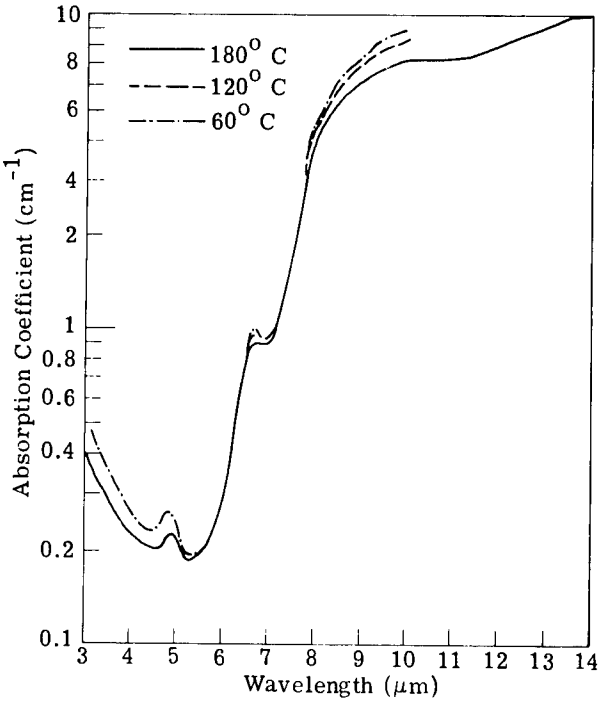


Fig. 1.19 The absorption coefficient of Irtran 1. Irtran® is a registered trademark of the Eastman Kodak Company.

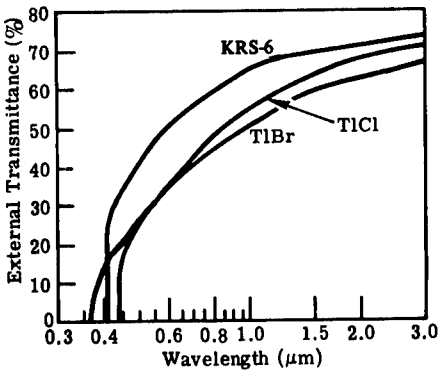


Fig. 1.20 The transmission of thallium bromide, thallium bromide-chlorine, and thallium chloride (1.65 mm thick). Thallium bromide can be ground a very small amount at a time without cracking or chipping. It bends like lead and is only slightly soluble in water.

1.3.2 Refractive Index

Values of the real part of the refractive index are measured most accurately by a minimum deviation technique. The values vary with wavelength and with temperature, and indeed the temperature variation changes with wavelength. Data are usually taken at well-defined spectral lines, and then dispersion equations are fit to the data so interpolations can be made. It is difficult to measure with an accuracy better than about one part in the fourth decimal, although some data are more accurate than that. The data or the equations and the references are reported here.

1.3.2.1 Alkali Halides

Li prepared an excellent review[2] of the alkali halides LiF, NaCl, KCl, KBr, KI, CsBr, and CsI. Table 1.2, adapted from that report, indicates the situation with respect to available refractive index data as of 1975. Some 100 documents were reviewed that included 283 data sets. They used the words *data sets* to indicate a group of values, for instance, index versus wavelength, presented by a particular investigator on a particular material. This table indicates, for instance, that there are many data sets (47) on the refractive index of lithium fluoride and few (3) on its temperature variation. It is also very safe to assume (unfortunately) that most of the data are for the visible spectrum. Other reviews include Ballard, McCarthy, and Wolfe[3] and Wolfe, Ballard, and McCarthy.[4]

Of these materials, it is clear that only the following may have sufficient data for the purposes of lens design: LiF, NaCl, KCl, KBr, KI, CsBr, and CsI. Therefore, they are considered further here. The report[1] has a nice theoretical development, leading to the calculations of n and dn/dT, but only measured values are used here.

Lithium Fluoride. The data of Tilton and Plyler[5] and Harting[6] are considered by Li to be the most reliable. These data and those of Gyulai[7] are useful.

The recommended dispersion and dn/dT equations are given here in a general form, and in a format somewhat different from that of Li, but the equations are the same:

$$n^2 = A_0 + \sum_{i=1}^{n} \frac{A_i \lambda^2}{\lambda^2 - \lambda_i^2} , \tag{1.33}$$

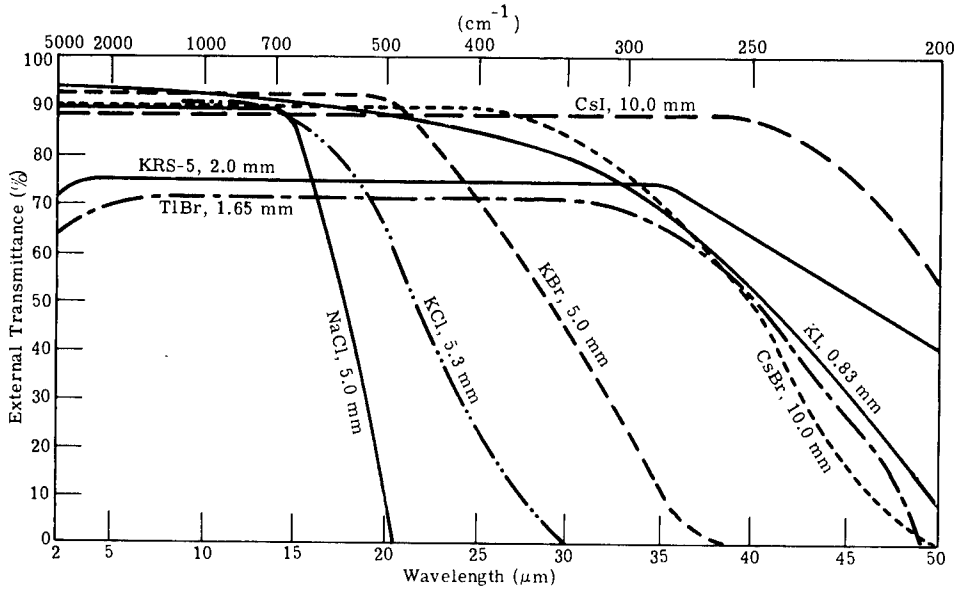


Fig. 1.21 The transmission of cesium iodide, potassium iodide, potassium bromide, thallium bromide, KRS-5, cesium bromide, sodium chloride, and potassium chloride.

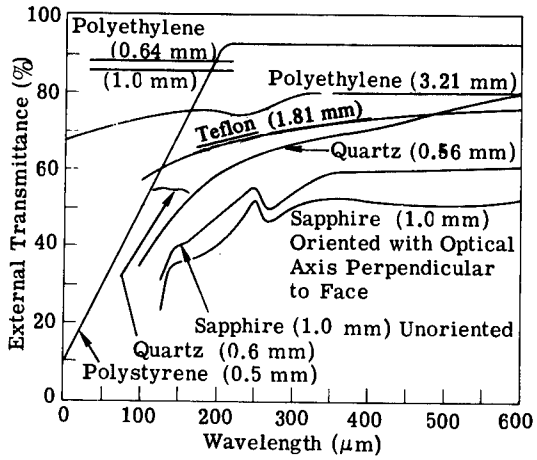


Fig. 1.22 The transmission of several long wave pass materials: quartz, sapphire, polystyrene, and polyethylene samples. Teflon® is a registered trademark of the Dupont Corporation.

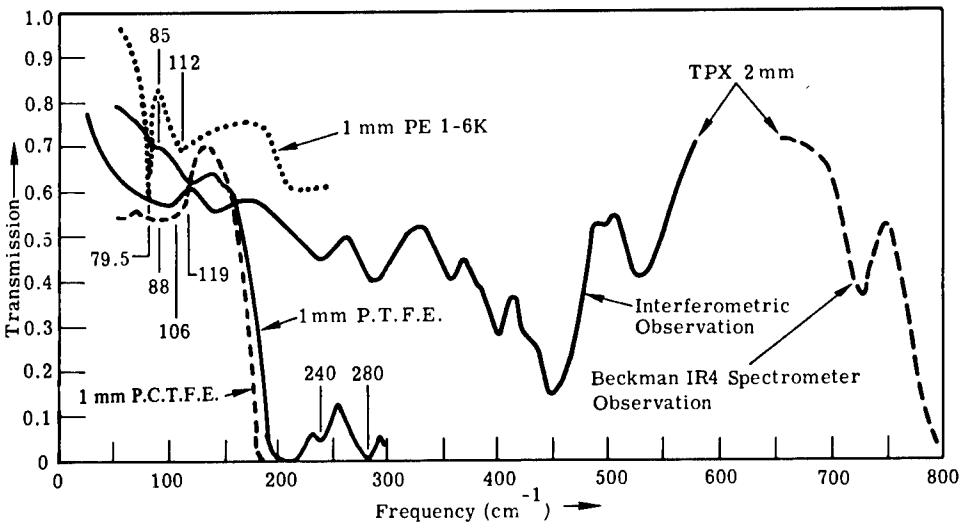


Fig. 1.23 The transmission of PE, PCTFE, and PTFE and the far-infrared absorption spectrum of TPX.

$$2n\frac{dn}{dT} = B_0 + B_1(n^2 - 1) + \sum_{i=2}^{n} \frac{B_i\lambda^4}{(\lambda^2 - \lambda_i^2)^2} \cdot \tag{1.34}$$

The constants for lithium fluoride are given in Tables 1.3 and 1.4. The equation is said to be valid within 50 K of room temperature. The estimated uncertainties in n and dn/dT are listed in Table 1.5.

The recommended equation of Li is compared in Table 1.6 with the Herzberger equation[8] for LiF and with the data given by several investigators. In Table 1.6 the first column is the wavelength of interest; the second is a calculation based on Li's dispersion equation; the third is from Herzberger's equation; the fourth is the data reported by various investigators; the fifth is the difference between the Li and Herzberger equations in units of 10^{-4}; and the sixth column is the difference between Li's calculation and the data. The first group of data in rows 1 through 9 is from Gyulai; the second group from Harting; the third from Tilton and Plyler; and the last from Hohls.[9] The agreement is very good in the middle infrared with the data of Tilton and Plyler. It never exceeds 3 in the fourth decimal place, except for one point that must have been a mistake. The four sets of data were taken at 293, 293, 296.6, and 291 K, but the temperature coefficient is $0.5 \times 10^{-4}\,C^{-1}$ at most, which makes a change of 2.8 in the fourth decimal place for the data adjustment between Tilton and Plyler and Hohls. It does not explain the discrepancy.

Lithium Bromide. Only two refractive index measurements are available for lithium bromide, one at the sodium D line on a crystalline specimen, using the immersion method, and one at the mercury green line on a liquid specimen. Li obtained refractive index values from 1 to 20 μm with an uncertainty as low as 0.008. The values are not cited here.

Table 1.2 Alkali Halide Status (adapted from Ref. 2)

	Index			*dn/dT*			
	Number	Δλ	Quality	Number	Δλ	Δ*T*	Quality
LiF	47	Wide	Good	3	Fair		Poor
LiCl	2	2	Poor				
LiBr	2	2	Poor				
LiI	1	1	Poor				
NaF	15	Wide	Good	2	Fair		Poor
NaCl	49	Wide	Good	10	Fair		Fair
NaBr	5	0.2–0.7	Fair				
NaI	1	1	Fair				
KF	5	0.2–0.6	Fair				
KCl	38	Wide	Good	8	Fair		Fair
KBr	27	Wide	Good	5	<1 μm		Poor
KI	21	Wide	Fair	2	<1 μm		Poor
RbF	1	1	Poor				
RbCl	3	0.2–0.7	Fair	1	1		Bad
RbBr	2	0.2–0.7	Fair				
RbI	3	0.2–0.7	Fair				
CsF	1	1	Poor				
CsCl	6	0.2–0.7	Fair				
CsBr	8	Wide	Good	1	Wide		Average
CsI	13	Wide	Good	1	Wide		Fair

Table 1.3 Constants for the Dispersion Equation for LiF

i	*A*	λ
0	1	
1	0.92549	0.07376
2	6.96747	32.79

Table 1.4 Constants for the *dn/dT* Equation for LiF

i	*B*	λ
0	−8.13	
1	−9.96	
2	12.09	0.00544
3	184.86	1075.18

Table 1.5 Estimated Uncertainties in *n* and *dn/dT* for LiF

Wavelength	Index	*dn/dT*
0.10– 0.15	0.01	0.9
0.15– 0.25	0.001	0.2
0.25– 0.35	0.0005	0.3
0.35– 3.00	0.002	0.3
3.00– 5.00	0.0005	0.3
5.00– 7.00	0.001	0.3
7.00–11.0	0.006	0.3

Table 1.6 Refractive Index Values for LiF

Wavelength	Index			Difference ($\times 10^{-4}$)	
	Li	Herz	Data	Herz	Li
0.19350	1.44312	2.03760	1.44500	−5944.8	18.8
0.19900	1.43966	1.84698	1.44130	−4073.1	16.4
0.20260	1.43758	1.76604	1.43900	−3284.6	14.2
0.20630	1.43557	1.70428	1.43670	−2687.1	11.3
0.21000	1.43368	1.65723	1.43460	−2235.5	9.2
0.21440	1.43158	1.61447	1.43190	−1828.8	3.2
0.21940	1.42937	1.57781	1.43000	−1484.3	6.3
0.22650	1.42652	1.54007	1.42680	−1135.6	2.8
0.23100	1.42486	1.52205	1.42440	−971.9	−4.6
0.25400	1.41785	1.46739	1.41792	−495.4	0.7
0.28000	1.41209	1.43920	1.41188	−271.2	−2.1
0.30200	1.40841	1.42608	1.40818	−176.7	−2.3
0.36600	1.40136	1.40790	1.40121	−65.5	−1.5
0.48610	1.39491	1.39663	1.39480	−17.2	−1.1
0.50000	1.39444	1.39596	1.39430	−15.2	−1.4
0.80000	1.38898	1.38918	1.38896	−1.9	−0.2
1.00000	1.38711	1.38719	1.38711	−0.8	0.0
1.50000	1.38316	1.38321	1.38320	−0.6	0.4
2.00000	1.37867	1.37876	1.37875	−0.9	0.8
2.50000	1.37316	1.37328	1.37327	−1.2	1.1
3.00000	1.36647	1.36662	1.36660	−1.5	1.3
3.50000	1.35853	1.35869	1.35868	−1.6	1.5
4.00000	1.34928	1.34942	1.34942	−1.5	1.4
4.50000	1.33865	1.33875	1.33387	−0.9	−47.8
5.00000	1.32659	1.32659	1.32661	−0.0	0.2
5.50000	1.31299	1.31286	1.31287	1.3	−1.2
6.00000	1.29779	1.29748	1.29745	3.1	−3.4
6.91000	1.26562	1.26491	1.26000	7.1	−56.2
7.53000	1.24003	1.23907	1.23900	9.7	−10.3
8.05000	1.21602	1.21491	1.21500	11.1	−10.2
8.60000	1.18780	1.18673	1.19000	10.8	22.0
9.18000	1.15456	1.15387	1.15500	6.9	4.4
9.79000	1.11520	1.11559	1.10900	−3.9	−62.0

Lithium Chloride. Lithium chloride is like lithium bromide, not very interesting, with two measurements, and with the same uncertainties in index.

Lithium Iodide. Lithium iodide is like lithium bromide, not very interesting, with two measurements, and with the same uncertainties in index.

Sodium Fluoride. The main data are from Harting[6] and Hohls.[9] The uncertainties are about the same as for lithium fluoride. It is true for sodium fluoride as well that the only data on index at temperatures below 290 K are for the spectral regions in which it is opaque.

Sodium Chloride. This material has been measured extensively, partly because of its ready availability and its wide spectral range of transmission, in spite of its inclination to be attacked by water. It has also served well as the disperser in many spectrometers. Li chooses the data of Martens,[10] Paschen,[11] Hohls,[9] Harting,[6] Rubens and Nichols,[12] and Rubens and Trowbridge[13] as the

most representative and accurate, partly because they are consistent. Ballard, McCarthy, and Wolfe[3] cite the data of Coblentz[14]; Wolfe and Ballard[4] cite Hohls and Coblentz and refer to others.

The appropriate equations, given by Li, are of the same form as earlier, and the constants are given in Tables 1.7 and 1.8. Table 1.9 shows the uncertainties. Although there are 53 references cited by Li, only those cited here are for prismatic samples, indicating enough accuracy, and only those above are for the spectral regions of interest. The data obtained for a calculation of dn/dT were obtained either above room temperature or, at the least, −26°C (247 K).

The recommended constants for the spectral and temperature variations of the refractive index are presented here. They compare favorably to those of Coblentz's data, which show agreement to better than 2 parts in the fourth—except at the highest wavelengths. The comparisons are shown in Table 1.10. Interestingly enough, Li's calculations seem to be always a little low.

Sodium Bromide. Sodium bromide is not a good candidate for infrared optical systems.

Potassium Fluoride. Potassium fluoride is not a suitable material for infrared instrumentation because of the difficulty associated with growing it and its extreme hygroscopicity. Several index measurements have been made, three at the sodium D line, one on the melt, and one from 0.21 to 0.58 μm.

Table 1.7 Constants for the Dispersion Equation for NaCl

i	A	λ
1	0.198	0.050
2	0.48398	0.100
3	0.38696	0.128
4	0.25998	0.158
5	0.08796	40.50
6	3.17064	60.98
7	0.30038	120.34

Table 1.8 Constants for the *dn/dT* Equation for NaCl

i	B	λ
0	−0.50	
1	−11.91	
2	6.118	0.02496
3	199.36	3718.56

Table 1.9 Uncertaintities in *n* and *dn/dT* for NaCl

Wavelength	Index	dn/dT
0.20– 0.25	0.006	0.8
0.25– 0.35	0.0005	
0.25– 4.00		0.2
4.00–15.00		0.4
0.35–10.0	0.0001	
10.00–15.00	0.0003	
15.00–25.00	0.0003	0.6
25.00–30.00	0.02	0.9

Table 1.10 Refractive Index Values for NaCl

Wavelength	Li	Coblentz	Difference ($\times 10^{-4}$)
1.0084	1.531895	1.53206	− 1.65
1.054	1.531356	1.53153	− 1.74
1.081	1.531067	1.53123	− 1.63
1.1058	1.530819	1.53098	− 1.61
1.142	1.530483	1.53063	− 1.47
1.1786	1.530171	1.53031	− 1.39
1.2016	1.529989	1.53014	− 1.51
1.2604	1.529561	1.52971	− 1.49
1.3126	1.529223	1.52937	− 1.47
1.4874	1.528298	1.52845	− 1.52
1.5552	1.528002	1.52815	− 1.48
1.6368	1.527681	1.52781	− 1.29
1.6848	1.527506	1.52764	− 1.34
1.767	1.527227	1.52736	− 1.33
2.0736	1.526342	1.52649	− 1.48
2.1824	1.526066	1.52621	− 1.44
2.2464	1.525909	1.52606	− 1.51
2.356	1.525647	1.52579	− 1.43
2.6505	1.524971	1.52512	− 1.49
2.9466	1.524302	1.52466	− 3.58
3.2736	1.52355	1.52371	− 1.60
3.5359	1.522923	1.52312	− 1.97
3.6288	1.522694	1.52286	− 1.66
3.8192	1.522215	1.52238	− 1.65
4.123	1.521415	1.52156	− 1.45
4.712	1.519729	1.51979	− 0.61
5.0092	1.518804	1.51883	− 0.26
5.3009	1.517847	1.5179	− 0.53
5.8932	1.515743	1.51593	− 1.87
6.4825	1.513432	1.51347	− 0.39
6.8	1.512093	1.512	0.93
7.0718	1.510895	1.51093	− 0.35
7.22	1.510221	1.5102	0.21
7.59	1.508473	1.5085	− 0.27
7.6611	1.508127	1.50822	− 0.93
7.9558	1.506654	1.50665	0.04
8.04	1.506222	1.5064	− 1.78
8.8398	1.50187	1.50192	− 0.50
9	1.500944	1.501	− 0.56
9.5	1.497932	1.4998	−18.68
10.0184	1.494613	1.49462	− 0.07
11.7864	1.481724	1.48171	0.14
12.5	1.475797	1.47568	1.17
12.965	1.471699	1.4716	0.99
13.5	1.466745	1.4666	1.45
14.1436	1.460436	1.46044	− 0.04
14.733	1.45431	1.45427	0.40
15.3223	1.447838	1.44743	4.08
15.9116	1.441005	1.4409	1.06
17.93	1.414641	1.4149	− 2.59
20.57	1.372149	1.3735	−13.51
22.3	1.338313	1.3403	−19.87

Table 1.11 Constants for the Dispersion Equation for KBr

i	A	λ
1	1.26486	0.100
2	0.30523	0.131
3	0.41620	0.162
4	0.18870	70.42

Table 1.12 Constants for the *dn/dT* Equation for KBr

i	B	λ
0	0.19	
1	−11.13	
2	3.393	0.02624
3	142.56	4958.98

Table 1.13 Uncertainties in *n* and *dn/dT* for KBr

Wavelength	Index	*dn/dT*
0.2 – 0.25	0.006	0.9
0.25– 0.35	0.0005	0.3
0.35– 0.4	0.0002	0.3
0.25– 4.0		0.3
4.0 –30.0		0.5
0.40–20.0	0.0001	
20.0 –26.0	0.0005	
26.0 –35.0	0.006	0.9
35.0 –42.0	0.008	0.9

Potassium Bromide. Potassium bromide is one of the good long-wavelength materials, but it is susceptible to water and chemical attack. This substance has been measured by some 27 investigators, but only the measurements of Spindler and Rodney,[15] Stephens et al.,[16] Forrest,[17] Harting,[6] and Gundelach[18] are useful for the index. In addition, only five sets of data on the temperature coefficient exist, all from 0.26 to 1.1 μm and all of these above 290 K. The constants for the dispersion equations are given in Table 1.11; those for *dn/dT* in Table 1.12; and the uncertainties are listed in Table 1.13. The data are compared with the dispersion equations of Li and of Stephens. The coefficient Stephens quotes for the wavelength squared term is an order of magnitude too high. Table 1.14 shows comparisons after this correction has been made.

Potassium Chloride. Although potassium chloride is not used frequently because of its poor mechanical and chemical characteristics, it is useful on occasion. Of the 41 data sets available, only 5 are of use in evaluating its refractive index in the infrared: those of Martens,[10] Paschen,[11] Hohls,[9] Harting,[6] and Rubens and Nichols,[12] which are the same good references for much of the other data. Values of *dn/dT* have been measured by four different investigators, Harting,[6] Liebreich,[19] and Koslovskii and Ustimenko,[20] in the temperature range from 223 to 308 K with a CO_2 laser. The dispersion equations given by Li are only for the ultraviolet and visible regions. The constants are given in Table 1.15; those for *dn/dT* in Table 1.16 and uncertainties in Table 1.17. Table 1.18 compares Li's equation with the data presented in Wolfe and Ballard.[4] In the 1- to 12-μm region, the fit is never worse than 3.5 in the fourth,

Table 1.14 Refractive Index Values for KBr

Wavelength	Index			Difference ($\times 10^{-4}$)	
	Li	Stephens	Data	Li	Stephens
1.014	1.54418	1.54408	1.54408	1.02	0.03
1.129	1.54270	1.54258	1.54258	1.21	0.03
1.367	1.54072	1.54057	1.54061	1.07	−0.36
1.701	1.53917	1.53901	1.53901	1.61	−0.02
2.440	1.53752	1.53734	1.53733	1.85	0.05
2.730	1.53711	1.53692	1.53693	1.77	−0.05
3.419	1.53631	1.53612	1.53612	1.85	0.03
4.258	1.53541	1.53523	1.53523	1.82	0.04
6.238	1.53303	1.53288	1.53288	1.52	−0.03
6.692	1.53239	1.53225	1.53225	1.43	−0.04
8.662	1.52915	1.52903	1.52903	1.15	0.00
9.724	1.52705	1.52696	1.52695	1.02	0.05
11.035	1.52412	1.52405	1.52404	0.80	0.05
11.862	1.52207	1.52200	1.52200	0.65	0.05
14.290	1.51508	1.51505	1.51505	0.28	0.03
14.980	1.51282	1.51280	1.51280	0.22	0.04
17.400	1.50390	1.50390	1.50390	−0.02	0.01
18.100	1.50101	1.50102	1.50076	2.51	2.58
19.010	1.49704	1.49705	1.49703	0.10	0.22
19.910	1.49286	1.49288	1.49288	−0.15	0.00
21.180	1.48653	1.48655	1.48655	−0.19	0.05
21.830	1.48308	1.48311	1.48311	−0.29	0.01
23.860	1.47134	1.47140	1.47140	−0.60	0.05
25.140	1.46313	1.46324	1.46324	−1.11	0.01

Table 1.15 Constants for the Dispersion Equation for KCl

i	A	λ
1	1.26486	0.100
2	0.30523	0.131
3	0.41620	0.162
4	0.18870	70.42

Table 1.16 Constants for the *dn/dT* Equation for KCl

i	B	λ
0	0.19	
1	−11.13	
2	3.393	0.02624
3	142.56	4958.98

Table 1.17 Uncertainties in *n* and *dn/dT* for KCl

Wavelength	Index	*dn/dT*
0.18– 0.20	0.01	0.9
0.20– 0.24	0.005	0.3
0.24– 0.35	0.0005	0.3
0.20– 4.00		0.3
0.35–10.0	0.0001	
10.0 –15.0	0.0002	0.5
15.0 –21.0	0.0005	0.9
21.0 –30.0	0.006	0.9
30.0 –35.0	0.008	0.9

Table 1.18 Refractive Index Values for KCl

Wavelength	Li	Data	Difference ($\times 10^{-4}$)
1.1786	1.478279	1.478311	− 0.32
1.768	1.475877	1.47589	− 0.13
2.3573	1.474715	1.474751	− 0.36
2.9466	1.473819	1.473834	− 0.15
3.5359	1.47295	1.473049	− 0.99
4.7146	1.471012	1.471122	− 1.10
5.3039	1.469892	1.470013	− 1.21
5.8932	1.468655	1.468804	− 1.49
8.2505	1.462446	1.462726	− 2.80
8.8398	1.460561	1.460858	− 2.97
10.0184	1.456371	1.45672	− 3.49
11.786	1.448997	1.44919	− 1.93
12.965	1.443314	1.44346	− 1.46
14.141	1.437003	1.43722	− 2.17
15.912	1.426213	1.42617	0.43
17.68	1.413782	1.41403	− 2.48
18.2	1.409788	1.409	7.88
18.8	1.40498	1.401	39.80
19.7	1.397347	1.398	− 6.53
20.4	1.391047	1.389	20.47

but it becomes 2 in the third at 20 μm where this material is more interesting, but has higher dispersion.

Potassium Iodide. Data measured by Gyulai,[7] Harting,[6] and Korth[21] are useful for estimating the index and for the temperature coefficient, but only around room temperature. Estimated uncertainties range from 0.002 to 0.009 for the index and 0.3 to 0.9 for the temperature coefficient.

Rubidium Halides. All of the rubidium halides are very hygroscopic and soft. They therefore take a poor polish, and are just not well suited to instrumentation, although they have good infrared transmission.

Cesium Fluoride. Data exist for the index of cesium fluoride at only one wavelength, and for both alpha and beta versions of the crystal. Nevertheless, Li estimates the uncertainty in index to be as good as 0.003 and that of dn/dT to be as good as 0.5, based on other physical data.

Cesium Chloride. Cesium chloride is not a good optical material. Few measurements have been made, and the uncertainties are as good as 0.001 and 0.4 for n and dn/dT, respectively.

Cesium Bromide. The earliest measurements were made just four years after the infrared region was discovered. The most meaningful data are those of Rodney and Spindler,[15] who reported two sets of measurements. The constants for the usual equations are given in Tables 1.19 and 1.20. The data uncertainties are listed in Table 1.21, but the temperature coefficient was measured only at about room temperature. The data of Rodney and Spindler are compared with their dispersion equation and that of Li. The results are shown in Table 1.22.

Table 1.19 Constants for the Dispersion Equation of CsBr

i	A_i	λ_i
0	1.14600	
1	1.26628	0.120
2	0.01137	0.146
3	0.00975	0.160
4	0.00672	0.173
5	0.34557	0.187
6	3.76339	136.05

Table 1.20 Constants for the *dn/dT* Equation for CsBr

i	B_i	λ_i
0	− 4.75	
1	−14.22	
2	2.172	0.03497
3	310.40	18509.6

Table 1.21 Uncertainties in *n* and *dn/dT* for CsBr

Wavelength	Index	*dn/dT*
0.21– 0.25	0.01	1.0+
0.25– 0.35	0.0003	0.4
0.35–30.0	0.0001	0.4
30.0 –40.0	0.0005	0.4
40.0 –55.0	0.006	0.9

Table 1.22 shows that Li differs from Rodney and Spindler by an almost constant 5 in the fourth decimal place.

Cesium Iodide. Rodney is the only investigator to measure in the spectral region of present interest. His data were all taken at about room temperature. The usual tables give the results: Table 1.23 gives the constants for the dispersion equation; Table 1.24 gives the constants for the thermal coefficient of the refractive index; and Table 1.25 gives the uncertainties for both.

1.3.2.2 Semiconductors

Semiconductors such as Ge, Si, GaAs, GaSb, CdS, Se, ZnS, and ZnSe play an important part in infrared technology, both as detectors and as transparent optical materials. Germanium and silicon have certainly maintained a dominant position in refractive elements, but ZnS and ZnSe are increasing in importance, and the gallium and cadmium compounds show great promise.

Germanium. Germanium has been one of the most used infrared optical materials. It has a spectacularly high refractive index value of about 4 and a concomitant low dispersion. The equivalent Abbe number in the 3- to 5-μm region is 101, and in the 8- to 12-μm region it is 990. The earliest data on its refractive index were reported by Briggs.[22] They were followed by Salzburg and Villa.[23] The latter data are about 5 parts in the third decimal place lower than those of Briggs. They also report that the refractive index of polycrystalline germanium differs from single crystal by a few parts in the fourth

Table 1.22 Refractive Index Values for CsBr

	Index			Difference ($\times 10^{-4}$)	
Wavelength	Li	Rodney	Data	Li	Rodney
1.01	1.678191	1.677659	1.677660	5.31	−0.01
1.13	1.676371	1.675839	1.675840	5.30	−0.01
1.53	1.672902	1.672375	1.672370	5.32	0.05
1.70	1.672095	1.671570	1.671580	5.15	−0.10
3.36	1.669174	1.668657	1.668660	5.14	−0.03
4.26	1.668446	1.667932	1.667940	5.06	−0.08
6.47	1.666708	1.666203	1.665870	8.38	3.33
9.72	1.663338	1.662849	1.662830	5.08	0.19
11.03	1.661632	1.661150	1.661180	4.52	−0.30
14.29	1.656460	1.655989	1.655940	5.20	0.49
14.98	1.655187	1.654717	1.654740	4.47	−0.23
15.48	1.654224	1.653755	1.653750	4.74	0.05
17.40	1.650216	1.649742	1.649670	5.46	0.72
18.16	1.648489	1.648011	1.647950	5.39	0.61
20.57	1.642474	1.641971	1.641840	6.34	1.31
21.80	1.639080	1.638555	1.638460	6.20	0.95
22.76	1.636274	1.635728	1.635650	6.24	0.78
23.86	1.632887	1.632309	1.632340	5.47	−0.31
25.16	1.628642	1.628016	1.628170	4.72	−1.54
25.97	1.625861	1.625199	1.625210	6.51	−0.11
26.63	1.623516	1.622821	1.622840	6.76	−0.19
29.81	1.611191	1.610278	1.610340	8.51	−0.62
30.54	1.608112	1.607134	1.607490	6.22	−3.56
30.91	1.606515	1.605501	1.605910	6.05	−4.09
31.70	1.603021	1.601924	1.601980	10.41	−0.56
33.00	1.597016	1.595762	1.595840	11.76	−0.78
34.48	1.589781	1.588313	1.588350	14.31	−0.37
35.45	1.584799	1.583170	1.582840	19.59	3.30
35.90	1.582422	1.580712	1.580690	17.32	0.22
37.52	1.573508	1.571465	1.571830	16.78	−3.65
39.22	1.563526	1.561065	1.559900	36.26	11.65
2.00	1.671130	1.670606	1.670410	7.20	1.96
3.00	1.669525	1.669006	1.668910	6.15	0.96
4.00	1.668645	1.668129	1.668060	5.85	0.69
6.00	1.667099	1.666592	1.666690	4.09	−0.98
7.00	1.666234	1.665731	1.665750	4.84	−0.19
8.00	1.665268	1.664770	1.664790	4.78	−0.20
10.00	1.662997	1.662509	1.662530	4.67	−0.21
12.00	1.660240	1.659762	1.659770	4.70	−0.08
13.00	1.658673	1.658199	1.658220	4.53	−0.21
14.00	1.656976	1.656505	1.656520	4.56	−0.15
16.00	1.653188	1.652718	1.652760	4.28	−0.42
18.00	1.648859	1.648382	1.648440	4.19	−0.58
20.00	1.643972	1.643477	1.643570	4.02	−0.93
22.00	1.638507	1.637978	1.638120	3.87	−1.42
24.00	1.632443	1.631860	1.632050	3.93	−1.90
26.00	1.625756	1.625093	1.625380	3.76	−2.87
28.00	1.618418	1.617642	1.617680	7.38	−0.38
30.00	1.610399	1.609470	1.609650	7.49	−1.80

Table 1.23 Constants for the Dispersion Equation of CsI

i	A_i	λ_i
0	1.27587	
1	0.68689	0.130
2	0.26090	0.147
3	0.06256	0.163
4	0.06527	0.177
5	0.14991	0.185
6	0.51818	0.206
7	0.01918	0.218
8	3.38229	161.29

Table 1.24 Constants for the *dn/dT* Equation for CsI

i	B_i	λ_i
0	− 5.53	
1	−14.70	
2	2.464	0.04752
3	242.76	26014.46

Table 1.25 Uncertainties in *n* and *dn/dT* for CsI

Wavelength	Index	*dn/dT*
0.25– 0.35	0.0002	0.8
0.35– 1.00	0.0001	0.5
0.35–20.0	0.0001	
1.00–50.0		0.3
20.0 –40.0	0.0005	0.4
50.0 –67.0	0.001	0.9

decimal place. Twenty values are reported between 2 and 16 μm and are an average of 0.7 μm apart. The relative thermal coefficient was reported as $6.7 \pm 0.4 \times 10^{-5}$.

The data of Salzburg and Villa are shown in Table 1.26, along with the dispersion calculations of Li[24] and of Herzberger. It is clear that Li is very high at short wavelengths, and a little low at long ones. Herzberger has ran-

Table 1.26 Refractive Index Values for Germanium

	Index			Difference ($\times 10^{-4}$)	
Wavelength	Li	Herz	Salz	Li-Salz	Herz-Salz
2.0581	4.1239	4.1016	4.1016	223	0
2.1526	4.1133	4.0920	4.0919	214	1
2.3126	4.0981	4.0787	4.0786	195	1
2.4374	4.0882	4.0702	4.0708	174	− 6
2.577	4.0789	4.0622	4.0609	180	13
2.7144	4.0710	4.0557	4.0552	158	5
2.998	4.0580	4.0450	4.0452	128	− 2
3.3033	4.0475	4.0366	4.0369	106	− 3
3.4188	4.0443	4.0340	4.0334	109	6
4.258	4.0280	4.0214	4.0216	64	− 2
4.866	4.0210	4.0161	4.0170	40	− 9
6.238	4.0121	4.0093	4.0094	27	− 1
8.66	4.0055	4.0044	4.0043	12	1
9.72	4.0040	4.0034	4.0034	6	− 0
11.04	4.0027	4.0026	4.0026	1	− 0
12.2	4.0019	4.0022	4.0023	−4	− 1
13.02	4.0015	4.0021	4.0021	−6	0
14.21	4.0010	4.0022	4.0015	−5	7
15.08	4.0007	4.0024	4.0014	−7	10
16	4.0004	4.0028	4.0012	−8	16

Table 1.27 Refractive Index Values for Germanium

	Index			Difference ($\times 10^{-4}$)	
Wavelength	Li	Herz	Wolfe	Li	Herz
2.554	4.08173	4.06346	4.06230	194.25	11.59
2.652	4.07582	4.05852	4.05754	182.76	9.79
2.732	4.07145	4.05490	4.05310	183.50	18.03
2.856	4.06539	4.04993	4.04947	159.16	4.56
2.958	4.06095	4.04632	4.04595	150.03	3.72
3.09	4.05585	4.04221	4.04292	129.29	− 7.13
4.12	4.03129	4.02291	4.02457	67.17	−16.63
5.19	4.01956	4.01397	4.01617	33.86	−22.00
8.23	4.00748	4.00497	4.00743	0.49	−24.62
10.27	4.00462	4.00300	4.00571	−10.90	−27.14
12.36	4.00303	4.00220	4.00527	−22.38	−30.68

domly distributed low and high values, but with rather high differences. Other data were measured by Icenogle, Platt, and Wolfe[25] that differ from those just mentioned. These are compared in Table 1.27 where it can be seen that there is good agreement at the shorter wavelengths, but disparities of about 0.003 at longer ones.

Edwin, Dudermel, and Lamare[26] of the National Physical Laboratories measured 10 different samples of germanium including monocrystalline, polycrystalline, high-resistivity, and low-resistivity samples. They also obtained a comparative measurement from the Institute Optique. The results are compared in Table 1.28. In no case were the differences much greater than 1 in the fourth decimal place.

Silicon. The Li and Herzberger dispersion equations are compared with the data reported by Salzburg and Villa. The equations are the same as for germanium, but with different constants, which are given in Table 1.29. The data sets are from Briggs,[22] Primak,[27] Cardona,[28] Lukes,[29] Salzburg and Villa,[23] and Icenogle, Platt, and Wolfe.[25] Briggs measured only from 1.05 to 2.60 μm. Cardona measured a prism at several temperatures and from 1 to 5 μm, but the data have only been reported in the form of a figure. The same is true for Lukes. From readings of the figure, however, the data of Cardona are seen to be about 40 parts in the fourth decimal place lower than Briggs. Lukes was lower than Salzburg and Villa by 15 but corresponded to Briggs and Icenogle.

Zinc Sulfide. The rather new material of zinc sulfide comes in single crystal, hot-pressed compact, and chemical vapor deposition (CVD) forms. The compact is made by Eastman Kodak, and has been measured by them and by Wolfe. Table 1.30 shows the difference between two samples measured by Wolfe and Korniski.[30] Table 1.31 shows the difference between the measurements of Wolfe and of Eastman Kodak. Figure 1.24 shows the change of index with temperature. It is not linear and a measurement made at room temperature would be misleading for lower temperatures.

Feldman[31] measured the refractive index of ZnS formed by a CVD process in 1978. No other equivalent measurements have been made. Mell[32] measured a greenish sphalerite; DeVore[33] measured a water white sphalerite; Bond[34]

Table 1.28 Comparison of the Refractive Indices of Several Samples of Germanium

Comparison of Two Intrinsic Monocrystalline Samples			
	Resistivities = 0.48 and 0.48		Difference ($\times 10^{-4}$)
8	4.00551	4.00536	1.5
9	4.00423	4.00404	1.9
10	4.00329	4.00311	1.8
12	4.00204	4.00188	1.6
13	4.00157	4.00148	0.9
14	4.00123	4.00112	1.1
Comparison of Intrinsic Polycrystalline and Monocrystalline Samples			
	Resistivities = 0.55 and 0.48		
8	4.00551	4.00551	0
9	4.00423	4.00425	−0.2
10	4.00329	4.00332	−0.3
12	4.00204	4.00206	−0.2
13	4.00157	4.00164	−0.7
14	4.00123	4.00128	−0.5
Comparison of Monocrystalline Samples			
	Resistivities = 0.48 and 0.21		
8	4.00551	4.00545	0.6
9	4.00423	4.00415	0.8
10	4.00329	4.00322	0.7
12	4.00204	4.00195	0.9
13	4.00157	4.00154	0.3
14	4.00123	4.00122	0.1
Comparison of Polycrystalline Samples			
	Resistivities = 0.55 and 0.20		
8	4.00551	4.00545	−0.6
9	4.00425	4.00417	−0.8
10	4.00332	4.00323	−0.9
12	4.00206	4.00195	−1.1
13	4.00164	4.00154	−1.
14	4.00128	4.00119	−0.9
Comparison of Measurements on Sample 1			
8	4.0058	4.0058	−2.9
9	4.0043	4.0043	−0.7
10	4.0032	4.0032	0.9
12	4.0017	4.0017	3.4
13	4.0013	4.0013	2.7
14	4.0011	4.0011	1.3

Table 1.29 Refractive Index Values for Silicon

Wavelength	Index			Difference ($\times 10^{-4}$)	
	Li	Herz	Data	Li	Herz
1.357	3.49957	3.49753	3.49750	20.74	0.28
1.3673	3.49830	3.49624	3.49620	20.97	0.36
1.3951	3.49499	3.49290	3.49290	20.89	0.01
1.5295	3.48142	3.47944	3.47950	19.23	−0.62
1.6606	3.47120	3.46951	3.46960	16.01	−0.89
1.7092	3.46799	3.46643	3.46640	15.89	0.31
1.8131	3.46196	3.46070	3.46080	11.60	−1.04
1.9701	3.45458	3.45376	3.45370	8.82	0.61
2.1526	3.44794	3.44759	3.44760	3.38	−0.06
2.3254	3.44302	3.44308	3.44300	0.24	0.78
2.4373	3.44038	3.44066	3.44080	− 4.18	−1.36
2.7144	3.43518	3.43594	3.43580	− 6.18	1.41
3	3.43125	3.43239	3.43200	− 7.47	3.94
3.3033	3.42814	3.42959	3.42970	−15.59	−1.08
3.4188	3.42717	3.42872	3.42860	−14.32	1.16
3.5	3.42654	3.42815	3.42840	−18.60	−2.49
4	3.42348	3.42539	3.42550	−20.22	−1.11
4.258	3.42230	3.42432	3.42420	−18.98	1.23
4.5	3.42138	3.42348	3.42360	−22.23	−1.19
5	3.41987	3.42210	3.42230	−24.26	−2.02
5.5	3.41876	3.42106	3.42130	−25.39	−2.39
6	3.41791	3.42026	3.42020	−22.86	0.60
6.5	3.41726	3.41963	3.41950	−22.45	1.29
7	3.41673	3.41913	3.41890	−21.68	2.25
7.5	3.41631	3.41872	3.41860	−22.90	1.20
8	3.41596	3.41840	3.41840	−24.35	−0.03
8.5	3.41568	3.41814	3.41820	−25.22	−0.57
10	3.41506	3.41774	3.41790	−28.38	−1.63
10.5	3.41491	3.41771	3.41780	−28.87	−0.88
11.04	3.41477	3.41775	3.41760	−28.26	1.49

measured a natural sample that became opaque at 2.4 μm, and he agreed with Feldman to 2 parts in the third decimal place. The rest of the measurements were in other spectral regions. Table 1.32 shows the agreement between the index values for two different samples and the two different dispersion equations of Feldman and of Li.

Zinc Selenide. Zinc selenide is an exceptional material that has a transmission region from the orange part of the visible to about 24 μm. It has been made recently by CVD. The other form is a hot-pressed compact, originally accomplished by Eastman Kodak and called Irtran 4. Probably the most reliable data on Irtran 4 are in Kodak's sales literature; Hilton and Jones[35] also measured Irtran 4.

References for ZnSe include Marple,[36] Rambauske,[37] Wunderlich and deShazer,[38] Feldman et al.,[31] and Thompson.[39] Because his measurements were made in 1963, Marple must have measured natural or grown crystals, and he measured only from 0.496 to 0.83 μm. Rambauske measured only between 0.4 to 0.644 μm. His two samples were of different purity and gave different results. Wunderlich and deShazer reported only for the visible region. Thompson measured only at two wavelengths. That leaves only the work reported by Feldman and by Wolfe. Table 1.33 shows results for three independent runs

Table 1.30 Difference between Two Irtran 2 Runs

Wavelength (μm)	Difference ($\times 10^{-4}$)
0.6328	2
1	9
1.5	1
2	0
2.5	1
3	− 1
3.5	− 4
4	1
4.5	3
5	4
5.5	2
6	8
6.5	4
7	8
7.5	10
8	5
8.5	1
9	4
9.5	2
10	6
10.5	3
11	3
11.5	3
12	1
12.5	5
13	1
13.5	1
14	1

Table 1.31 Difference between Wolfe and Eastman Kodak for Irtran 2 for the Refractive Index of Irtran 2

Wavelength (μm)	Difference ($\times 10^{-4}$)
1	42
1.5	48
2	41
2.5	45
3	43
3.5	43
4	41
4.5	41
5	39
5.5	37
6	40
6.5	36
7	33
7.5	35
8	34
8.5	37
9	34
9.5	32
10	28

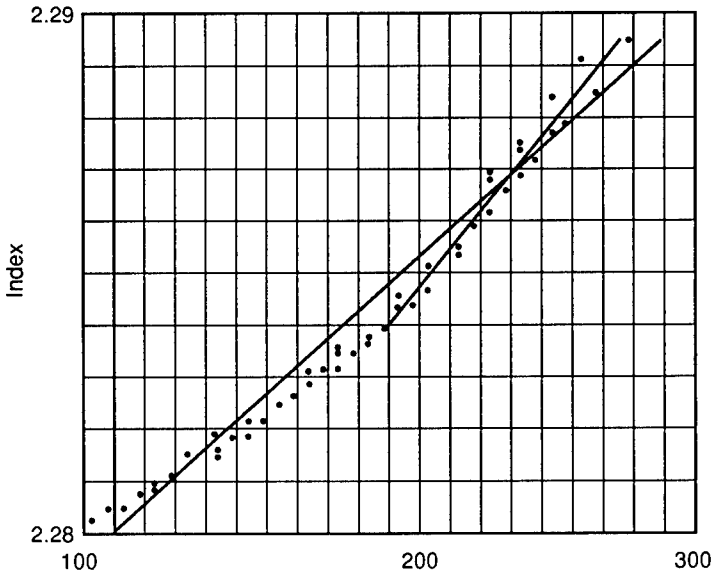


Fig. 1.24 Refractive index of Irtran 2 as a function of temperature.

Table 1.32 Refractive Index Values for ZnS

Wavelength	Index			Difference ($\times 10^{-4}$)	
	Feld1	Feld2	Li	F1 − F2	Li − F1
1.00	2.292262	2.292297	2.292065	−0.35	1.97
2.00	2.264512	2.264529	2.264698	−0.17	− 1.87
3.00	2.257190	2.257187	2.257427	0.03	− 2.37
4.00	2.251815	2.251783	2.252003	0.32	− 1.88
5.00	2.246161	2.246096	2.246261	0.65	− 1.00
6.00	2.239630	2.239531	2.239627	0.99	0.03
7.00	2.231957	2.231826	2.231855	1.31	1.01
8.00	2.222957	2.222805	2.222783	1.52	1.74
9.00	2.212464	2.212306	2.212268	1.58	1.96
10.00	2.200291	2.200156	2.200157	1.35	1.34
11.00	2.186225	2.186154	2.186280	0.71	− 0.55
12.00	2.170005	2.170058	2.170433	−0.54	− 4.28
13.00	2.151311	2.151576	2.152374	−2.64	−10.62
14.00	2.129745	2.130342	2.131806	−5.96	−20.61

Table 1.33 Refractive Index Values for Zinc Selenide

Wave-length	Index			Standard Deviation ($\times 10^{-4}$)
	Run 1	Run 2	Run 3	
3.0	2.43492	2.43538	2.43496	2.55
3.5	2.43253	2.43283	2.43220	3.15
4.5	2.42869	2.42874	2.42886	0.87
4.6	2.42829	2.42821	2.42834	0.66
5.0	2.42690	2.42701	2.42701	0.64
5.4	2.42581	2.42558	2.42570	1.15
5.8	2.42443	2.42401	2.42433	2.19
6.2	2.42287	2.42246	2.42271	2.07
6.6	2.42135	2.42089	2.42118	2.33
7.0	2.41959	2.41934	2.41963	1.57
7.2	2.41882	2.41849	2.41882	1.91
7.4	2.41791	2.41759	2.41767	1.67
7.6	2.41700	2.41673	2.41688	1.35
7.8	2.41609	2.41583	2.41596	1.30
8.0	2.41523	2.41487	2.41517	1.93
8.2	2.41423	2.41382	2.41416	2.19
8.4	2.41326	2.41288	2.41309	1.90
8.6	2.41214	2.41184	2.41215	1.76
8.8	2.41114	2.41089	2.41113	1.42
9.0	2.41015	2.40985	2.41015	1.73
9.2	2.40901	2.40859	2.40902	2.45
9.4	2.40789	2.40766	2.40792	1.42
9.6	2.40677	2.40660	2.40675	0.93
9.8	2.40590	2.40544	2.40562	2.32
10.0	2.40439	2.40421	2.40440	1.07
10.2	2.40328	2.40298	2.40328	1.73
10.4	2.40211	2.40173	2.40194	1.90
10.6	2.40081	2.40023	2.40078	3.27
10.8	2.39954	2.39898	2.39932	2.82
11.0	2.39819	2.39769	2.39803	2.55
11.2	2.39682	2.39640	2.39665	2.11
11.4	2.39540	2.39504	2.39532	1.89
11.6	2.39396	2.39370	2.39386	1.31
11.8	2.39246	2.39209	2.39225	1.86
12.0	2.39104	2.39075	2.39083	1.50

Table 1.34 Refractive Index Values for ZnSe

	Index			Differences ($\times 10^{-4}$)	
Wavelength	Feld1	Feld2	Li	F1 − F2	F1 − Li
1.00	2.488824	2.489152	2.488788	−3.28	0.36
2.00	2.446198	2.446238	2.446179	−0.40	0.20
3.00	2.437579	2.437554	2.437626	0.25	−0.47
4.00	2.433159	2.433120	2.433228	0.39	−0.69
5.00	2.429527	2.429492	2.429599	0.36	−0.72
6.00	2.425844	2.425822	2.425910	0.22	−0.65
7.00	2.421809	2.421808	2.421865	0.02	−0.55
8.00	2.417283	2.417307	2.417326	−0.24	−0.44
9.00	2.412179	2.412235	2.412213	−0.56	−0.33
10.00	2.406437	2.406530	2.406462	−0.93	−0.25
11.00	2.399999	2.400135	2.400023	−1.36	−0.23
12.00	2.392810	2.392995	2.392839	−1.85	−0.29
13.00	2.384809	2.385049	2.384856	−2.40	−0.46
14.00	2.375931	2.376233	2.376009	−3.03	−0.79
15.00	2.366099	2.366472	2.366231	−3.73	−1.32
16.00	2.355229	2.355681	2.355440	−4.52	−2.11
17.00	2.343221	2.343762	2.343545	−5.42	−3.25
18.00	2.329960	2.330602	2.330440	−6.42	−4.80

by Wolfe. Table 1.34 provides the same comparison made with Feldman's two samples and Li's values, all computed from a dispersion equation.

The refractive index and its change with temperature have been measured recently. The change in refractive index with temperature has the same shape as the Irtran material, gradually flattening with decreasing temperature. At about 150 K, in the center of the curve, the new coefficient is 5.3×10^{-5}, compared to an average value of 5×10^{-5}, a very good agreement. The temperature variation is, of course, a second-order quantity.

It is appropriate to compare the recent measurements of Wolfe to those of Feldman. They are shown in Table 1.35. The difference between the values, even after correction for the index of air, is almost 2 in the third decimal place. It must be related to differences in the material or systematic differences in the measurements.

Gallium Antimonide. The refractive index data for gallium antimonide have been reported by Seraphin and Bennett.[40] The data are given in Table 1.36. The data beyond 7.00 μm are suspicious in that they stay constant and then increase.

Gallium Arsenide. The refractive index data for gallium arsenide have been reported,[41] but are rather limited in accuracy and extent for purposes of lens design. They are given in Table 1.37. The wavelengths are reported to be uncertain by ± 0.05 over all but the ends and the refractive index values by ± 0.04 over the entire range.

Cadmium Sulfide. The dispersion equations for this birefringent hexagonal material are[42]:

Table 1.35 Refractive Index Values for ZnSe

	Index			Differences ($\times 10^{-4}$)	
Wavelength	Feld1	Li	Wolfe	F1 − WW	Li − WW
3.0	2.437579	2.437626	2.435651	19.28	19.76
3.50	2.435170	2.435231	2.433260	19.10	19.72
4.50	2.431318	2.431390	2.429419	18.99	19.71
4.60	2.430959	2.431031	2.429018	19.40	20.12
5.00	2.429527	2.429599	2.427628	18.99	19.71
5.40	2.428083	2.428153	2.426538	15.45	16.15
5.80	2.426602	2.426670	2.425157	14.45	15.13
6.20	2.425071	2.425135	2.423597	14.74	15.38
6.60	2.423476	2.423536	2.422076	14.00	14.59
7.00	2.421809	2.421865	2.420316	14.93	15.49
7.20	2.420947	2.421000	2.419546	14.01	14.54
7.40	2.420063	2.420114	2.418636	14.27	14.78
7.60	2.419158	2.419207	2.417725	14.33	14.81
7.80	2.418232	2.418278	2.416815	14.17	14.63
8.00	2.417283	2.417326	2.415955	13.28	13.72
8.20	2.416310	2.416352	2.414954	13.56	13.98
8.40	2.415314	2.415354	2.413984	13.30	13.69
8.60	2.414294	2.414331	2.412864	14.31	14.67
8.80	2.413249	2.413285	2.411863	13.86	14.21
9.00	2.412179	2.412213	2.410873	13.06	13.39
9.20	2.411084	2.411115	2.409733	13.51	13.82
9.40	2.409962	2.409992	2.408612	13.50	13.79
9.60	2.408815	2.408842	2.407492	13.23	13.50
9.80	2.407640	2.407666	2.406622	10.18	10.44
10.00	2.406437	2.406462	2.405112	13.26	13.51
10.20	2.405207	2.405231	2.404001	12.06	12.30
10.40	2.403949	2.403972	2.402831	11.18	11.41
10.60	2.402662	2.402685	2.401530	11.31	11.54
10.80	2.401345	2.401368	2.400260	10.85	11.08
11.00	2.399999	2.400023	2.398910	10.90	11.13
11.20	2.398623	2.398647	2.397539	10.84	11.08
11.40	2.397217	2.397241	2.396119	10.98	11.22
11.60	2.395780	2.395805	2.394678	11.01	11.27
11.80	2.394311	2.394338	2.393178	11.33	11.60
12.00	2.392810	2.392839	2.391757	10.53	10.81

$$n_o^2 = 5.235 + \frac{1.819 \times 10^7}{\lambda^2 - 1.651 \times 10^7}, \tag{1.35}$$

$$n_e^2 = 5.239 + \frac{2.076 \times 10^7}{\lambda^2 - 1.651 \times 10^7}, \tag{1.36}$$

where the wavelengths are in angstroms.

1.3.2.3 Glasses

Several glasses have been formulated for the infrared. They are mainly fused quartz, calcium aluminate, arsenic sulfur (arsenic trisulfide), and AMTIR and are discussed here.

Table 1.36 The Refractive Index of Gallium Antimonide

λ (μm)	n	λ (μm)	n
1.80	3.820	4.00	3.833
1.90	3.802	5.00	3.824
2.00	3.789	6.00	3.824
2.10	3.780	7.00	3.843
2.20	3.764	8.00	3.843
2.30	3.758	9.00	3.843
2.40	3.755	10.00	3.843
2.50	3.749	12.00	3.843
3.00	3.898	14.00	3.861
3.50	3.861	14.90	3.880

Table 1.37 The Refractive Index of Gallium Arsenide

λ (μm)	n	λ (μm)	n
0.78	3.34	14.5	2.82
8.0	3.34	15.0	2.73
10.0	3.135	17.0	2.59
11.0	3.045	19.0	2.41
13.0	2.97	21.9	2.12
13.7	2.895		

Table 1.38 The Refractive Index of Arsenic Trisulfide Glass

i	λ_{i2}	K_i
1	0.0225	1.8983678
2	0.0625	1.9222979
3	0.1225	0.8765134
4	0.2025	0.1188704
5	750.0000	0.9569903

Arsenic Trisulfide. The dispersion equation is the summation form given earlier for the alkali halides. The constants are given[43] in Table 1.38.

AMTIR-1. This chalcogenide glass, $Ge_{33}As_{12}Se_{55}$, is of increasing interest in the infrared. It was measured by Nofziger and Wolfe and by Hilton in 1984, and by both more recently. Since it is a glass, one wonders about batch variations as well as nominal values. The data are shown in Table 1.39, as reported by AMTIR[44] and by Hilton.[45] The data of Wolfe[46] are shown in Table 1.40. The change in refractive index with respect to temperature dn/dT is given in units of 10^{-6} K^{-1} by AMTIR as 101 at 1.15 μm, 77 at 3.39 μm, and 72 at 10.6 μm.

AMTIR-3. This chalcogenide glass, $Ge_{28}Sb_{12}Se_{60}$, has a higher index than its counterpart, AMTIR-1. Data, as reported by AMTIR,[44] are presented in Table 1.41. There are no available independent measurements. The values of the temperature coefficient are given in Table 1.42. The uncertainties are ±2 for the low temperatures and about +18 for the high ones.

Fused Silica. The dispersion formula for fused silica may be written as

$$n^2 - 1 = \sum \frac{K_i \lambda^2}{\lambda^2 - \lambda_i^2}, \tag{1.37}$$

Table 1.39 Refractive Index and Absorption Coefficient of AMTIR-1

Wavelength (μm)	Index	Absorption Coefficient (cm^{-1})
1.0	2.6055	0.07
1.064	2.5933	0.03
1.5	2.5469	0.01
2.0	2.5310	0.004
2.4	2.5250	0.003
3.0	2.5184	0.003
4.0	2.5146	0.002
5.0	2.5112	~0.001
6.0	2.5086	~0.001
7.0	2.5062	~0.001
8.0	2.5036	~0.001
9.0	2.5008	0.006
10.0	2.4977	0.008
11.0	2.4942	0.03
12.0	2.4902	0.15
13.0	2.4862	0.15
14.0	2.4825	0.13

where the coefficients are given[47] in Table 1.43. There are variations among batches and from one manufacturer to another. The user is advised to check and perform a batch check if the application is critical. The variations range to a few parts in the fourth decimal place of the index value.

Arsenic Modified Selenium Glass. Data for arsenic modified selenium glass are dependent on the batch; two different prisms varied by about one part in the third decimal. Data for one sample are given[48] in Table 1.44.

Irtrans. These materials are hot-pressed compacts. With enough pressure and temperature they can be formed in domes up to about 25 cm in diameter and in other shapes as well. They approach theoretical density. They were developed by Eastman Kodak, but the hot-pressed forms of the materials are available from other vendors. Table 1.45 provides the data on Irtrans, as provided by the Eastman Kodak Company.[49] Values were initially determined at selected wavelengths. These values are a result of the dispersion equation; all values beyond 10 μm are extrapolated.

Table 1.40 Refractive Index Runs for Sample 001A of AMTIR-1

Wavelength	Index				Standard Deviation
	Run 1	Run 2	Run 3	Run 4	
4.6	2.51118	2.51122	2.51133	2.51119	0.69
5.0	2.51021	2.51019	2.51024	2.51034	0.67
5.4	2.50932	2.50933	2.50925	2.50926	0.41
5.8	2.50859	2.50841	2.50839	2.50841	0.94
6.2	2.50729	2.50726	2.50729	2.50732	0.24
6.6	2.50621	2.50634	2.50623	2.50634	0.70
7.0	2.50542	2.50545	2.50537	2.50543	0.34
7.2	2.50508	2.50496	2.50495	2.50507	0.70
7.4	2.50480	2.50442	2.50440	2.50443	1.92
7.6	2.50388	2.50390	2.50388	2.50377	0.59
7.8	2.50358	2.50343	2.50339	2.50333	1.07
8.0	2.50295	2.50288	2.50298	2.50289	0.48
8.2	2.50234	2.50235	2.50231	2.50233	0.17
8.4	2.50181	2.50167	2.50175	2.50171	0.60
8.6	2.50116	2.50115	2.50121	2.50113	0.34
8.8	2.50061	2.50058	2.50064	2.50063	0.26
9.0	2.50007	2.50013	2.50010	2.50011	0.25
9.2	2.49963	2.49953	2.49946	2.49951	0.71
9.4	2.49892	2.49883	2.49882	2.49882	0.49
9.6	2.49824	2.49817	2.49827	2.49822	0.42
9.8	2.49765	2.49758	2.49754	2.49763	0.50
10.0	2.49696	2.49682	2.49693	2.49685	0.66
10.2	2.49628	2.49609	2.49621	2.49630	0.95
10.4	2.49562	2.49549	2.49548	2.49546	0.73
10.6	2.49485	2.49474	2.49490	2.49476	0.75
10.8	2.49423	2.49401	2.49420	2.49406	1.07
11.0	2.49349	2.49335	2.49355	2.49348	0.84
11.2	2.49283	2.49254	2.49266	2.49265	1.20
11.4	2.49200	2.49188	2.49176	2.49226	2.14
11.6	2.49109	2.49103	2.49104	2.49111	0.39
11.8	2.49030	2.49027	2.49007	2.49040	1.38
12.0	2.48960	2.48958	2.48944	2.48956	0.72

Table 1.41 The Refractive Index and Absorption Coefficient of AMTIR-3

Wavelength (μm)	Index	Absorption Coefficient (cm^{-1})
3	2.6266	0.002
4	2.6210	0.001
5	2.6173	0.001
6	2.6142	0.001
7	2.6117	0.001
8	2.6088	0.002
9	2.6055	0.004
10	2.6023	0.008
11	2.5983	0.03
12	2.5942	0.13
13	2.5892	0.20
14	2.5843	0.20

Table 1.42 The Temperature Coefficient of Refractive Index of AMTIR-3

Wavelength (μm)	Low $dn/dT \times 10^6$ (76 to 298 K)	High $dn/dT \times 10^6$ (298 to 423 K)
3	58	98
5	57	92
8	55	87
10	56	91
12	56	93

Table 1.43 The Refractive Index of Fused Silica

i	λ_i	K_i
1	0.0684043	0.6961663
2	0.1162414	0.4079426
3	9.896161	0.8974794

Table 1.44 The Refractive Index of Arsenic-Selenium Glass

λ (μm)	n	λ (μm)	n
1.0140	2.5783	7.00	2.4787
1.1286	2.5565	7.50	2.4784
1.3622	2.5294	8.10	2.4778
1.5295	2.5183	8.50	2.4775
1.7012	2.5100	9.10	2.4771
2.1526	2.4973	9.50	2.476
3.00	2.4882	10.00	2.4767
3.4188	2.4858	10.50	2.4759
4.00	2.4835	11.00	2.4758
4.50	2.4822	11.50	2.4753
5.00	2.4811	12.00	2.4749
5.50	2.4804	13.00	2.4760 (sic)
6.00	2.4798	13.50	2.4748
6.50	2.4792	14.00	2.4743

Table 1.45 The Refractive Indices of Irtran Materials

Wavelength	Irtran 1 MgF_2	Irtran 2 ZnS	Irtran 3 CdF_2	Irtran 4 ZnSe	Irtran 5 MgO
7.2500	1.2865	2.2282	1.3648	2.422	1.5307
7.5000	1.2792	2.2260	1.3600	2.421	1.5154
7.7500	1.2715	2.2237	1.3550	2.419	1.4993
8.0000	1.2634	2.2213	1.3498	2.418	1.4824
8.2500	1.2549	2.2188	1.3445	2.417	1.4646
8.5000	1.2460	2.2162	1.3388	2.416	1.4460
8.7500	1.2367	2.2135	1.3330	2.415	1.4265
9.0000	1.2269	2.2107	1.3269	2.413	1.4060
9.2500		2.2078	1.3206	2.411	
9.5000		2.2048	1.3141	2.410	
9.7500		2.2018	1.3073	2.409	
10.000		2.1986	1.2694	2.407	
11.000		2.1846		2.401	
12.000		2.1688		2.394	
13.000		2.1508		2.386	
14.000				2.378	
15.000				2.370	
16.000				2.361	
17.000				2.352	
18.000				2.343	
19.000				2.333	
20.000				2.323	

1.3.2.4 Miscellaneous Materials

Some of the materials used or considered for use in infrared instrumentation are not semiconductors or alkali halides, and do not fit into a convenient category. They are described here.

Calcium Carbonate. Calcium carbonate, also known as calcite, has been a useful birefringent crystal for many years. Its refractive indices have been reported by several investigators. They are listed in Table 1.46 for the infrared part of the spectrum.[50] The temperature coefficients in the visible (0.211- to 0.643-μm region) for the ordinary and extraordinary indices are about 0.3 and $1.3 \times 10^{-6}\ \mathrm{K}^{-1}$, respectively.

Calcium Fluoride. The dispersion equation for calcium fluoride is the same as that for fused silica; the constants are given[51] in Table 1.47. Values for the change of refractive index with temperature have also been reported. Values for the infrared are reported in Table 1.48 for a temperature of approximately 333 K.

Magnesium Oxide. The dispersion equation is[52]:

$$n^2 - 1 = 1.956362 - 0.01062387\lambda^2 - 0.0000204968\lambda^4 - \frac{0.02195770}{\lambda^2 - 0.01428322} \quad (1.38)$$

The temperature coefficients are given in Table 1.49.

Sapphire. The dispersion equation for the ordinary ray of sapphire is the same as that for fused silica; the coefficients are given[53] in Table 1.50. The

Table 1.46 The Refractive Indices of Calcite

λ (μm)	n_o	n_e	λ (μm)	n_o	n_e
1.042	1.64276	1.47985	1.609	1.63261	
1.097	1.64167	1.47948	1.615		1.47695
1.159	1.64051	1.47910	1.682	1.63127	
1.229	1.63926	1.47870	1.749		1.47638
1.273	1.63849		1.761	1.62974	
1.307	1.63789	1.47831	1.849	1.62800	
1.320	1.63767		1.900		1.47573
1.369	1.63681		1.946	1.62602	
1.396	1.63637	1.47789	2.053	1.62372	
1.422	1.63590		2.100		1.47492
1.479	1.63490		2.172	1.62099	
1.497	1.63457	1.47744	3.324		1.47392
1.541	1.63381				

Table 1.47 The Refractive Index of Calcium Fluoride

i	λ_i	K_i
1	0.050263605	0.5675888
2	0.1003909	0.4710914
3	34.649040	3.8484723

Table 1.48 The Temperature Change of the Refractive Index for Calcium Fluoride

λ	dn/dT
1.2	−1.040
1.25	−1.029
1.30	−1.018
2.0	−0.932
3.16	−0.881
4.2	−0.831
5.3	−0.821
6.5	−0.787

Table 1.49 Temperature Change of Refractive Index for Magnesium Oxide

	dn/dT (10^{-6} °C^{-1}) 14.8				
λ (μm)	20°C	25°C	30°C	35°C	40°C
7.679	13.6	13.7	13.8	13.9	14.0
7.065	14.1	14.2	14.3	14.4	14.5
6.678	14.4	14.5	14.6	14.7	14.8
6.563	14.5	14.6	14.7	14.8	14.9
5.893	15.3	15.4	15.5	15.6	15.7
5.461	15.9	16.0	16.1	16.2	16.3
4.861	16.9	17.0	17.1	17.2	17.3
4.358	18.0	18.1	18.2	18.3	18.4
4.047	18.9	19.0	19.1	19.2	19.3

Table 1.50 Dispersion Constants for Sapphire

i	λ_{i2}	K_i
1	0.00377588	1.023798
2	0.0122544	1.058264
3	321.3616	5.280792

temperature coefficient of refractive index between 19 and 24°C decreases from about 20×10^{-6} at the short wavelengths to 10×10^{-6} near 4 μm.

Barium Fluoride. The dispersion equation for barium fluoride is the same form as that for fused silica; the coefficients are given[54] in Table 1.51.

Diamond. The refractive index of diamond is reported[55] in Table 1.52.

1.3.3 Permittivity

Permittivity or dielectric constant is defined in Sec. 1.2.13. Table 1.53 provides the values for the materials, the frequency and temperature at which the values were measured, and references for further study.

Table 1.51 The Dispersion Constants for Barium Fluoride

i	λ_{i2}	K_i
1	0.0033396	0.63356
2	0.012030	0.506762
3	2151.70	3.8261

Table 1.52 The Refractive Index of Diamond

λ (μm)	n
0.480	2.4368
0.486	2.4354
0.546	2.4235
0.589	2.4175
0.644	2.4114
0.656	2.4104

1.3.4 Hardness

Hardness measures are discussed in Sec. 1.2.7. Knoop values are included in Table 1.54 for the most part. The temperature at which the measurement was made and the load applied to the indenter are given where available. The references provide further information about the data.

1.3.5 Thermal Properties

As mentioned in Sec. 1.2.5, although almost all properties of materials are functions of temperature, these thermal properties are probably foremost: melting or softening temperature, specific heat, Debye temperature, thermal expansion, and thermal conductivity. The summary data are provided in the following tables.

The melting or softening temperatures are defined and discussed in Sec. 1.2.5. Summary data are given in Table 1.55. Crystals melt; glasses soften. No material should be used at temperatures close to its melting or softening point. The higher the specific heat, the more heat that can be absorbed without a temperature rise. Specific heat is dimensionless, because it is ratioed to the value for water. The temperature listed is the temperature at which the specific heat was measured. All references are given first for specific heat.

The Debye temperature is discussed in Sec. 1.2.6. Data for it are given in Table 1.56 in kelvins. There are several methods of measurement, and these are also listed. If the measurement was obtained by measuring the specific heat (as a function of temperature), then c_p is indicated; by elastic coefficient methods, c_{ij}; but calorimetry, cal; by bulk modulus, β; by temperature, T; by thermal conductivity, k; and by density, ρ.

Thermal conductivity and the temperature at which it was measured are listed in Table 1.57. Thermal conductivity is discussed in Sec. 1.2.5. The constant, linear, and quadratic coefficients of linear thermal expansion are given in the same table, along with the temperature of measurement and references. Thermal expansion is also discussed in Sec. 1.2.5.

Table 1.53 Permittivity (Relative Dielectric Constant)

Material	Permittivity	Frequency (Hz)	Temperature (K)	Reference
Al				
Ag				
AgCl	12.3	10^6	293	56
Al_2O_3	10.55p, 8.6s	10^2 to 10^8	298	57
As_2S_3	8.1	10^3 to 10^6	—	58
Au				
BaF_2	7.33	2×10^6	—	59
$BaTiO_3$	1240–1100	10^2 to 10^8	298	57
C				
CaF_2	6.76	10^5	—	59
$CaCO_3$	8.5s, 8.0p	10^4	290 to 295	60
$CaTiO_3$	140.0	1.5×10^6	294	
CdTe	11	10^5	$5.5 \times 10^{13}/cm^3$	61
CsBr	6.51	2×10^6	293	59
CsI	5.65	10^6	298	59
Cu				
CuBr	8.0	3×10^6	293	
CuCl	10.0	5×10^5	293	
CuTe				
GaAs	11.1	—	—	
GaSb				
Ge	16.6	9.37×10^9	9 Ω cm	62
Irtran 1	5.1	10^{10}		63
Irtran 2	8–8.5	10^{10}		63
Irtran 3				
Irtran 4				
Irtran 5				
KBr	4.9	10^2 to 10^{10}	298	57
KCl	4.64	10^6	302.5	59
KI	4.94	2×10^6	—	59
KRS-5	32.9–32.5	10^2 to 10^7	298	57
LiF	9.00, 9.11	10^2 to 10^{10}	298, 360	57
MgF_2	4.87p, 5.45s	10^5 to 10^7		64
MgO	9.65	10^2 to 10^8	298	57
$MgO.3.5Al_2O_3$	8.0–9.0		—	
NaCl	5.90, 6.35 to 5.97	10^2 to 2.5×10^{10}	298, 358	57
NaF	6.0	2×10^6	292	59
$NaNO_3$				
Se	6.0	10^2 to 10^{10}	298	57
Si	13.0	9.37×10^9	—	62
Se(As)	234–230	10^2 to 10^{10}	298	
SiO_2 (crystal)	4.34s, 4.27s	3×10^7	290 to 295	60
SiO_2 (fused)	3.78	10^2 to 10^{10}	298	57
$SrTiO_3$				
TiO_2	200–160	10^4 to 10^7	298	
TlBr	30.3	10^3 to 10^7	298	
TlCl	31.9	2×10^6	298	
ZnS				
ZnSe				
ZnTe				

Table 1.54 Hardness of Optical Materials

Material	Knoop Value	Temperature	Load	Reference
AgCl	9.5	—	200	65
Al_2O_3	1370	—	1000	66
As_2S_3	109	—	100	67
AMTIR-1	170			68
AMTIR-3	150			68
BaF_2	82	—	500	69
C	8820	110	—	
$CaCO_3$	Moh 3			70
CaF_2				60
CdS	55, 80	—	—	
CdSe	90, 44, 66	—	—	
CdTe	56	—	—	
CsBr	19.5	—	200	69
Cu	48, 17.5, 8	293, 773, 973	—	
CuBr	21.2	—	—	
CuTe	19.2	—	—	
GaAs	721	—	—	
GaSb	469	—	—	
Ge	176, 83, 80, 24	873, 973, 1023, 1223	—	
InAs	330	—	—	
InSb	225	—	—	
InP	430	—	—	
Irtran 1	576 Moh 6			71
Irtran 2	354			71
KBr	5.9, 7.0	—	200	65
KCl	7, 2, 9.3	—	200	65
KRS-5	40.2, 39.8, 33.2	—	200, 500, 500	65
LiF	102–113	—	600	72
MgO	692	—	600	65
$MgO.3.5Al_2O_3$	1140	—	1000	
NaCl	15.2, 18.2	—	200	65
$NaNO_3$	19.2	—	200	65
Si	1000, 500, 128	293, 773, 1273	—	73
SiO_2	461, 741	—	200, 500	65
$SrTiO_3$	595	—	200, 500	74
TlBr	11.9	—	500	65
TlCl	12.8	—	500	65
TiO_2	879, 792		500, 1000	71
ZnS	178	—	—	
ZnSe	137	—	—	
ZnTe	82	—	—	

Table 1.55 Melting or Softening Temperature and Specific Heat for Various Materials

Material	Melting or Softening Temp.	Specific Heat	Temperature	References
Ag	1233.8	—		
AgCl	730.7	0.0848	298	75
Ag^2S	—	0.072	273	
Al	933.2	0.218	298	
Al_2O_3	2303	0.180, 0.174	298, 273	
AlSb		—	—	71
As_2S_3	483	—	—	58
BaF_2	1553	—	—	60
C (diamond)	>3773	0.122	300	59
$CaCO_3$	1612.0 @ 100 atm	0.203, 0.214	273	59
CaF_2	1633	0.204, 0.212	273, 373	60
CdS	1560.0 @ 100 atm	0.08820	273	76
CdSe				
CdTe	1314–1323	0.01875	323	77
CsBr	909.0	0.6300	293	60
CsI	894.0	0.04800	293	60
Cu	1356.0	0.092	300	
GaAs	1511.0		—	78
GaSb	993.0	0.01828	—	79
Ge	1209	0.074	273–373	80
InAs	1215	3.4, 5.2, 7.01	78–290, 573–673	
InSb	796.0	0.023, 0.248	180, 300	81
InP	1343			79
Irtran 1	1396	0.22		
Irtran 2	>1073			
KBr	1003.0	0.104, 0.108	273, 373	60
KCl	1049.0	0.162, 0.168	273, 373	59
KI	996.0	0.73, 0.75	200, 270	59
KRS-5	687.5			82
LiF	1143.0	0.373	283	60
MgO	3073.0	0.209	273	
$MgO.3.5Al_2O_3$	2030–2060	0.03	308	71
MgF_2	1255			
NaCl	1074.0	0.204, 0.217	273, 373	60
$NaNO_3$	579.8	0.247, 0.270	273, 373	
Pt	2046.5	0.0318	273	
Se	308	0.068	276	83
Se(As)	~343			84
Si	1693	0.177	298	80
SiO_2	1743	0.188	285–373	85
$SrTiO_3$	2353			
Te	722.8	0.0483	561–646	86
TiO_2	2093.0	0.17	293	
TlBr	733.0	0.045	293	
TlCl	703.0	0.0520	273	
ZnS				
ZnSe				
ZnTe				

Table 1.56 Debye Temperatures for Various Materials

Material	Debye Temperature	Methods
Ag	212, 220, 212, 220, 203	c_p, c_{ij}, cal, ρ
AgCl	180	—
Ag^2S	—	c_p, c_{ij}, cal, β, ρ, T, k
Al	385, 399, 396	c_p, c_{ij}, cal
Al_2O_3		
AlSb		—
As_2S_3		—
BaF_2		—
C (diamond)	2050, 1860, 1491, 2200	c_p, β, ρ
$CaCO_3$		
CaF_2	470, 474	c_p, β
CdS		
CdSe		
CdTe		
CsBr		
CsI		
Cu	310–330, 310, 329, 313, 325, 333	c_p, c_{ij}, cal, β, ρ
GaAs		
GaSb		
Ge	360, 370	—
InAs		
InSb	208, 200 ± 5	c_{ij}, c_p
InP		
KBr	152–183, 180, 185, 171, 176, 162	cal, c_p
KCl	218–235, 230, 233, 229, 226, 203	—
KI	195, 162, 119	—
KRS-5		
LiF	607–750, 680, 686, 1020, 845, 440	c_p, cal, k, T
MgO	800	—
$MgO.3.5Al_2O_3$		
NaCl	275–300, 280, 292, 294, 276, 235	c_p, cal, k, T
$NaNO_3$		
Pt		
Se		
Se(As)		
Si		
SiO_2	255	—
$SrTiO_3$		
Te	130	—
TiO_2	450	
TlBr		
TlCl	703.0	0.0520
ZnS	260	—
ZnSe		
ZnTe		
ZrN		

1.3.6 Solubility, Molecular Weight, and Density (Specific Gravity)

Some of the long-wavelength materials are soluble in water to a degree that is serious for instrumentation. The data are given in Table 1.58 in grams per 100 ml, which, since water has a density of 1 g/ml, is a form of percentage. Specific gravity is the density of a material with respect to that of water. These two concepts are discussed in more detail in Secs. 1.2.8 and 1.2.11. *Insoluble* means a solubility less than 0.001 in these units.

1.3.7 Elastic Coefficients

The fundamental mechanical properties of crystals listed in Table 1.59 can be used to calculate engineering moduli and other important characteristics (of crystals). They are presented in bar or dyne cm^{-2} and are discussed in more detail in Sec. 1.2.10.

1.3.8 Engineering Moduli

Table 1.60 shows the properties that are important in the calculation of the durability of a window or lens under various loads. They are given in the time-honored units of pounds per square inch.

1.4 MIRROR DATA

Mirrors are made from many different substrates. Most are then covered with a metallic coating to give them a surface with high reflectivity throughout the desired spectrum. The coatings most often applied are aluminum, gold, and silver, although rhodium and tantalum are sometimes used. Aluminum is not stable, and soon becomes overcoated with a layer of aluminum oxide, which may not be stoichiometric sapphire. It is, however, a good, hard coat. Silver is more inert, but it too does not retain the same high reflectivity after exposure to air. Gold is an inert, noble metal that retains its very high reflectivity and, consequently, low emissivity. The specular spectral reflectivity representative of these coatings is given[150] in Fig. 1.25.

One of the important considerations in the choice of a mirror is its final weight. The weight is determined by both the required thickness to obtain a blank that will not sag or print through and by the density of the material. Table 1.61 is a compilation of the most useful mirror blank materials, their densities, and their Young's moduli. The third column is a figure of merit by which these materials may be judged, at least in part. It is the "strength-to-weight" ratio. In considering the required weight of a mirror blank, you should be aware of the fact that most can be "lightweighted" by removing material from the back in an egg-crate pattern, and more than 75% or the material can usually be removed (or not included in the first place) without affecting the surface figure.

The scatter of a mirror is due mostly to the roughness of its substrate surface. In the infrared, it may be dominated by a few, large disparities, since the scatter from surface roughness may be very low at these longer wavelengths. Therefore, it is often necessary to obtain the smoothest possible surface before the coating process begins. It is very rare indeed for a mirror to get smoother

Table 1.57 Coefficients of Linear Thermal Expansion and Thermal Conductivity

Material	Conductivity (m cal cm^{-1} K^{-1})	Temperature (K)	Reference	Temperature (K)	A	B	C	References
Al								
AgCl	2.75	295	87	30	293–333			99
Al_2O_3	60p 55s	299 296	88	323 323	6.7p 5.0s			100
AMTIR-1	0.6		80		12			68
AMTIR-3			80		13.5			68
As_2S_3	0.4	313	89	306–438	24.62			87
BaF_2	28	286	90	273–1073	18.4			101
$BaTiO_3$	1.6 3.2	401 room	91	193–253 283–343 393–453	16.0 19.0 13.0			102
C	34.0 8.6 6.59	76 194 272	90 90 90	17–300	1.38			98
CaF_2	93.2 36 24.7 23.2 19.1	83 200 273 309 373	92	181–280 316–900	18.38 1.851	2.511 1.481 −21.10 21.52		
$CaCO_3$	13.2p 11.1s	273 273	93 93	123–273 123–273 323 323 638 348–673	24.39p − 5.68s 26.6 5.2 − 3.8 24.71	0.533 0.0333	−30.7 − 4.58	
$CaTiO_3$								
CdS	38	287	89	300–343	4.2			89
CdTe	15			323 873	4.5 5.9			

Table 1.57 *(continued)*

Material	Conductivity (m cal cm^{-1} K^{-1})	Temperature (K)	Reference	Temperature (K)	A	B	C	References
CsBr	2.3 2.2 2.6	298 318 338	90	293–323 134–573	47.9 46.6			99
CsI	2.7	298	68	298–323	50.0			99
Cu								
CuBr				293–423	19.0			
CuCl				313–413	10.0			
CuTe								
GaAs					5.7			103
GaSb	130	300			6.9			
Ge	105 140	300 293	91	40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240	0.07 0.20 0.39 0.67 1.05 1.54 2.20 2.79 3.25 3.62 3.91 4.12 4.29 4.45 4.58 4.70 4.82 4.93 5.03 5.13 5.23			

(continued)

Table 1.57 *(continued)*

Material	Conductivity (m cal $cm^{-1} K^{-1}$)	Temperature (K)	Reference	Temperature (K)	A	B	C	References
Ge	140	293	91	250 260 270 280 290 300	5.32 5.42 5.50 5.59 5.67 5.75			
Irtran 1	35 26	329 452		298–573	11.0			
Irtran 2	37 26	327 447		298–573	6.9			
Irtran 3	19 15	353 449		298–573	20.0			
Irtran 4	31 16	327 695		298–573	7.7			
Irtran 5	104 70	298 441		298–573	12.0			
KBr	6.98 11.5	299 319	94 92	113–573 318–953 293–333	27.6 37.99 43.0	4.1 1.263	5.256	
KCl	15.6	315	92	293–333	36.0			99
KI	5.0	299	94	313	42.6			104
KRS-5	1.3	293	90	223–293 293–373	61.0 58.0			99
LiF	27	314	95	273–373 123–273 320–1067	37.0 31.95 33.17	 5.049 3.075	−4.070 2.399	95
MgF_2								
MgO	2620 7560	10 30		323–988 300	10.98 11.2			

Table 1.57 *(continued)*

Material	Conductivity (m cal cm^{-1} K^{-1})	Temperature (K)	Reference	Temperature (K)	A	B	C	References
MgO	6380 140 77	100 300 500	481 659 825 964	12.3 13.5 14.6 15.4 16.0				
$MgO.3.5Al_2O_3$	33	308	88	313	5.9			100
NaCl	15.5	289	90	223–473	44.0			99
NaF	124 25.2 22	83 273 298			36.0			105
$NaNO_3$				323	12p 11s			93
Se	2.6			195–292 195–273 273–294 293–373 478	20.3 42.7 48.7 22.9 45.2			
Si	390	313	96	50–100 50–100 100–200 100–200 200–300 200–300 400–500 400–500 500–600 500–600 600–700 600–700 25–900	2.5⟨111⟩ 2.7⟨110⟩ 3.1⟨111⟩ 3.5⟨110⟩ 3.9⟨111⟩ 3.8⟨110⟩ 4.3⟨111⟩ 4.1⟨110⟩ 4.7⟨111⟩ 4.4⟨110⟩ 5.0⟨111⟩ 4.5⟨110⟩ 3.0024			
Se(As)					34.0			

(continued)

Table 1.57 *(continued)*

Material	Conductivity (m cal cm^{-1} K^{-1})	Temperature (K)	Reference	Temperature (K)	A	B	C	References
SiO_2 (crystal)	25.5p to c 14.8s to c	323 323	97 97	173–310 283–607 273–633 633–723 273–353 273–353	7.067p 7.067p 25.80p 7.97p 13.37s			 105 105
SiO_2 (fused)	2.82	314	98	293–1173	0.5			107
$SrTiO_2$					~ 9.4			108
TiO_2	30p 33p 21s 17s	309 340 317 340	94	313 313	9.19p 7.14s			100
TlBr	1.4	316	94	293–353	51.0			109
TlCl	1.8	311	94	293–333	53.0			109
ZnS								
ZnSe								
ZnTe								

Table 1.58 Solubility, Molecular Weight, and Density

Material	Solubility (g/100 ml H_2O)	Molecular Weight	Specific Gravity	Temperature	References
Al	Insoluble				
AMTIR-1			4.4		110
AMTIR-3			4.6		110
Ag	Insoluble				
AgCl	8.9×10^{-5} @ 283 K	143.34	5.589 5.56	273 293	111
Al_2O_3	Insoluble	101.94	3.98		112
As_2S_3	Insoluble	364.02	3.198		113
Au	Insoluble				
BaF_2	0.17	175.36	4.83	293	114, 114
$BaTiO_3$		232.96	5.90		
C	Insoluble				
CaF_2	0.0017 @ 299 K	78.08	3.179	298	
$CaCO_3$	1.4×10^{-3} @ 298 K	100.09	2.7102	293	
$CaTiO_3$		135.98	4.10	293	
CdS		144.48	4.82	293	
CdTe	Insoluble	240.02	5.854		
CsBr	124.3 @ 298 K	212.83	4.44	293	
CsI		259.83	4.526		114, 115
Cu	Insoluble				
CuBr	Insoluble	143.46	4.718	293	114, 114
CuCl			3.53	293	
CuTe					
GaAs	Insoluble	144.63	5.3161	298	
GaSb	Insoluble	191.48			
Ge	Insoluble; soluble in hot sulfuric acid and aqua regia	72.60	5.327	298	
Irtran 1	Insoluble	62.32	3.18		
Irtran 2	Insoluble	97.45	4.09		
Irtran 3	Insoluble	78.08	3.18		
Irtran 4	Insoluble	144.34	5.27		
Irtran 5	6.2×10^{-4}	40.32	3.58		
KBr	53.48 @ 273 K 102 @ 373 K	119.01	2.75	298	113, 116
KCl	34.7 @ 293 K	74.55	1.984	293	
KI	127.5 @ 273 K	116.02	3.13		
KRS-5	0.05 @ room temp		7.371	289	117
LiF	0.27 @ 291 K	25.94	2.639	298	114, 118
MgF_2	Insoluble				
MgO	Insoluble; soluble in acids and ammonia salts	40.32			114, 119

(continued)

Table 1.58 *(continued)*

Material	Solubility (g/100 ml H_2O)	Molecular Weight	Specific Gravity	Temperature	References
$MgO.3.5Al_2O_3$	Insoluble; soluble in acids and ammonia salts	356.74	3.61		
NaCl	35.7 @ 273 K 39.12 @ 373 K Soluble in glycerine; slightly soluble in alcohol; insoluble in HCl	58.45	2.164	293	
NaF	4.22 @ 291 K	42.00	2.79	293	
$NaNO_3$	73 @ 273 K 180 @ 373 K	85.01	2.261		
Se	Insoluble		4.82		
Se(As)	Insoluble				
Si	Insoluble				
SiO_2 (crystal)	Insoluble	60.06	2.648	298	
SiO_2 (fused)	Insoluble	60.06	2.202	293	120
$SrTiO_3$	Insoluble	183.53	5.122	293	
Te	Insoluble				
TiO_2	Insoluble; soluble in acids	79.90	4.25		112
TlBr	0.05 @ 298 K	284.31	7.453	298	
TlCl	0.32 @ 293 K	238.85	7.018	298	
ZnS					
ZnSe					
ZnTe					

Table 1.59 Elastic Coefficients

Material	T	c_{11}	c_{12}	c_{13}	c_{33}	c_{44}	References
Al							
Ag							
AgCl		6.01	3.62			0.625	121
Al_2O_3	298	49.68	16.36		49.81	14.74	
As_2S_3							
Au							
BaF_2		9.01	4.03			2.49	
$BaTiO_3$	298	8.18	2.98	1.95	6.76	18.30	122
C		95	5.02			3.47	123
CaF_2		16.4	5.3			3.37	124
$CaCO_3$		13.71	4.56	4.51	7.97	3.42	124, 125
$CaTiO_3$							
CdS		8.432	5.212	4.638	9.397	1.489	126

Table 1.59 *(continued)*

Material	T	c_{11}	c_{12}	c_{13}	c_{33}	c_{44}	References
CdTe		5.351	3.681			1.994	
CsBr		3.097	0.403			0.75	127
CsI		2.46	0.67			0.624	127
Cu							
CuBr							
CuCl							
CuTe							
GaAs		1.192	0.5986			0.538	
GaSb		8.849	4.037			4.325	128
Ge		1.29	4.83			6.71	129
Irtran 1							
Irtran 2							
Irtran 3							
Irtran 4							
Irtran 5							
KBr		3.45	0.54			0.508	130
KCl		3.98	0.62				131
KI		2.69	0.43			0.362	132
KRS-5		3.31	1.32			0.579	133
LiF		9.74	4.04			5.54	134
MgF_2							
MgO		2.90	0.876			1.55	135
$MgO.3.5Al_2O_3$							
NaCl		4.85	1.23			1.26	132
NaF		9.09	2.64			1.27	
$NaNO_3$		8.67	1.63	1.60	3.74	2.13	
Se							
Se(As)							
Si		1.67	0.65			0.80	136
SiO_2 (crystal) SiO_2 (fused)		11.60	1.67	3.606	3.28	11.04	132
$SrTiO_3$		31.56	10.27			12.15	
Te	300	3.265	0.195	2.493	7.22	3.121	
TiO_2		35.8	26.7	17.0	47.9	12.5	
TlBr		3.78	1.48			0.756	133
TlCl		4.01	1.53			0.760	133
ZnS		9.45	5.70			4.36	
ZnSe							
ZnTe							

Table 1.60 Engineering Elastic Moduli

Material	Young's Modulus (Mpsi)	Rigidity (Mpsi)	Bulk (Mpsi)	Rupture (kpsi)	Apparent Elastic Limit (kpsi)	References
Al						
Ag						
AgCl	0.02	1.03	6.39		3.8	137, 137, *,[a] *
Al_2O_3	50.00	21.50	0.30			138, 138, 138
As_2S_3	2.30	0.94		2.4		139, 139, 139
Au						
BaF_2	7.70			3.9×10^4		140, 140
$BaTiO_3$	4.90	18.30	23.50			141, *, 142
C						
CaF_2	11.00	4.90	12.00	5.3	5.3	143, *, *, 143, *
$CaCO_3$						
$CaTiO_3$						
CdS						
CdTe				0.850		144
CsBr	2.30			0.0239	1.22	140, 140, 140
CsI	0.769				0.81	137, 137
Cu						
CuBr						
CuCl						
CuTe						
GaAs				10.436	16	
GaSb	9.19	6.28	8.19			
Ge	14.90	9.73	11.30			*, *, 145
Irtran 1	16			21.8		
Irtran 2	14			14.1		
Irtran 3						
Irtran 4						
Irtran 5						
KBr	3.90	0.737	2.18	0.48	0.16	137, 137, *, 137, 137
KCl	4.30	0.906	2.52	0.64	0.33	137, 137, *, 137, 137
KI	4.57	0.90	124			*, *, 146
KRS-5	2.30	0.840	2.87	18.1	38	137, 137, *, 137, 137
LiF	9.40	8.00	9.00	2.0×10^3	$1.62 \times 10_4$	143, *, *, 143, *,
MgF_2						
MgO	36.10	22.40	22.40			
$MgO.3.5Al_2O_3$						

Table 1.60 *(continued)*

Material	Young's Modulus (Mpsi)	Rigidity (Mpsi)	Bulk (Mpsi)	Rupture (kpsi)	Apparent Elastic Limit (kpsi)	References
NaCl	5.80	1.83	3.53	0.57	0.35	137, *, *, 137, 137
NaF						
$NaNO_3$			3.80			147
Se						
Se(As)						
Si	19	11.6	14.8			*, *, 145
SiO_2 (crystal)	11.10s 14.10p	5.28				148, 147
SiO_2 (fused)	10.6	4.52				149, 148
$SrTiO_3$						
Te						
TiO_2						
TlBr	4.28	1.10	3.26			
TlCl	4.60	1.10	3.42			
ZnS						
ZnSe						
ZnTe						

[a] * = no reference.

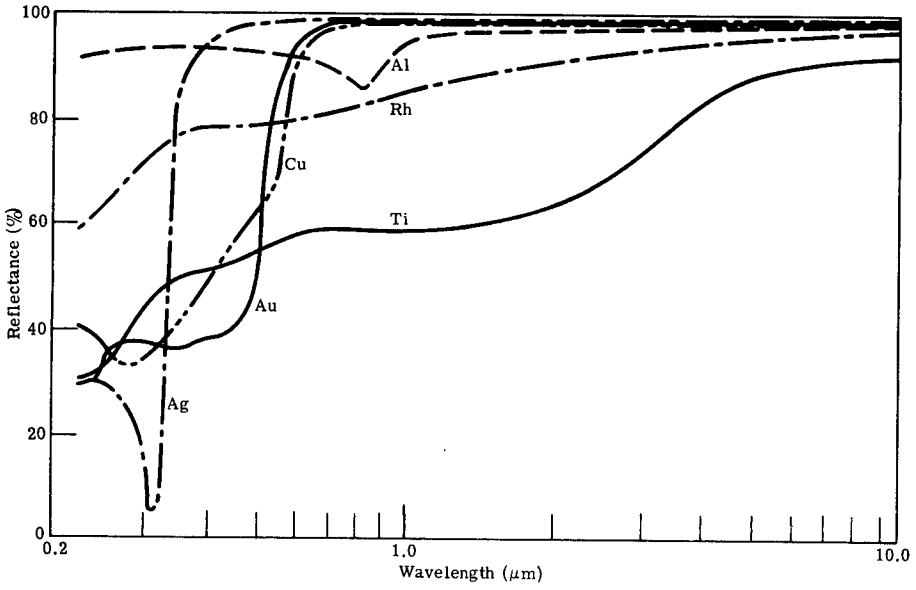


Fig. 1.25 Reflectance of various films of silver, gold, aluminum, copper, rhodium, and titanium.

as a result of the overcoat. In general, glass-type substrates can be polished smoother than metallic ones. One important exception to this is the electroless nickel coating that is often applied to beryllium mirrors. These coatings are sufficiently thick and serve a purpose other than just the generation of high reflectivity; they really become the substrate layer on the beryllium, on which is coated aluminum, gold, or another high-reflectance coating.

For many infrared applications the mirror is used at quite cold temperatures, below 100 K, at which point the properties of thermal expansion (contraction) become important. In general, metallic mirrors have higher expansion coefficients than "glass" ones and higher thermal conductivity. Thus, to some extent, the introduction or subtraction of heat results in a uniform decrease and a quicker equilibration with a structure or heat sink for a metallic mirror than a vitreous one. Some of the vitreous substrates have very low expansion coefficients because they have been manufactured with that in mind. These include ultralow-expansion fused silica, ULE, and Zerodur. Although the coefficients reported may be very low, and in some temperature regimes are actually zero, the important consideration is how much the materials actually expand or contract over the temperature range of importance. The degree of expansion is a function of the temperature.

For some applications, immunity to irradiation by high-energy particles (gamma rays, neutrons, etc.) is an important consideration. This fluence of high-energy radiation, either natural or artificial, will impart heat to the surface as the particles are absorbed and can cause the films to wrinkle and peel. One solution is to eliminate the reflective layer, and polish the substrate. Of course, the blank must be a metal. Usually such surfaces are not as smooth as the coated ones, but they can withstand some levels of irradiation, particularly if they have low enough atomic numbers, as does beryllium.

1.4.1 Density and Young's Modulus

Table 1.61 lists the density and Young's modulus of several mirror blank materials. The ratio is also given as an indication of the strength-to-weight ratio. This is merely a guide; other parameters enter into the choice of material.

1.4.2 Thermal Expansion and Thermal Conductivity

Mirrors are used for many different purposes, including astronomy, for high-power lasers, and for various sensor systems. In some applications the absorption and conduction of heat from the mirror—and the resultant distortions from resultant temperature increases—are indeed important. Table 1.62 lists the thermal expansion coefficients and the thermal conductivity of some candidate materials.

Silicon Carbide. Silicon carbide is often manufactured by hot pressing, but better uniformity and smaller grain size can be obtained by chemical vapor deposition (CVD) techniques. One example[151] obtained a density that differed insignificantly from theoretical. The thermal conductivity, specific heat, and thermal expansion for silicon carbide are shown in Figs. 1.26, 1.27, and 1.28. The hardness, measured by both Vickers and Knoop techniques, and with different loads, ranges from about 2400 to 2500 kg/mm^2. The fracture toughness done by Vickers was 3.2 to 3.5 MN m^{-1}. There is a reststrahlen reflectivity

Table 1.61 Density, Young's Modulus, and the Strength-to-Weight Ratio

Material	Specific Gravity	Young's Modulus, E ($g\ cm^{-1}\ s^{-2}$)	E/ρ ($cm^2\ s^{-2} \times 10^{12}$)	References
Alumina	3.85	3.5	0.91	
Aluminum	2.70	0.69	0.256	
Beryllium	1.82	2.8	1.54	
Fused silica	2.20	0.73	0.33	
Magnesium	1.74	0.45	0.26	
Pyrex	2.35	0.68	0.29	
SiC (CVD)	3.213	467 GPa		151
SXA				
ULE fused silica	2.21	0.68	0.32	
Zerodur				

Table 1.62 Thermal Expansion and Conductivity of Substrate Materials

Material	Expansion α ($10^6\ K^{-1}$)	Conductivity, k ($cal\ cm^{-1}\ s^{-1}\ K^{-1}$)	T	α/k	Specific Heat, c_p ($cal\ g^{-1}\ K^{-1}$)	References
Alumina	6.0	0.041		150		
Aluminum	24.0	0.53		45	0.22	*,[a] 152
Beryllium	12.0	0.38		32	0.45	*, 152
BK7	7					152
Brass	20					152
Copper	16.7					152
Fused silica	0.56	0.0033		170	0.18	*, 152
7940	0.472		26.0			153
Homosil	0.495		26.5			153
Magnesium	26.0	0.38		68		
Nickel, electroless	13–14.5					152
Pyrex	3.2	0.0027		1200		
SF6	8					152
SiC	0.6, 1.5, 5	30, 90 W/m K	−100, 0, 500, 75, 250			151
Silver	18					
Steel	12					
Superinvar	−0.200		27.5			153
SXA						
ULE	0.040	0.0031	26.5	11	0.18	153, 152
Zerodur	−0.040		15.5		0.20	153, 152

[a] * = no reference.

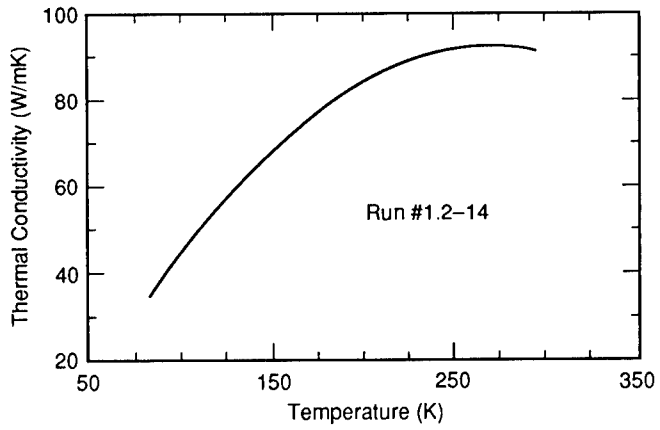


Fig. 1.26 Calculated thermal conductivity as a function of temperature for CVD SiC. Notice the conductivity appears to reach a maximum value at about 270 K (−3°C).

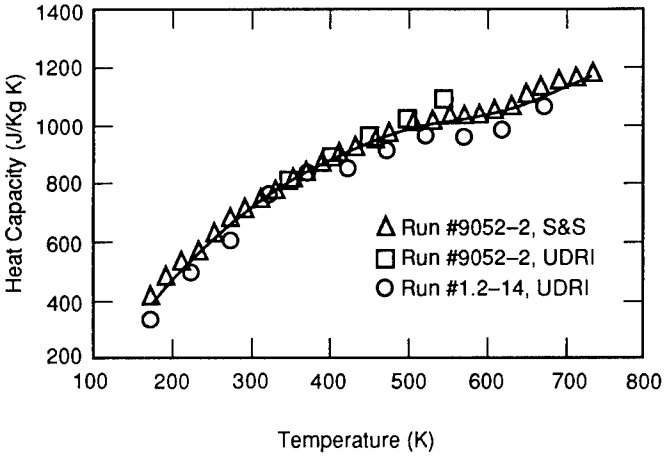


Fig. 1.27 Heat capacity of CVD SiC measured by (DSC). Measurements were performed at UDRI and Skinner and Sherman Laboratories (S&S) using the sapphire standard. Note good agreement in data at low temperatures and slight deviation observed at higher temperatures. The solid line drawn through points is a least-squares fifth-order regression fit to the data.

maximum of about 0.95 from 11 to 13 μm^3. Typical surfaces[151,154] have rms roughness values of about 1 nm, although one surface has been made about three times smoother. Curved mirrors as large as 40 cm in diameter with a silicon faceplate and areal density of 25 kg m^{-2} have been made.[154]

Fused Quartz. Generally, the term fused quartz indicates an amorphous form of silica, but one form is remelted crystals, while the other is a synthetic form generated by some high-temperature technique involving the constituents rather than the compound as base materials. Some terms are now commonly accepted for the four different types of this material. The first two are remelts; the second two are synthetics. Type 1, remelted in an electric furnace and repre-

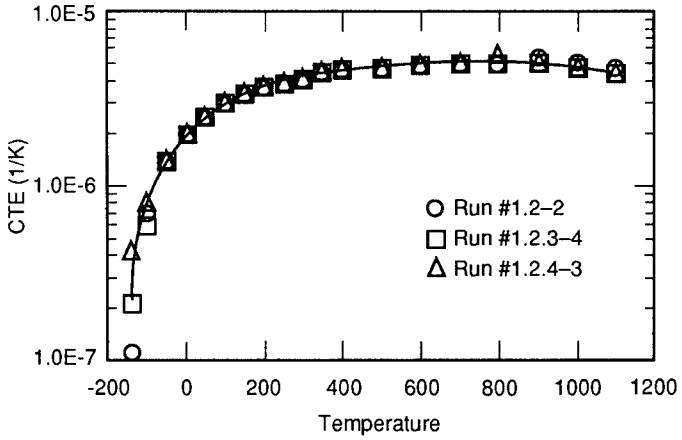


Fig. 1.28 Coefficient of thermal expansion (CTE) plotted versus temperature for various samples of CVD SiC. Material from Run 1.2–2 was produced in a small research CVD furnace; Run 1.2.3–4 in a pilot plant facility; and Run 1.2.4–3 in a manufacturing facility. The uncertainty in the measured value of CTE is $\pm 1 \times 10^{-7}\,K^{-1}$.

sented by Infrasil, is the lowest quality quartz with the highest content of inclusions and granularities. Type 2, remelted with an oxygen-hydrogen flame, is of higher quality and represented by Optosil, Homosil, and Ultrasil. Type 3 is made by hydrolyzing $SiCl_4$ in an oxygen-hydrogen flame. It is purer still and represented by Suprasil. It has little scatter (almost that of Rayleigh) but still retains a very strong OH absorption at about 2.7 μm. Type 4 uses an oxygen plasma burner (no hydrogen) and is represented by Suprasil-W. It seems clear that type 1 material is perfectly satisfactory for mirror blanks, although for the best scatter performance, type 1 might provide too many surface imperfections.

Beryllium. Much has been done to improve the performance of beryllium. It is extremely light and strong and relatively impervious to high-energy irradiation. However, it does not take a good polish on its own surface. For applications not requiring such performance, an electroless nickel coating is used. Its main requirements are that it adhere well, be smooth, and have an expansion coefficient that is close to that of beryllium. Nickel has been found to be the best material, but it is not perfect. Beryllium is highly toxic in a powder form. Extreme care must be taken when grinding and polishing it. The use of vacuum systems near all tools is commonplace, and facility approval is required.

Recent techniques for improving the uniformity of the material include the use of spherical particles in the compact and hot isostatic pressing (HIP) rather than a directional press. Some examples are given here of the scattering from materials that have been polished bare and that have had a sintered coating of beryllium on beryllium. The area is the subject of intense development, and results obtained after the publication of this handbook may be available. One example[155] of scattering from a beryllium mirror made by the HIP process is shown in Figure 1.29.

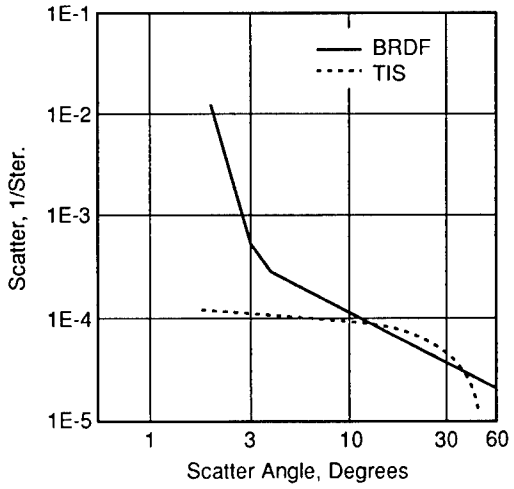


Fig. 1.29 The BRDF of a 50-cm-diam HIP beryllium mirror measured at 10.6 μm.

1.5 BLACKS DATA

Many different paints and finishes are used to control stray light. They are described by their directional-hemispherical emissivity or reflectivity and by their bidirectional reflectivity over spectral bands and spectrally. Bidirectional reflectivity is the best way of characterizing the optical properties of a black. It is defined as the radiance reflected in a specific direction to the incident irradiance. The units are reciprocal steradians. It is a function of both the angles of incidence and the angles of reflection (or scatter). The directional-hemispherical reflectance is the ratio of the power or flux density that is collected over the entire overlying hemisphere to the power or flux density that is incident on a sample. It is a function of only the angles of incidence. It is equivalent to the hemispherical-directional reflectivity by the Helmholtz reciprocity theorem. Many authors shorten the expression to simply *hemispherical reflectance*. An isotropic (Lambertian) reflector has a bidirectional reflectivity that is equal to $1/\pi$ times the hemispherical reflectivity. Some summary data[156] are given in Fig. 1.30, and the various materials are described in the following sections.

In addition to the optical properties, durability and outgassing are important properties of blacks, especially for space instrumentation. These properties are harder to quantify, but they are described qualitatively.

Several blacks have been measured spectrally.[157,158] The results are shown in Figs. 1.31, 1.32, and 1.33. The spectral absorption is quoted. This refers to the absorptivity, probably measured with a spectrometer. Therefore, it is a directional absorptivity. However, because the samples are close to Lambertian, they may be considered the one-complement of the hemispherical reflectivity, to a good approximation.

Additional data are given by Harris and Cuff[159] for acetylene black, Dupont flat black (unspecified), Eastman Kodak NOD-18, gold-black, lampblack "soot" and "lacquer," and for Globars, boron nitride, silicon carbide, zirconia, magnesia, alumina, aluminum, inconel, stainless steel, and for some enamels.[160]

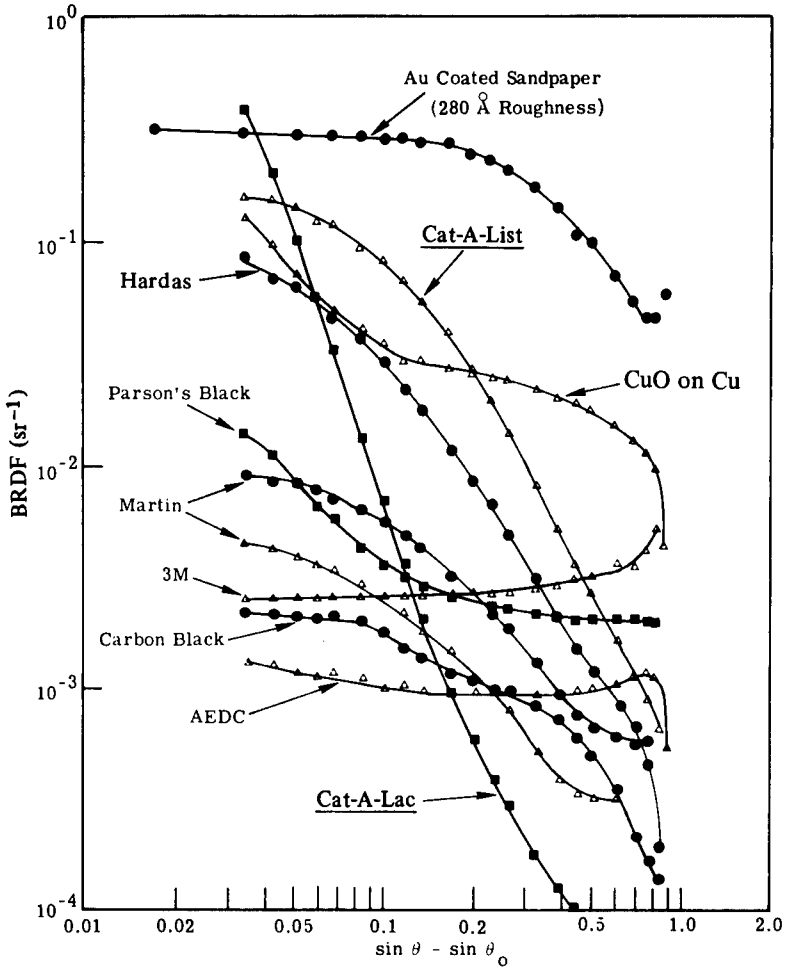


Fig. 1.30 The bidirectional reflectivity measured by the University of Arizona Optical Sciences Center with He-Ne laser radiation for Au-coated sandpaper, Cat-A-List, Hardas, CuO on Cu, Parson's black, Martin, 3M, carbon black, AEDC, and Cat-A-Lac. Cat-A-Lac® and Cat-A-List® are registered trademarks of Bostic-Finch, Inc. (AEDC = Arnold Engineering Development Co.)

Studies at long wavelengths have been carried out,[161] comparing Nextel Black Velvet, Herbert 100 2 E, Cornell black, ECP 2200 SiC, and Martin black. Specular reflectivities at 10.6 and 337 μm for the materials, respectively, are (in percent): 0.0011, 51.3; 0.0055, 0.3; 0.01, 33.5; 0.023, 45; and 0.003, 64.5. Outgassing tests[8] produced total mass loss and collected volatile condensed material measurements, again respectively and for each material in turn, as follows (in percent): 7, 0.18; 1.55, 0.34; 4.08, 0.08; 0.79, 0.01; and 0, 0. Other long-wave data have been presented by Smith[162] and Pompea et al.[163]

Chemglaze Z306. Chemglaze Z306 is a flat-black, oil-free polyurethane paint manufactured by Lord Corporation of Erie, Pennsylvania. It has been inves-

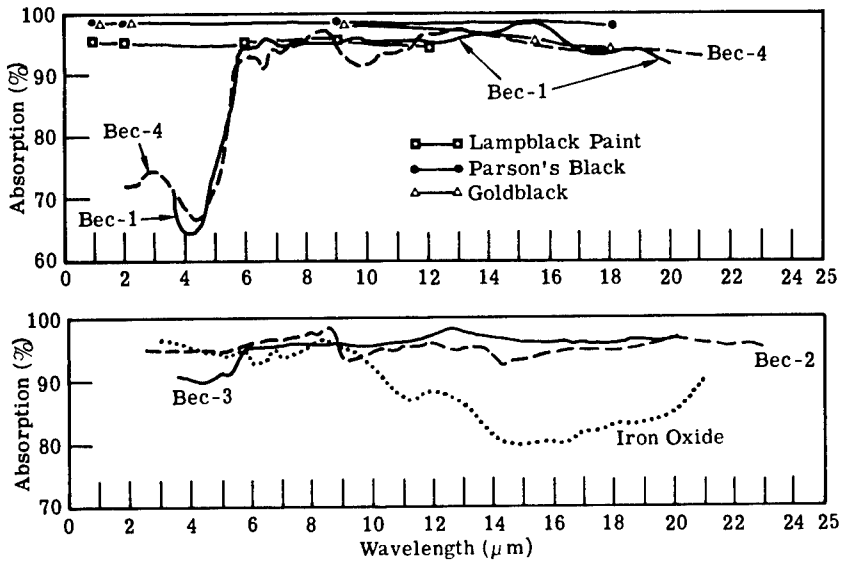


Fig. 1.31 The spectral absorption of lampblack paint, Parson's black, gold-black, and BEC-1, 2, 3, and 4. (BEC = Barnes Engineering Co.)

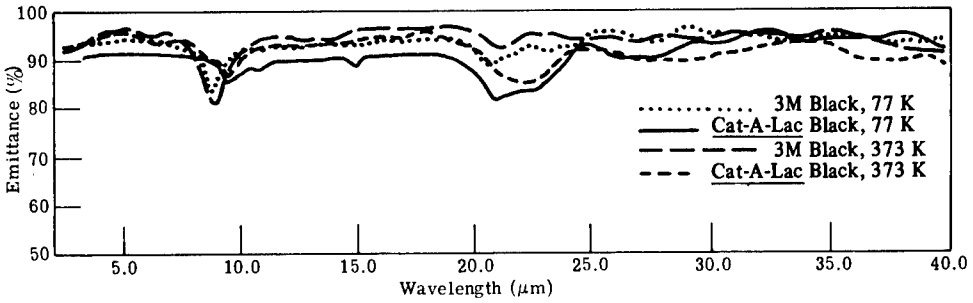


Fig. 1.32 The spectral absorption of 3M black and Cat-A-Lac black at 77 and 373 K. Cat-A-Lac® is a registered trademark of Bostic-Finch, Inc.

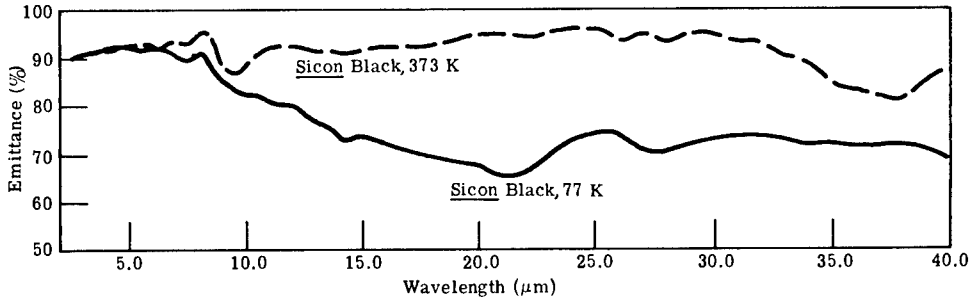


Fig. 1.33 The spectral absorption of Sicon black at 77 and 373 K. Sicon® is a registered trademark of Midland Industrial Finishes.

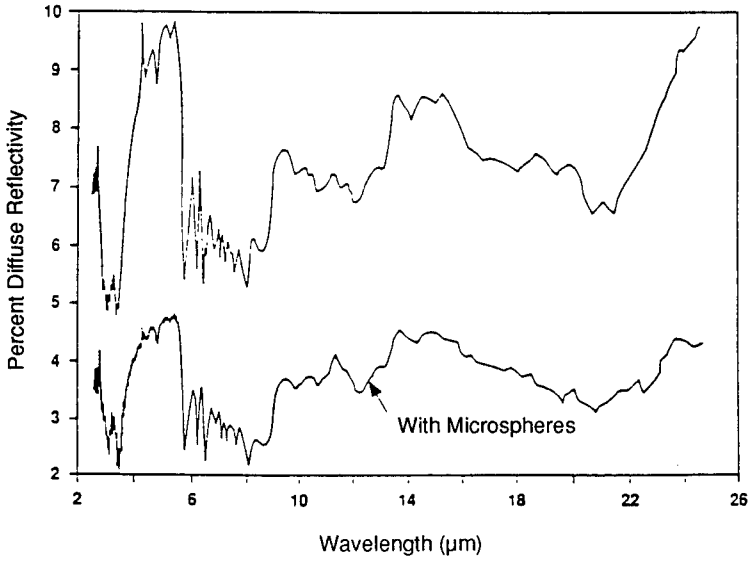


Fig. 1.34 The hemispherical reflectivity of two types of Chemglaze enamel.

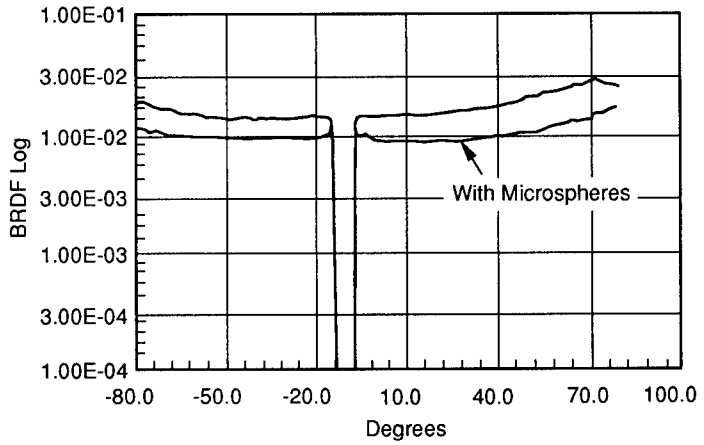


Fig. 1.35 The bidirectional reflectivity of Chemglaze at 0.6328 μm and for a 5-deg angle of incidence.

tigated in its standard form and with the inclusion of glass microspheres.[164] The ones reported here[165] were Scotchlite C15/250 soda lime borosilicate microspheres manufactured by 3M Industrial Specialties of St. Paul, Minnesota. An alternative is Q-CEL 2116 manufactured by PQ Corporation of Valley Forge, Pennsylvania. Outgassing is reduced considerably if the paint is first cured to the highest operational temperature or above, although the highest cure temperature is 373 K. Figure 1.34 shows the spectral distribution of hemispherical reflectivity for Chemglaze with and without the addition of microspheres.

The bidirectional reflectivities of the same materials are shown in Figs. 1.35 and 1.36 for a wavelength of 0.6328 μm, and for near-normal incidence and

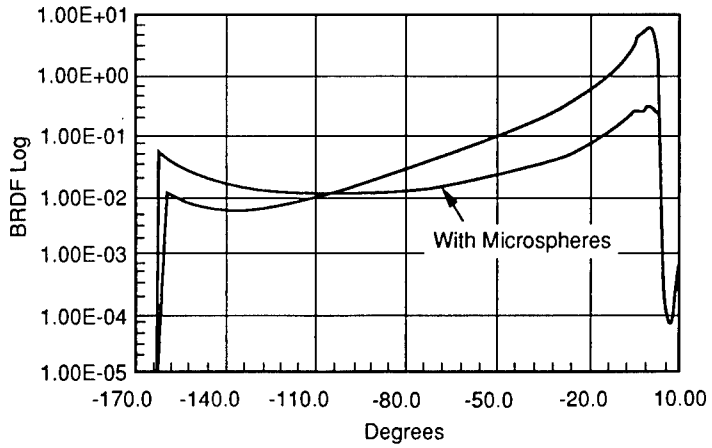


Fig. 1.36 The bidirectional reflectivity of Chemglaze for 0.6328 μm and an 80-deg angle of incidence.

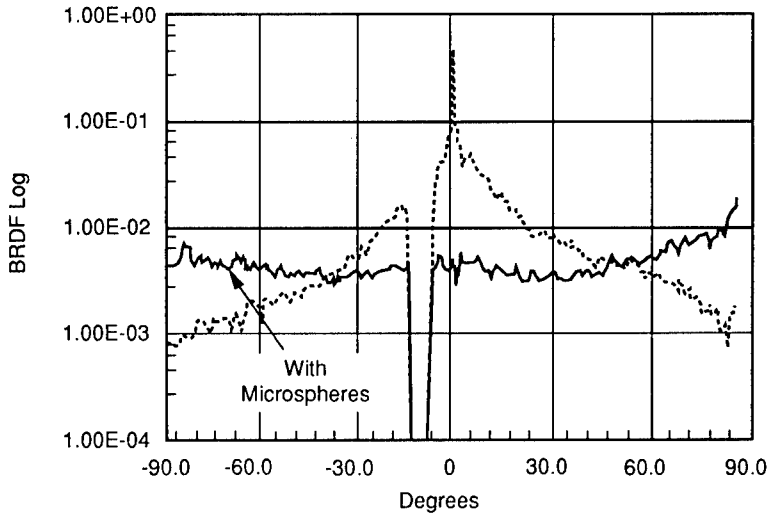


Fig. 1.37 The bidirectional reflectivity of Chemglaze for 10.6 μm and a 5-deg angle of incidence.

near-grazing incidence. The same information is given in the same manner for a wavelength of 10.6 μm in Figs. 1.37 and 1.38.

Parsons Optical Black. Parsons optical black[166] lacquer consists of an undercoating that is mostly carbon black dispersed in nitrocellulose and an overcoating of aniline black and ethyl cellulose. The undercoat is usually mixed with organic solvents and sprayed onto a clean surface. The topcoat is applied after drying with an airbrush as well. Parsons black appears visibly to be blacker than lampblack. It does, as do most of these blacks, have many deep pits of about 10 μm in the surface. Table 1.63 shows the hemispherical reflectivities for Parsons black and some close relatives for different wavelengths.

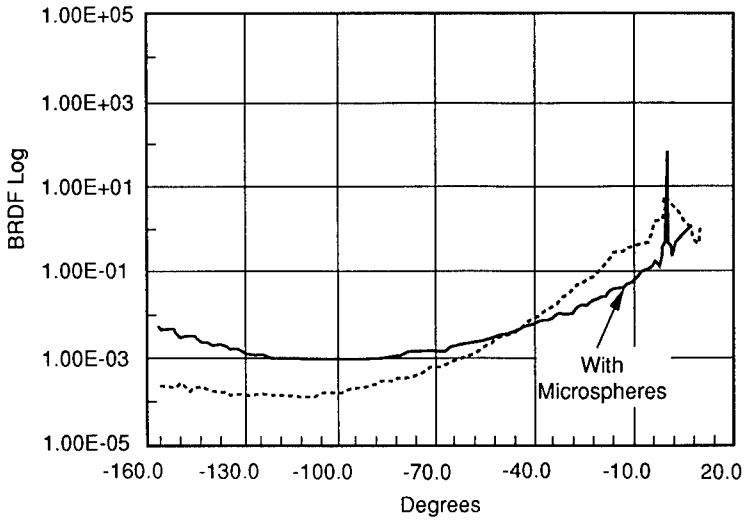


Fig. 1.38 The bidirectional reflectivity for Chemglaze for 10.6 μm and an 80-deg angle of incidence.

Table 1.63 Hemispherical Reflectivities of Some Blacks

Material/ Wavelength (μm)	0.254	0.6	0.7	0.95	4.4	8.8	12
Acetylene	0.009	0.011	0.010	0.004	0.007	0.012	0.03
Lamp	0.045	0.041	0.039	0.035	0.032	0.042	0.044
Lamp + Soot	0.014	0.016	0.016	0.010	0.010	0.013	0.057
Parsons		0.012	0.012	0.04		0.022	

Cat-A-Lac and Cat-A-List. These are paints made by the Bostic-Finch Company in Torrance, California.

Carbon Black. Carbon black is nothing more than black soot applied by a candle. It is easy to use and cheap, but it is not very durable, and it does outgas.

Martin Black. Martin black is a proprietary process of the Martin Company in Denver, Colorado. It is essentially a deep-etch anodic process on aluminum that provides a needle-like structure, which is blackened with a dye. It is quite black, but it is fragile and does outgas some. The surface is easily marred and degraded by a touch that crushes the needles. A severe shaking usually loosens some of the needles. A good process for a space application is to provide excessive shaking (and cleaning) prior to launch. Martin black, like most blacks, has a reflectivity that is a function of incidence angle; the reflectivity increases as the angle of incidence increases.

Black Velvet Nextel. Black Velvet Nextel, formerly manufactured by the 3M Corporation with a code of 401-C10, has been discontinued.

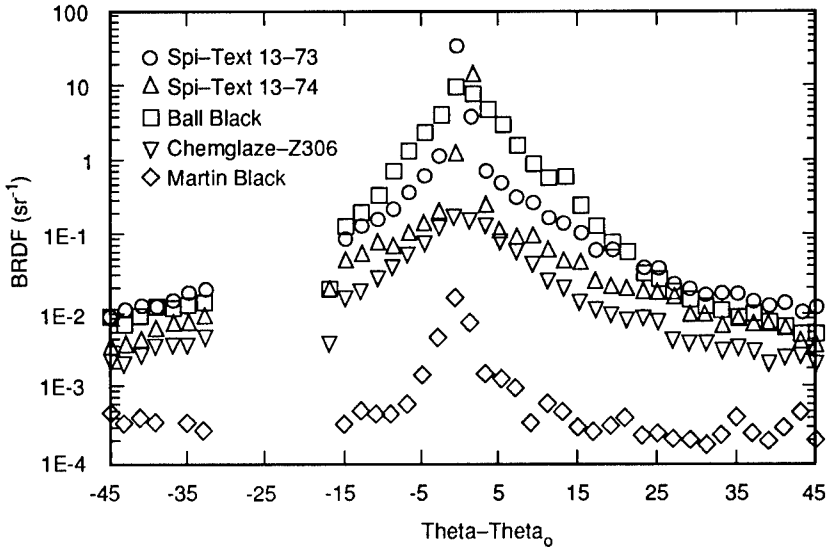


Fig. 1.39 Comparison of several blacks at 10.6 μm.

Cornell Black. Cornell black[167] is a mixture of carbon, vehicles, and small spherical balls.

Other Blacks. Many different formulations have been tried to obtain good, very diffuse black materials. The approaches are to use materials that do absorb and structures that are rough and porous so that the light is diffusely reflected and is reflected many times within the material. Some of these have been measured[168] including 401C10 black paints 1 and 2, silicon carbide foam and aluminum foam, as well as Spi-Text 13-73 and -74 and Ball black.[168] The Spi materials are experimental coatings from the Spire Company, Boston, Massachusetts. Ball black is made by the Ball Aerospace Company, Boulder, Colorado. A comparison of these with Z306 and Martin is given in Fig. 1.39.

References

1. W. L. Wolfe and G. J. Zissis, Eds., *The Infrared Handbook*, Environmental Research Institute of Michigan, Ann Arbor, MI (Revised 1985).
2. H. H. Li, "Refractive index of alkali halides and its wavelength and temperature derivatives," Center for Information and Numerical Data Analysis and Synthesis, Purdue University, West Lafayette, IN (May 1975).
3. S. S. Ballard, K. A. McCarthy, and W. L. Wolfe, *Optical Materials for Infrared Instrumentation*, University of Michigan, Ann Arbor, MI (1958).
4. W. L. Wolfe, S. S. Ballard, and K. A. McCarthy, "Refractive index of special crystals and certain glasses," Chap. 6b in *The Handbook of Optics*, Optical Society of America (1978).
5. L. W. Tilton and E. K. Plyler, "Refractivity of lithium fluoride with application to the calibration of spectrometers," *Journal of Research of the National Bureau of Standards* **47,** 25 (1951).
6. H. Harting, "The optical aspects of some crystals and their representation by the Hartmann equation," *Zeitschrift für Instrumentkunde* **63,** 125 (1943).

7. Z. Gyulai, "The dispersion of some alkali halides in the ultraviolet region," *Zeitschrift für Physik* **46,** 80 (1927).
8. M. Herzberger and C. D. Salzburg, "Refractive indices of infrared optical materials and color correction of infrared lenses," *Journal of the Optical Society of America* **52,** 420 (1962).
9. H. W. Hohls, "Dispersion and absorption of lithium fluoride and sodium fluoride in the infrared," *Annalen der Physik* **32,** 433 (1937).
10. F. F. Martens, "The dispersion of ultraviolet radiation," *Annals of Physics* **6,** 603 (1901).
11. F. Paschen, "The dispersion of rocksalt and sylvite in infrared," *Annals of Physics* **26,** 120 (1908).
12. H. Rubens and E. F. Nichols, "Experiments with heat radiation of long wavelengths," *Annals of Physical Chemistry* **60,** 418 (1897).
13. H. Rubens and A. Trowbridge, "Contribution to the knowledge on the refractive index and absorption of infrared radiation for rocksalt and sylvite," *Annals of Physical Chemistry* **60,** 724 (1897).
14. W. W. Coblentz, "Transmission and refraction data on standard lens and prism material with special reference to infrared spectrometry," *Journal of the Optical Society of America* **4,** 432 (1920).
15. R. J. Spindler and W. S. Rodney, "Refractivity of potassium bromide for visible wavelengths," *Journal of Research of the National Bureau of Standards* **49,** 253 (1952).
16. R. E. Stephens, E. K. Plyler, W. S. Rodney, and R. J. Spindler, "Refractive index of potassium bromide for infrared radiant energy," *Journal of the Optical Society of America* **43,** 110 (1953).
17. J. W. Forrest, "Refractive index values for potassium bromide," *Journal of the Optical Society of America* **32,** 382 (1942).
18. E. Gundelach, "The dispersion of KBr crystal in the infrared," *Zeitschrift für Physik* **66,** 775 (1930).
19. E. Liebreich, "The optical temperature coefficient for rocksalt, sylvite and fluoride in the region of lower temperatures," *Verhandlungen der Deutchen Physikalische Gesellschaft* **13,** 700 (1911).
20. O. A. Koslovskii and L. N. Ustimenko, "Measurement of the temperature coefficient of refractive index of infrared materials using a CO2 laser," *Optics and Spectroscopy* **33,** 430 (1972).
21. K. Korth, "Dispersion measurements on potassium bromide and potassium iodide in the infrared region," *Zeitschrift für Physik* **84,** 677 (1933).
22. H. B. Briggs, "Optical effects in bulk silicon and germanium," *Physical Review* **77,** 287 (1950).
23. C. D. Salzburg and J. J. Villa, "Infrared refractive indexes of silicon germanium and modified selenium glass," *Journal of the Optical Society of America* **47,** 244 (1957); and "Corrections," *Journal of the Optical Society of America* **48,** 579 (1958).
24. H. H. Li, "Refractive index of silicon and germanium and its wavelength and temperature derivative," *Journal of Physical and Chemical Reference Data* **9,** 561 (1980).
25. H. W. Icenogle, B. C. Platt, and W. L. Wolfe, "Refractive indexes and temperature coefficients of germanium and silicon," *Applied Optics* **15,** 2348 (1976).
26. R. P. Edwin, M. T. Dudermel, and M. Lamare, "Refractive index measurements of ten germanium samples," *Applied Optics* **21,** 878 (1982).
27. W. Primak, "The refractive index of silicon," *Applied Optics* **10,** 759 (1971).
28. M. Cardona, W. Paul, and H. Brooks, "Dielectric constant of germanium and silicon as a function of volume," *Physics and Chemistry of Solids* **8,** 204 (1959).
29. F. Lukes, "The temperature-dependence of the refractive index of silicon," *Physics and Chemistry of Solids* **11,** 342 (1959).
30. W. L. Wolfe and R. Korniski, "Refractive index of Irtran2 as a function of wavelength and temperature," *Applied Optics* **17,** 1547 (1978).
31. A. Feldman, D. Horowitz, R. M. Waxler, and M. J. Dodge, "Optical materials characterization," *National Bureau of Standards Technical Notes* **993,** 63 (1979).
32. M. Mell, "Refraction and absorption of light in ZnS at temperatures up to 700C," *Zeitschrift für Physik* **16,** 244 (1923).
33. J. R. DeVore, "Refractive indices of rutile and sphalerite," *Journal of the Optical Society of America* **41,** 416 (1951).

34. W. L. Bond, "Measurements of the refractive indices of several crystals," *Journal of Applied Physics* **36,** 1674 (1965).

35. A. R. Hilton and C. E. Jones, "The thermal change in the nondispersive refractive index of optical materials," *Applied Optics* **6,** 1513 (1967).

36. D. T. F. Marple, "Refractive index of ZnSe, ZnTe and CdTe," *Journal of Applied Physics* **35,** 539 (1964).

37. W. R. Rambauske, "Optical dispersion of zinc selenide," *Journal of Applied Physics* **35,** 2958 (1964).

38. J. A. Wunderlich and L. G. deShazer, "Visible optical isolator using ZnSe," *Applied Optics* **6,** 1513 (1977).

39. C. J. C. Thompson, A. G. DeBell, and W. L. Wolfe, "Refractive index of ZnSe at 3.85um and 10.6um from 80K to 300K," *Applied Optics* **18,** 2085 (1979).

40. B. Seraphin and H. Bennett, *Optical Constants in Semiconductors and Semimetals,* Vol. 3, R. Willardson and A. Beer, Eds., Academic Press, New York (1967).

41. L. Barcus, A. Perlmutter, J. Callaway, "Effective mass of electrons in gallium arsenide," *Physical Review* **111,** 167 (1958).

42. S. Czyzak, W. Baker, R. Crane, J. Howe, "Refractive index of single synthetic zinc sulfide and cadmium sulfide crystals," *Journal of the Optical Society of America* **47,** 240 (1957).

43. W. Rodney, I. Malitson, T. King, "Refractive index of arsenic trisulfide," *Journal of the Optical Society of America* **48,** 633 (1958).

44. Technical Data Sheets, AMTIR Inc., Garland, TX.

45. A. R. Hilton, "Precise refractive index measurements of infrared materials," *Proceedings of the SPIE* **1307,** 516 (1990).

46. W. L. Wolfe, "The status and needs of infrared optical property information for optical designers," *Proceedings of the SPIE* **1354,** 696–741 (1990).

47. I. Malitson, "Interspecimen comparison of the refractive index of fused silica," *Journal of the Optical Society of America* **55,** 1205 (1965).

48. C. Salzberg, J. Villa, "Infrared refractive indexes of silicon, germanium and modified selenium glass," *Journal of the Optical Society of America* **47,** 244 (1957).

49. Condensed Data for Kodak Irtran Infrared Optical Materials, Publication No. U-71, Eastman Kodak Company, Rochester, NY (Revised May 1975).

50. F. Micheli, *Annalen der Physik* **4,** 772 (1902).

51. I. Malitson, "A redetermination of some optical properties of calcium fluoride," *Applied Optics* **2,** 1103 (1963).

52. R. Stephens, I. Malitson, *Journal of Research of the National Bureau of Standards* **49,** 249 (1952).

53. I. Malitson, "Refraction and dispersion of synthetic sapphire," *Journal of the Optical Society of America* **52,** 1377 (1962).

54. I. Malitson, "Refractive properties of barium fluoride," *Journal of the Optical Society of America* **54,** 628 (1964).

55. S. Rosch, "Die Optik des Fabulit, die Farbe des Brewster Winkels und das Farbspielmoment," *Optica Acta* **12,** 253 (1965).

56. A. Eucken, A. Buchner, *Zeitschrift für Physikalische Chemie* **27B,** 321 (1934).

57. A. R. von Hippel (Ed.), *Dielectric Materials and Applications,* John Wiley & Sons, New York (1954).

58. Data Sheets, American Optical Company.

59. D. E. Gray, *American Institute of Physics Handbook,* McGraw-Hill, New York (1957).

60. *Handbook of Chemistry and Physics,* 39th ed., Chemical Rubber Publishing Company (1957).

61. D. deNobel and D. Huffman, "The dielectric constant of CdTe," *Physica* **22,** 252 (1956).

62. Quarterly Progress Report #4 on AF33(616)78, University of Pennsylvania (June 1953).

63. W. McKusick, W. Parsons, T. Hale, and R. Krueger, Eastman Kodak, private communication (1959).

64. A. Duncanson and R. Stephenson, "Some properties of magnesium fluoride crystallized from the melt," *Proceedings of the Physical Society, London* **72,** 1001 (1958).

65. L. Combes, S. Ballard, and K. McCarthy, "Mechanical and thermal properties of certain

optical crystallized materials," *Journal of the Optical Society of America* **41,** 215 (1951).

66. W. McKusick, Eastman Kodak Company, private communication (1959).
67. L. Combes, Tufts University, private communication (1953).
68. Technical Data Sheets, AMTIR Inc., Garland, TX.
69. S. Ballard, L. Combes, and K. McCarthy, "A comparison of the physical properties of barium fluoride and calcium fluoride," *Journal of the Optical Society of America* **42,** 684 (1952).
70. W. Forsythe, *Smithsonian Data Tables,* 9th ed., Smithsonian Institution, Washington, DC, p. 228.
71. Technical Data Sheets, R. Kebler, Linde Air Products.
72. S. Ballard, L. Combes, and K. McCarthy, "Gross variations in the physical properties of synthetic crystalline lithium fluoride," *Journal of the Optical Society of America* **41,** 772 (1951).
73. Technical Data Sheets, Texas Instruments, Garland, TX.
74. S. Levin, N. Field, F. Plock, and L. Merker, "Some optical properties of strontium titanate crystal," *Journal of the Optical Society of America* **45,** 737 (1955).
75. H. Kremers, "Optical silver chloride," *Journal of the Optical Society of America* **37,** 337 (1947).
76. A. Adamiano, *Journal of Physical Chemistry* **61,** 1253 (1957).
77. J. Mellor, *A Comprehensive Treatment on Inorganic and Theoretical Chemistry,* Vol. 11, Longmans Green and Company, London (1931).
78. H. Welker and H. Weiss, Vol. 3 in *Solid State Physics,* F. Seitz and D. Turnbull, Eds., Academic Press, New York (1956).
79. H. Welker, "Semiconducting intermetallic compounds," *Physica* **20,** 893 (1954).
80. G. Pearson and W. Brattain, *Proceedings of the Institute of Radio Engineers* **43,** 1794 (1955).
81. R. Breckenridge, R. Blaut, W. Hosler, H. Frederikse, J. Becher, and W. Oshinsky, "Electrical and optical properties of intermetallic compounds: indium antimonide," *Physical Review* 96, 571 (1954).
82. A. Smakula, MIT, private communication (1953).
83. *Proceedings of the Conference on Infrared Optical Materials, Filters and Films,* Engineering Research and Development Laboratories (1955).
84. W. DeSorbo, *Journal of Chemical Physics* **21,** 876 (1953).
85. R. Sosman, *The Properties of Silica,* The Chemical Catalog Company **(city?)** (1927).
86. R. Machol and E. Westrum, "The triple-point temperature of tellurium," *Journal of Physical Chemistry* **62,** 361 (1958).
87. Data Sheets, American Optical Company.
88. K. McCarthy and S. Ballard, "New data on the thermal conductivity of optical crystals," *Journal of the Optical Society of America* **41,** 1062 (1951).
89. S. Ballard, K. McCarthy, E. Bray, and W. Little, Tufts University Measurements (1954).
90. D. E. Gray, *American Institute of Physics Handbook,* McGraw-Hill, New York (1957).
91. K. McCarthy and S. Ballard, "Thermal conductivity of germanium at ambient temperatures," *Physical Review* **99,** 1104 (1955).
92. S. Ballard, K. McCarthy, and W. Davis, *Review of Scientific Instruments* **21,** 905 (1950).
93. Landolt Bornstein, *Physikalische-Chemische Tabellen,* Vols. I–IV, Springer-Verlag, Berlin, p. 711 (1927).
94. R. Srinivasan, *Journal of Indian Instrument Society* **37,** 200 (1955).
95. S. Ballard, L. Combes, and K. McCarthy, "Gross variations in the physical properties of synthetic crystal lithium fluoride," *Journal of the Optical Society of America* **41,** 772 (1951).
96. A. Stuckes and R. Chasmar, "Report of the meeting on semiconductors," *Proceedings of the Physical Society, London* **123** (1956).
97. *International Critical Tables,* McGraw-Hill, New York (1929).
98. M. Straumanis and E. Aka, "Lattice parameter coefficients of thermal expansion and atomic weights of purest silicon and germanium," *Journal of Applied Physics* **23,** 330 (1952).
99. L. Combes, S. Ballard and K. McCarthy, "Mechanical and thermal properties of certain optical crystalline materials," *Journal of the Optical Society of America* **41,** 215 (1951).

100. Technical Data Sheets, R. Kebler, Linde Air Products Company.
101. R. Stow, private communication.
102. A. von Hippel, Ed., *Dielectric Materials and Applications*, John Wiley & Sons, New York (1954).
103. H. Welker and H. Weiss, Vol. 3 in *Solid State Physics*, F. Seitz and D. Turnbull, Eds., Academic Press, New York (1956).
104. R. Srinivasan, *Journal of Indian Instrument Society* **A37,** 200 (1955); see also *International Critical Tables III*, p. 43, McGraw-Hill, New York (1929).
105. W. A. Wooster, *A Textbook on Crystal Physics*, University Press, Cambridge, MA (1935).
106. *Handbook of Chemistry and Physics* 39th ed., Chemical Rubber Publishing Company (1957).
107. J. Saunders, *Journal of Research of the National Bureau of Standards* **28,** 51 (1942).
108. A. Linz, Data Sheets, The National Lead Company.
109. G. Jones and F. Jelen, *Journal of the American Ceramic Society* **57,** 2532 (1935).
110. Technical Data Sheets, AMTIR Inc., Garland, TX.
111. P. Bridgman, *Proceedings of the American Academy of Arts and Sciences* **74,** 21 (1940).
112. Technical Data Sheets, R. Kebler, Linde Air Products Company.
113. Data Sheets, American Optical Company.
114. *Handbook of Chemistry and Physics* 39th ed., Chemical Rubber Publishing Company (1957).
115. A. Smakula and V. Sils, "Densities and imperfections of single crystals," *Physical Review* **99,** 1744 (1955).
116. A. Smakula, "Physical Properties of Optical Crystals with Special Reference to Infrared," Defense Technical Information Center Number 111052. U.S. Department of Commerce (1952).
117. A. Smakula, "Resolving power of prisms made from homogeneous thallium bromide crystals," *Journal of the Optical Society of America* **45,** 1086 (1955).
118. C. Hutchison and H. Johnston, *Journal of the American Chemical Society* **62,** 3165 (1940).
119. H. Johnston and C. Hutchison, *Physical Review* **62,** 32 (1942).
120. D. Hughes and C. Maurette, "Dynamic elastic moduli of iron aluminum and fused quartz," *Journal of Applied Physics* **27,** 1184 (1956).
121. D. Arenberg, *Journal of Applied Physics* **21,** 941 (1950).
122. W. Bond, W. Mason, and H. McSkimin, "Elastic and electromechanical coupling coefficients in single-crystal barium titanate," *Physical Review* **82,** 442 (1951).
123. Bhagavantan, *Proceedings of the Indian Academy of Science* **A41** (1955).
124. W. Voigt, *Lehrbuch der Kristallphysik*, Edwards Brothers, Ann Arbor, MI (1946).
125. J. Bhimasenachar, *Proceedings of the Indian Academy of Science* **A22,** 203 (1955).
126. D. Bolef and M. Menes, *Bulletin of the American Physical Society II* **5,** 169 (1960).
127. D. Bolef and M. Menes, *Bulletin of the American Physical Society II* **4,** 427 (1959).
128. H. Frederiske, Syracuse University, private communication (1952).
129. H. McSkimin, "Measurement of elastic constants at low temperature by means of ultrasonic waves—data for silicon and germanium single crystals, and for fused silica," *Journal of Applied Physics* **24,** 988 (1953).
130. H. Huntington, Vol. 7 in *Solid State Physics*, F. Seitz and D. Turnbull, Eds., Academic Press, New York (1958).
131. J. Galt, "Mechanical properties of NaCl, KBr, KCl," *Physical Review* **73,** 1460 (1948).
132. H. Huntington, Vol. 7 in *Solid State Physics*, F. Seitz and D. Turnbull, Eds., Academic Press, New York (1958).
133. D. Arenberg, Naval Research Report (1948).
134. H. Huntington, "Ultrasonic measurements on single crystals," *Physical Review* **72,** 321 (1947).
135. M. Durand, "The temperature variation of the elastic moduli of NaCl, KCl and MgO," *Physical Review* **50,** 449 (1936).
136. W. Binnie, "Calculation of the mean Debye temperature of cubic crystals," *Physical Review* **103,** 579 (1957).
137. L. Combes, S. Ballard, K. McCarthy, "Mechanical and thermal properties of certain optical crystalline materials," *Journal of the Optical Society of America* **41,** 215 (1951).

138. Technical Data Sheets, R. Kebler, Linde Air Products Company.

139. Technical Data Sheets, American Optical Company.

140. S. Ballard, L. Combes, and K. McCarthy, "A comparison of the physical properties of barium fluoride and calcium fluoride," *Journal of the Optical Society of America* **42,** 684 (1952).

141. C. Hilsum, Royal Radar Establishment, Malvern, UK, private communication (1952).

142. D. E. Gray, *American Institute of Physics Handbook*, pp. 3-96, McGraw-Hill, New York (1957).

143. S. Ballard, L. Combes, and K. McCarthy, "Gross variations in the physical properties of synthetic crystalline lithium fluoride," *Journal of the Optical Society of America* **41,** 772 (1951).

144. T. Shilliday, Battelle, Columbus, OH, private communication (1957).

145. G. Pearson, W. Brattain, *Proceedings of the Institute of Radio Engineers* **43,** 1794 (1955).

146. H. Huntington, Vol. 7 in *Solid State Physics*, F. Seitz and D. Turnbull, Eds., Academic Press, New York (1958).

147. *International Critical Tables*, McGraw-Hill, New York (1929).

148. R. Sosman, *The Properties of Silica*, The Chemical Catalog Company, New York (1927).

149. H. McSkimin, "Measurement of elastic constants at low temperature by means of ultrasonic waves—data for silicon and germanium single crystals, and for fused silica," *Journal of Applied Physics* **24,** 988 (1953).

150. G. Hass and A. Turner, *Ergebnisse der Hochvakuumtechnik und der Physik duenner Schichten*, M. Auwater, Ed., Wissenschaftliche Verlagsgesellschaft (1957).

151. M. Pickering, R. Taylor, and J. Keeley, "Chemically vapor deposited silicon carbide (SiC) for optical applications," *Proceedings SPIE* **1118,** 2 (1989).

152. W. English, "Quartz glass for space optical applications," *Proceedings of the SPIE* **1118,** 42 (1989).

153. J. Berthold, "Dimensional Stability of Low Expansivity Materials—Time Dependent Changes in Optical-Contact Interfaces and Phase Shifts on Reflection from Multilayer Dielectrics," PhD dissertation, University of Arizona (1976).

154. J. Goela and R. Taylor, "Large scale fabrication of lightweight Si/SiC LIDAR mirrors," *Proceedings of the SPIE* **1118,** 14 (1989); and J. Goela and R. Taylor, private communication (1989).

155. E. Gossett, J. Marder, R. Kendrick, and O. Cross, "Evaluation of hot isostatic pressed beryllium for low scatter cryogenic optics," *Proceedings of the SPIE* **1118,** 50 (1989).

156. T. Heslin, J. Heaney, and M. Harper, NASA Technical Note D7643, Goddard Space Flight Center, Beltsville, MD (May 1974).

157. D. Stierwalt, "Infrared spectral emittance measurements of optical material," *Applied Optics* **5,** 1911 (1966).

158. D. Stierwalt, J. Bernstein, and D. Kirk, "Measurement of the infrared spectral absorptance of optical materials," *Applied Optics* **2,** 1169 (1963).

159. L. Harris and K. Cuff, "Reflectance of goldblack deposits and some other materials of low reflectance from 254 mu to 100 mu, the scattering-unit-size in goldblack deposits," *Journal of the Optical Society of America* **46,** 160 (1956).

160. H. Blau, E. Chaffee, J. Jasperse, and W. Martin, *High Temperature Thermal Properties of Materials*, Arthur D. Little, Cambridge, MA (1960).

161. S. Ungar, J. Mangin, M. Lutz, G. Jeandel, and B. Wyncke, "Infrared black paints for room and cryogenic temperatures," *Proceedings of the SPIE* **1157,** 369 (1989).

162. S. Smith, "Specular reflectance of optical black coatings in the far infrared," *Applied Optics* **23,** 2311 (1984).

163. S. Pompea, D. Bergener, D. Shepard, S. Russak, and W. Wolfe, "Reflectance measurements on an improved optical black for stray light rejection from 0.3 to 500 μm," *Optical Engineering* **23**(2), 149 (1984); D. Bergener, S. Pompea, D. Shepard, and R. Breault, "Stray light rejection performance of SIRTF: a comparison," *Proceedings of the SPIE* **511,** 65 (1984).

164. W. Boerum, "Specifications for the Application of Chemglaze Z306 Coatings Containing Microballoons," IUE 701-73-010, Goddard Space Flight Center.

165. A. James, "Z306 black paint measurements," *Proceedings of the SPIE* **1331,** 299 (1990).

166. L. Harris, *The Optical Properties of Metal Blacks and Carbon Blacks*, The Eppley Foundation and MIT, Newport, RI (1967).

167. J. Pipper and J. Houck, "Black paints for far infrared cryogenics use," *Applied Optics* **10,** 567 (1971).

168. L. Scherr, J. Schmidt, and K. Sorensen, "BRDF of silicon carbide and aluminum foam compared to black paint at 3.39 microns," *Proceedings of the SPIE* **1165,** 204 (1989).

169. A. Lompado, B. Murray, J. Wollam, and J. Meroth, "Characterization of optical baffle materials," *Proceedings of the SPIE* **1165,** 212 (1989).

CHAPTER 2

Optical Design

Warren J. Smith
Kaiser Electro-Optics
Carlsbad, California

CONTENTS

This chapter has been reprinted from *The Infrared Handbook,* W. L. Wolfe and G. J. Zissis, Eds., Environmental Research Institute of Michigan, Ann Arbor, MI (revised 1985), with the exception of the bibliography, which has been updated to include current literature.

2.1 INTRODUCTION

This chapter deals with optical design, lenses, mirrors, and combinations of these elements. The media of propagation are always considered isotropic. Unless otherwise indicated, optical systems are assumed to be axially symmetrical, that is, to be composed of surfaces that are figures of rotation, whose axes of symmetry coincide with the optical axis.

In general, where both uppercase and lowercase symbols are used for the same quantity, the uppercase symbol represents the trigonometric (or "exact") quantity; the lowercase symbol represents the corresponding paraxial, or first-order, value. Primed symbols refer to quantities after refraction (or reflection) by a surface or by a lens, or to quantities associated with the image. Subscripts are used to indicate the surface or element with which a symbol is associated, or to indicate a particular ray. Table 2.1 lists most of the symbols, nomenclature, and units used in this chapter. Symbols that appear as parameters in a particular calculation, however, have been omitted.

2.2 DEFINITIONS

Axis, Optical. The common axis of symmetry of an optical system; in an element, the line between the centers of curvature of the two (axially symmetric) surfaces.

Eye Relief. In a visual instrument (e.g., telescope or microscope), the distance from the last optical surface to the (usually external) exit pupil, thus the clearance or "relief" between the instrument and the eye.

Invariant, Optical. When two unrelated (i.e., with different axial intercepts) paraxial rays are traced through an optical system, their data are sufficient to define the system completely. If the ray height and ray slope data of the marginal and principal rays are identified by y, u, and y_p, u_p, the expression

$$\mathcal{J} = n(y_p u - y u_p) \tag{2.1}$$

(where n is the index of refraction of the medium) is invariant across any surface or space of the optical system. At an object or image plane (where $y = 0$ and $y_p = h$), the invariant reduces to the Lagrange invariant:

$$\mathcal{J} = hnu = h'n'u' \tag{2.2}$$

where h and h' are the object and image height, respectively, and nu and $n'u'$ are the ray slope-index products at the object and image, respectively.

Magnification, Lateral or Linear. The ratio between the size of an image (measured perpendicular to the optical axis) and the size of the corresponding (conjugate) object.

Magnification, Longitudinal. The ratio between the length or depth (measured along the optical axis) of an image and the length of the corresponding object.

Magnification, Angular. The ratio of the angular size of an image (produced by an afocal optical system) to the angular size of the corresponding object.

Table 2.1 Symbols, Nomenclature, and Units

Symbol	*Nomenclature*	*Unit*
A	Area	m^2
$A(y,z)$	The complex amplitude of the wavefront emerging from the optical system	$V\ m^{-1}$
AIM	Aerial image modulation	–
A_n	Coefficient of the nth order aspheric deformation term	$m^{(n-1)}$
B	The blur diameter	m
$B(u,v)$	Amplitude factor proportional to the square root of the flux density at point (u,v) in the pupil	–
BFL, bfl	Back focal length–the distance from the back vertex of the optical system to the back (or second) focal point (See Figure 2-2.)	m
C	Surface curvature, reciprocal of surface radius when subscripted for a particular surface, also total curvature of an element	m^{-1}
CC	Coma contribution (TOA)	m
C_e	Equivalent curvature	m^{-1}
Coma_s	Coma, sagittal	m
Coma_t	Coma, tangential	m
D	Diameter	m
D_e	Diameter for $(1/e^2)$ of optical beam	m
$D_{j,k}$	Distance along ray from surface j to surface k	m
DC	Distortion contribution (TOA)	m
d	Axial distance, especially between (thin) elements, or between principal points	m
E_v	Illuminance	$lm\ m^{-2}$
F, f, EFL, efl	Focal length (effective) or effective focal length–the distance from the second principal point to the back (or second) focal point. Also, the distance from the front (or first) focal point to the first principal point. (See Figure 2-2.)	m
FFL, ffl	Front focal length–the distance from the front vertex of optical system to the front focal point. (See Figure 2-2.)	m
$F/\#$	Relative aperture, speed. The ratio of focal length to entrance pupil diameter. If the object is at infinity, $(F/\#) = [2(NA)]^{-1}$	–
f, f_y, f_z	Spatial frequency in cycles per unit length in the direction indicated	m^{-1}
H, h	Image height	m
$H(\omega_y, \omega_z)$	Optical transfer function	–
$h(y,z)$	Point-spread function (impulse response)	–
I	Angle of incidence which is equal to (minus) the angle of reflection (See Figure 2-1 .)	rad
$I(\omega_y, \omega_z)$	Spatial frequency spectrum of an image	–

(continued)

Table 2.1 *(continued)*

Symbol	*Nomenclature*	*Unit*
I'	Angle of refraction, defined by Snell's Law	rad
$\mathcal{J}$	Optical invariant	rad m
L	Radiance	W m^{-2} sr^{-1}
L, l	Distance from surface to intersection of ray with axis, before refraction; L', l' same after refraction	m
LA	Spherical aberration	m
LA_m	Marginal spherical aberration (longitudinal)	m
LA_z	Zonal spherical aberration (longitudinal)	m
MTF	Modulation transfer function	m
M_j	Distance (vector) from vertex of surface j (perpendicular) to ray	m
M_x	The x component of M_j	m
m	Magnification, lateral	–
m_t	Magnification, longitudinal	–
NA	Numerical aperture, given by $n \sin U$ where n is the final index in an optical system and U is the slope angle of the axial ray at the image. If the object is at infinity, (NA) = $[2(F/\#)]^{-1}$	–
n	Index of refraction	–
$O(\omega_y, \omega_z)$	Spatial frequency spectrum of an object	–
OPD	Optical path difference	m
OSC	Offense against the sine condition	–
OTF	Optical transfer function	–
P_1, P_2	First and second principal points, respectively	–
p (subscript)	Used to denote a principal or chief ray	–
Q	Stop-shift ratio	–
R	Radius of curvature	m
r	Semi-aperture = $(y^2 + z^2)^{1/2}$	m
S	Nominal distance at which a system is focused	m
s	Distance, first principal point to object	m
s'	Distance, second principal point to image	m
TA	Transverse aberration	m
TAC	Transverse version of aberration contribution for astigmatism	m
TchA	Lateral chromatic aberration, chromatic difference of image size	m
TchC	Lateral chromatic aberration contribution	m
TLchC	Transverse version of aberration contribution for axial color	m
$T/\#$	The speed of a lens, taking its transmission into account $(T/\#) = (F/\#)$ (transmission)$^{-1/2}$	–
TOA	Third-order aberration	–
TPC	Transverse version of aberration contribution for Petzval curvature	m

(continued)

Table 2.1 *(continued)*

Symbol	*Nomenclature*	*Unit*
TSC	Transverse version of aberration contribution for spherical aberration	m
t	Thickness; axial spacing between surfaces	m
U,u	Ray-slope angle, before refraction; U', u' same after refraction	rad
u,v	Spatial coordinates in exit pupil	m
V	Abbe number; reciprocal relative dispersion	–
v,v'	u/y and u'/y	m^{-1}
W_{ABC}	Numerical coefficient in the wave-aberration polynomial	–
W	Semi-diameter of a Gaussian beam (e^{-2} points); also, width of slit	m
x	Distance from first focal point to object	m
x'	Distance from second focal point to image	m
x_p	The longitudinal curvature (sag) of the image field: Petzval	m
x_s	The longitudinal curvature (sag) of the image field: sagittal	m
x_t	The longitudinal curvature (sag) of the image field: tangential	m
X,Y,Z	Direction cosines	–
x,y,z	Coordinate system; x is the optical axis; x and y define the meridional plane. The origin is at the vertex of the surface	m
Y,y	Height of intersection of ray with surface	m
β	Angular diameter of image blur spot	rad
$\Delta\phi$	Optical phase difference	m
δ, or Δ	The change in any quantity, such as Δn, δn and δs	–
δS	The longitudinal distance from the position of best focus	m
λ	Wavelength	μm
ν	Frequency of radiation	sec^{-1}
$\tilde{\nu}$	Wavenumber of radiation	cm^{-1} or wavenumber
ρ	Reflectivity	–
ρ,ϕ	Polar coordinates	m, rad
Φ	Radiant power or flux	W
$\phi(u,\mathrm{v})$	The wave-aberration function, equal to the OPD of the ray through point (u,v)	m
ϕ	Lens power, i.e., reciprocal focal length $\equiv 1/F$ or $1/f$	diopter, m^{-1}
$\psi(\omega_y,\omega_z)$	Phase transfer function	–
ω_y, ω_z	Radian spatial frequency in the y- and z-directions, i.e., $\omega_y = 2\pi f_y$ and $\omega_z = 2\pi f_z$	m^{-1}

Magnification, Microscopic. The ratio of the angular size of an image to the angular size of the object, if the object were to be viewed at a conventional distance. For visual work, the conventional distance is 250 mm (10 in.).

Paraxial. Pertaining to an infinitesimal, thread-like region about the optical axis.

Plane of Incidence. The incident ray, the normal to the surface at the point of incidence, and the refracted (or reflected) ray all lie in the same plane of incidence (see Fig. 2.1).

Planes, Principal. If each ray of a bundle, incident on an optical system parallel to the axis, is extended to meet the backward extension of the same ray after it has passed through the system, the locus of the intersections of all the rays is called a *principal plane.* The first principal plane is formed by rays from the right. The second principal plane is formed by rays incident from the left. The principal planes are planes only in the paraxial region; at any finite distance from the axis they are figures of rotation, frequently approximating spherical surfaces.

Points, Cardinal. The focal points, principal points, and nodal points (see Fig. 2.2).

Point, Focal. The point to which (paraxial) rays, parallel to the axis, converge, or appear to converge, after passing through the optical system (see Fig. 2.2).

Point, Front (First) Focal. The focal point to which rays incident from the right are converged (see Fig. 2.2).

Point, Back (Second) Focal. The focal point to which rays incident from the left are converged (see Fig. 2.2).

Points, Principal. The intersection of the principal planes with the optical axis (see Fig. 2.2).

Points, Nodal. Two axial points of an optical system, located such that an oblique ray directed toward the first appears to emerge from the second, parallel

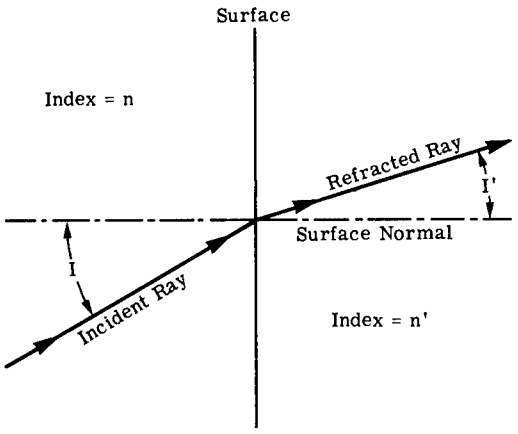


Fig. 2.1 Refraction at an optical surface. The plane of incidence and refraction is the plane of the paper.

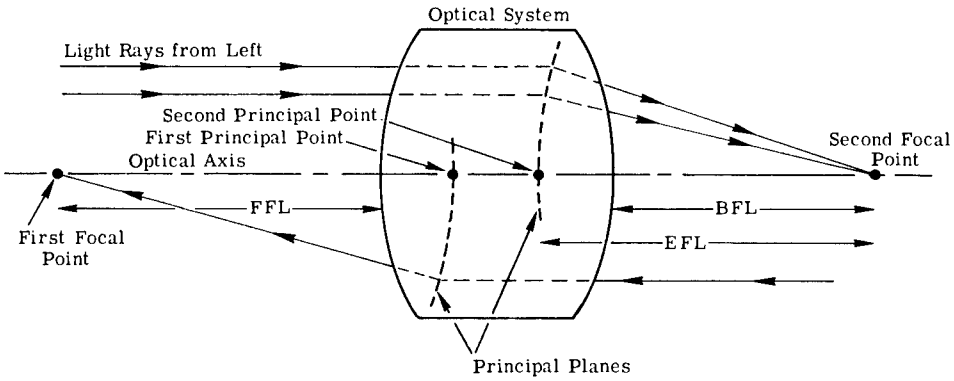


Fig. 2.2 The focal points and principal points of a generalized optical system.

to its original direction. For systems in air, the nodal points coincide with the principal points.

Pupil, Entrance. The image of the aperture stop formed by the optical elements (if any) between the aperture stop and the object. The image of the aperture stop as "seen" from the object.

Pupil, Exit. The image of the aperture stop formed by the optical elements (if any) behind the aperture stop.

Ray, Chief. A ray directed toward the center of the entrance pupil of the optical system.

Ray, Principal. Strictly, a ray directed toward the first principal point, but commonly used to refer to the chief ray.

Ray, marginal. The ray from the axial point on the object that intersects the rim of the aperture stop.

Sign Conventions. Light rays are assumed to progress from left to right. Radii and curvatures are positive if the center of curvature is to the right of the surface. Surfaces or elements have positive power if they converge light. Distances upward (or to the right) are positive, that is, points that lie above the axis (or to the right of an element, surface, or another point) are considered to be a positive distance away. Slope angles are positive if the ray is rotated counterclockwise to reach the axis. (This is the reverse of the usual geometrical convention.) Angles of incidence, refraction, and reflection are positive if the ray is rotated clockwise to reach the normal to the surface. The index of refraction is positive when the light travels in the normal left-to-right direction. When the light travels from right to left, for instance, after a reflection, the index is taken as negative (as is the distance to the "next" surface, since it is to the left).

Snell's Law. The angles of incidence, I, and refraction, I', and the refractive indices, n and n', on either side of an optical surface are related by Snell's law: $n \sin I = n' \sin I'$ (see Fig. 2.1).

Stop, Aperture. The physical diameter that limits the size of the cone of radiation an optical system will accept from an axial point on the object. For off-axis points, the limiting aperture may be defined by more than one physical feature of the optical system.

Stop, Field. The physical diameter that limits the angular field of view of an optical system.

2.3 FIRST-ORDER (GAUSSIAN) OPTICAL LAYOUT

2.3.1 Image Size and Location

The following equations apply rigorously and exactly to the paraxial characteristics of any optical system, simple or complex. Although these paraxial relationships are strictly valid for only a thin, thread-like, infinitesimal region near the optical axis, most well-corrected systems closely approximate these relationships.

Image position (see Fig. 2.3):

$$\frac{1}{s'} = \frac{1}{s} + \frac{1}{f} , \tag{2.3}$$

$$x' = -\frac{f^2}{x} . \tag{2.4}$$

Image size; lateral magnification (see Fig. 2.3):

$$m = \frac{h'}{h} = \frac{s'}{s} = \frac{f}{x} = -\frac{x'}{f} . \tag{2.5}$$

Image size; longitudinal magnification (see Fig. 2.4):

$$m_t = \frac{s_2' - s_1'}{s_2 - s_1} = \frac{s_1' s_2'}{s_1 s_2} = m_1 m_2 \approx m^2 . \tag{2.6}$$

2.3.2 Thick Elements

The power, focal length, and back focal length of a single element in air are given by (see Fig. 2.5):

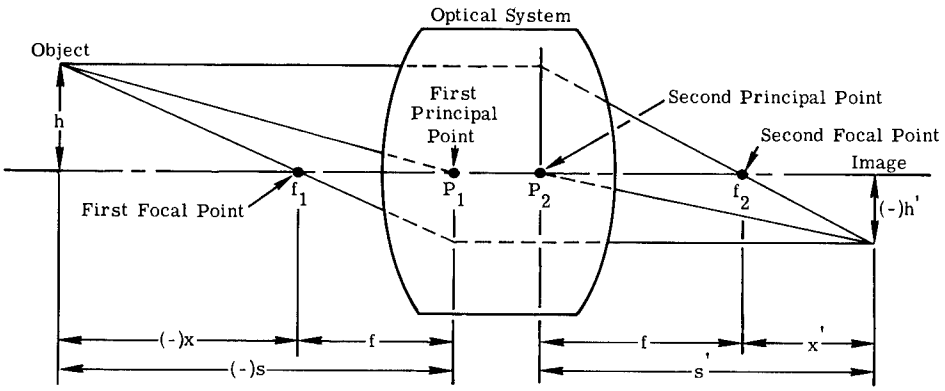


Fig. 2.3 Object and image relationships.

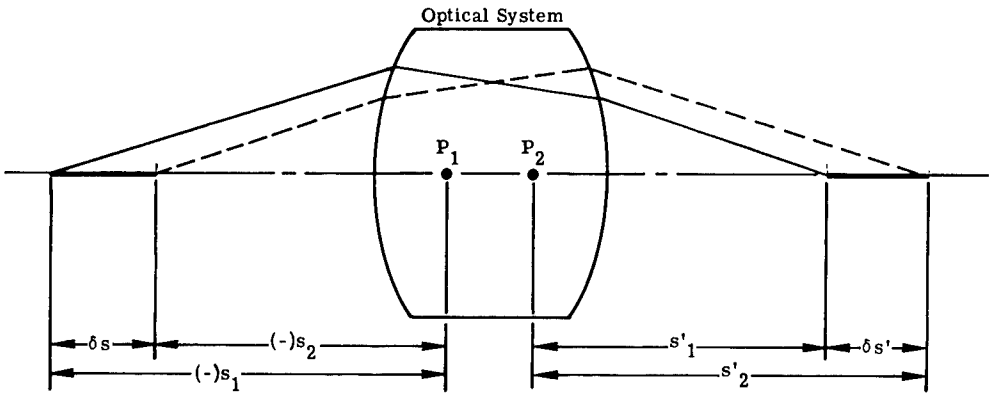


Fig. 2.4 Longitudinal magnification m_l.

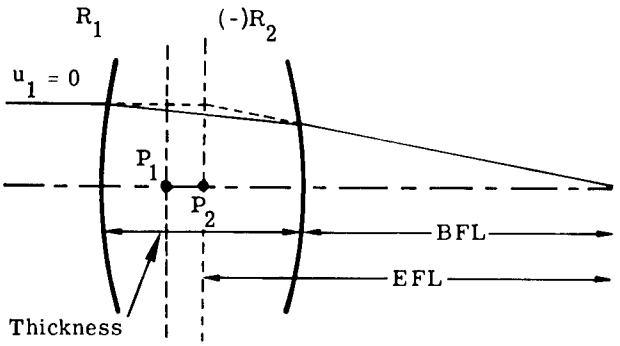


Fig. 2.5 The thick element and its second cardinal points.

$$\phi = \frac{1}{f} = (n-1)\left[C_1 - C_2 + \frac{tC_1C_2(n-1)}{n}\right]$$

$$= (n-1)\left[\frac{1}{R_1} - \frac{1}{R_2} + \frac{t(n-1)}{nR_1R_2}\right], \tag{2.7}$$

$$\text{BFL} = f\left[1 - \frac{tC_1(n-1)}{n}\right] = f\left[1 - \frac{t(n-1)}{nR_1}\right]. \tag{2.8}$$

2.3.3 Thin Lenses

When the thickness of the element is negligible, Eq. (2.7) reduces to

$$\phi = (n-1)(C_1 - C_2)$$

$$= (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right). \tag{2.9}$$

2.3.4 Two-Component Systems

When a system consists of two components, a and b, the following explicit expressions may be applied (see Fig. 2.6). Components a and b may be simple

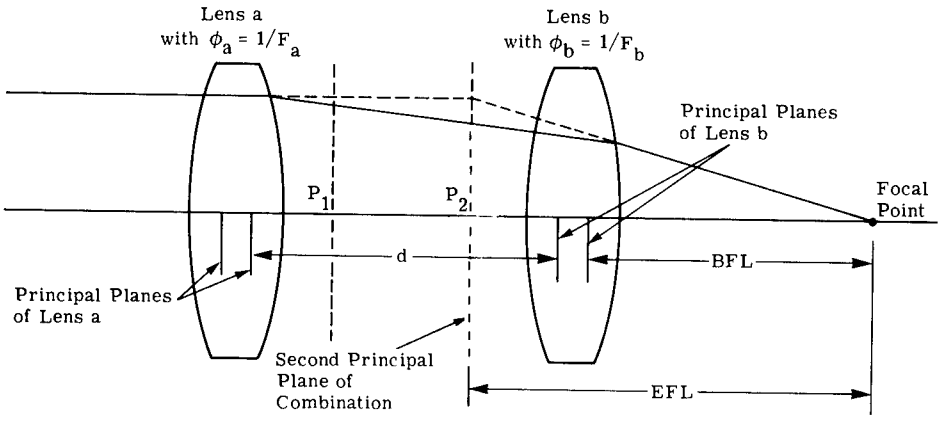


Fig. 2.6 Two-component system.

elements, mirrors, compound lenses, or complex systems in their own right. In Eqs. (2.10) through (2.16), the distances (d, BFL, FFL) are measured from the principal points of components a and b:

$$\phi_{ab} = \frac{1}{F_{ab}} = \phi_a + \phi_b - d\phi_a\phi_b = \frac{1}{F_a} + \frac{1}{F_b} - \frac{d}{F_aF_b} , \tag{2.10}$$

$$F_{ab} = \text{EFL}_{ab} = \frac{F_aF_b}{F_a + F_b - d} , \tag{2.11}$$

$$\text{BFL}_{ab} = F_b\left(\frac{F_a - d}{F_a + F_b - d}\right) = F_{ab}\left(\frac{F_a - d}{F_a}\right) , \tag{2.12}$$

$$-\text{FFL} = F_{ab}\left(\frac{F_b - d}{F_b}\right) . \tag{2.13}$$

The powers of the components that will produce a desired set of system characteristics can be determined from the following equations:

$$F_a = d\left(\frac{F_{ab}}{F_{ab} - \text{BFL}}\right) , \tag{2.14}$$

$$F_b = -d\left(\frac{\text{BFL}}{F_{ab} - \text{BFL} - d}\right) , \tag{2.15}$$

$$d = F_b\left(\frac{\text{BFL}}{F_b - \text{BFL}}\right) = F_a + F_b - \frac{F_aF_b}{F_{ab}} . \tag{2.16}$$

2.3.5 Paraxial Ray-Tracing Equations

Paraxial Image Location, Single Surface (see Fig. 2.7):

$$\frac{n'}{l'} = \frac{n}{l} + \frac{(n' - n)}{r} = \frac{n}{l} + (n' - n)C \ , \tag{2.17}$$

$$m = \frac{h'}{h} = \frac{nl'}{n'l} \ . \tag{2.18}$$

Paraxial Ray-Tracing. The following equations are more convenient for tracing ray paths than Eqs. (2.17) and (2.18) (see Fig. 2.8):

Opening equations (relating object to the first surface):

$$n_1 u_1 = \frac{n_1 y_1}{l_1} \tag{2.19}$$

or

$$n_1 u_1 = \frac{n_1 h_1}{(l_1 - s_1)} \ . \tag{2.20}$$

Iterative equations (applied to each surface in turn, $j = 1, 2, ..., k$):

$$n_j' u_j' = n_j u_j + (n_j' - n_j) y_j C_j \ , \tag{2.21}$$

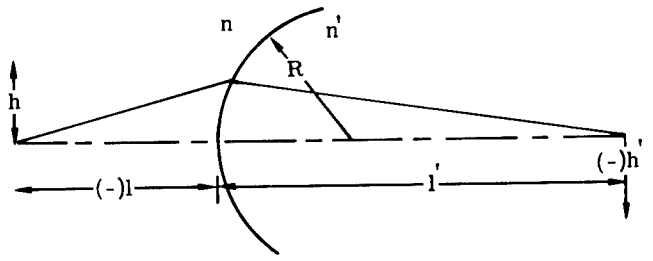


Fig. 2.7 Image formation at a single surface by Eq. (2.17).

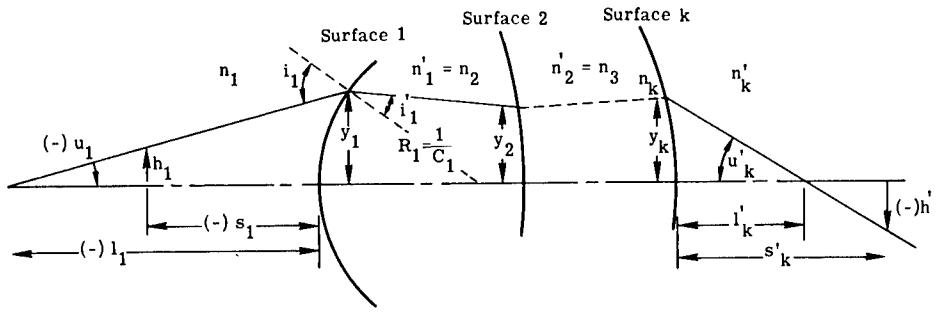


Fig. 2.8 Illustrating the nomenclature of the paraxial ray-tracing equations; Eqs. (2.19) through (2.28).

$$y_{j+1} = y_j - \frac{t_j' n_j' u_j'}{n_j'} \ . \tag{2.22}$$

Alternate iterative equations:

$$i_j = y_j C_j - u_j \ , \tag{2.23}$$

$$i_j' = \frac{n_j i_j}{n_j'} \ , \tag{2.24}$$

$$u_j' = u_j + i_j - i_j' = u_j + i_j\left(1 - \frac{n_j}{n_j'}\right) \ , \tag{2.25}$$

$$y_{j+1} = y_j - t_j' u_j' \ . \tag{2.26}$$

Closing equations (relating last surface to image):

$$l_k' = \frac{n_k' y_k}{n_k' u_k'} = \frac{y_k}{u_k'} \ , \tag{2.27}$$

$$h_k' = y_k - u_k' s_k' = u_k'(l_k' - s_k') \ . \tag{2.28}$$

2.3.6 Multielement Systems

Although paraxial rays may be traced through complete systems, one surface at a time by using Eqs. (2.19) through (2.28), it is frequently more convenient to treat a system as a set of components separated by air (Fig. 2.9). The object and image for each component (in turn) may be determined by using Eqs. (2.3) or (2.4) and (2.5). Even more convenient would be to trace rays through the system component-by-component using the following: ϕ_j, the power of the j'th component (or element); y_j, the height at which the ray strikes the principal planes of the j'th component; and d_j', the distance from the second principal plane of the j'th component to the first principal plane of the $(j + 1)$'th com-

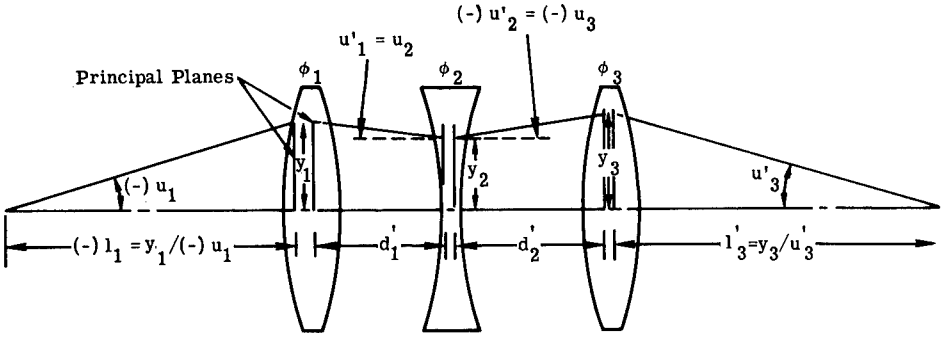


Fig. 2.9 Illustrating the ray-tracing nomenclature for use with the component-by-component ray-tracing Eqs. (2.9) and (2.10).

ponent. The ray slope after refraction by the j'th component, u_j', is determined from

$$u_j' = u_j + y_j\phi_j \ . \tag{2.29}$$

The ray height at the next component is given by

$$y_{j+1} = y_j - d_j'u_j' \ . \tag{2.30}$$

If the elements are thin, d is the space between them. If the ray from the axial intercept of the object has been traced, the magnification can be determined from the Lagrange invariant (Sec. 2.2).

2.4 EXACT RAY TRACING

2.4.1 The General, or Skew, Ray[1-4]

A general ray is defined by its direction cosines (X, Y, and Z) and by the coordinates (x, y, and z) of its intersection with a surface of the optical system (see Fig. 2.10). The subscript notation for this section is illustrated in Fig. 2.11.

Spherical Surfaces. Opening (At the initial reference surface):

$$C(x^2 + y^2 + z^2) - 2x = 0 \ , \tag{2.31}$$

$$X^2 + Y^2 + Z^2 = 1.0 \ . \tag{2.32}$$

Intersection of ray with next surface:

$$e = tX - (xX + yY + zZ) \ , \tag{2.33}$$

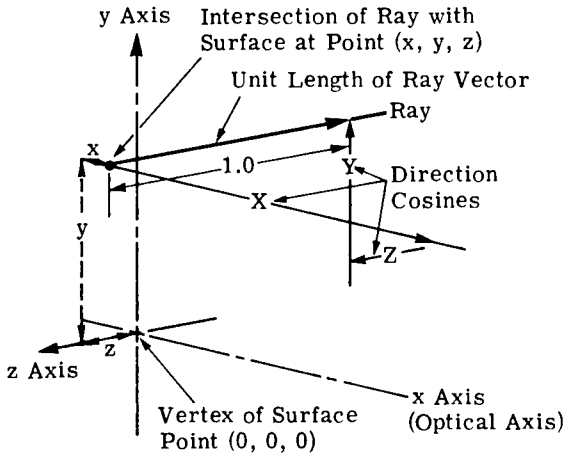


Fig. 2.10 Illustrating the symbols used in the general ray-tracing equations of Sec. 2.4. The spatial coordinates of the intersection point of the ray with the surface are x, y, and z. The ray direction cosines are X, Y, and Z (Ref. 5).

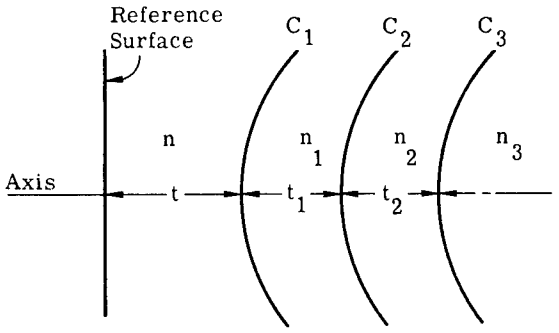


Fig. 2.11 The subscript notation used for the constructional parameters in Sec. 2.4.1 (Ref. 5).

$$M_{1x} = x + eX - t \ , \tag{2.34}$$

$$M_1^2 = x^2 + y^2 + z^2 - e^2 + t^2 - 2tx \ , \tag{2.35}$$

$$\cos I_1 = E_1 = [X^2 - C_1(C_1M_1^2 - 2M_{1x})]^{1/2} \ , \tag{2.36}$$

$$D_{0,1} = e + \left(\frac{C_1M_1^2 - 2M_{1x}}{X + E_1}\right) \ , \tag{2.37}$$

$$x_1 = x + D_{0,1}X - t \ , \tag{2.38}$$

$$y_1 = y + D_{0,1}Y \ , \tag{2.39}$$

$$z_1 = z + D_{0,1}Z \ . \tag{2.40}$$

Direction cosines of ray after refraction:

$$\cos I_1' = E_1' = \left[1 - (1 - E_1^2)\left(\frac{n}{n_1}\right)^2\right]^{1/2} \ , \tag{2.41}$$

$$g_1 = E_1' - \frac{n}{n_1}E_1 \ , \tag{2.42}$$

$$X_1 = \left(\frac{n}{n_1}\right)X - g_1C_1x_1 + g_1 \ , \tag{2.43}$$

$$Y_1 = \left(\frac{n}{n_1}\right)Y - g_1C_1y_1 \ , \tag{2.44}$$

$$Z_1 = \left(\frac{n}{n_1}\right)Z - g_1C_1z_1 \ . \tag{2.45}$$

Equations (2.33) through (2.45) are repeated with the subscripts advanced by one for the next surface. The process is continued until the final (image) surface is reached.

Aspheric Surfaces. An aspheric surface of revolution may be represented as a sphere of curvature C deformed by a series of terms in even powers of the semidiameter r:

$$x = \frac{Cr^2}{1 + (1 - C^2r^2)^{1/2}} + A_2r^2 + A_4r^4 + \cdots + A_jr^j \ , \tag{2.46}$$

where $r^2 = y^2 + z^2$ and j is an even integer.

Intersection of Ray with Aspheric. The sphere of curvature C is presumed to be a fair approximation to the aspheric. The intersection of the ray with the sphere at (x_0, y_0, z_0) is found using Eqs. (2.33) through (2.40). The actual x coordinate of the aspheric surface corresponding to this distance from the axis is found by substituting the y and z coordinates of the ray intersection with the sphere into Eq. (2.46) to get

$$r_0^2 = y_0^2 + z_0^2 \ , \tag{2.47}$$

$$\tilde{x}_0 = \frac{Cr_0^2}{1 + (1 - C^2r_0^2)^{1/2}} + A_2r_0^2 + \cdots \ . \tag{2.48}$$

Thus, a measure of the approximation error is the difference $(\tilde{x} - x)$ between the true sag of the aspheric and the approximation. Then one computes

$$l_0 = (1 - C^2r_0^2)^{1/2} \ , \tag{2.49}$$

$$m_0 = -y_0[C + l_0(2A_2 + 4A_4r_0^2 + \cdots + jA_jr_0^{(j-2)})] \ , \tag{2.50}$$

$$n_0 = -z_0[C + l_0(2A_2 + 4A_4r_0^2 + \cdots + jA_jr_0^{(j-2)})] \ . \tag{2.51}$$

An improved approximation to the intersection of the ray with the aspheric (Fig. 2.12) can be obtained from

$$G_0 = \frac{l_0(\tilde{x}_0 - x_0)}{(Xl_0 + Ym_0 + Zn_0)} \ , \tag{2.52}$$

$$x_1 = G_0X + x_0 \ , \tag{2.53}$$

$$y_1 = G_0Y + y_0 \ , \tag{2.54}$$

$$z_1 = G_0Z + z_0 \ , \tag{2.55}$$

and the new error is $(\tilde{x}_1 - x_1)$. This approximation process is repeated, from Eqs. (2.47) to (2.55), with the subscripts advanced by one at each iteration, until the error $(\tilde{x}_k - x_k)$ after the k'th iteration is negligible. Refraction at the aspheric surface is then

$$P^2 = l_k^2 + m_k^2 + n_k^2 \ , \tag{2.56}$$

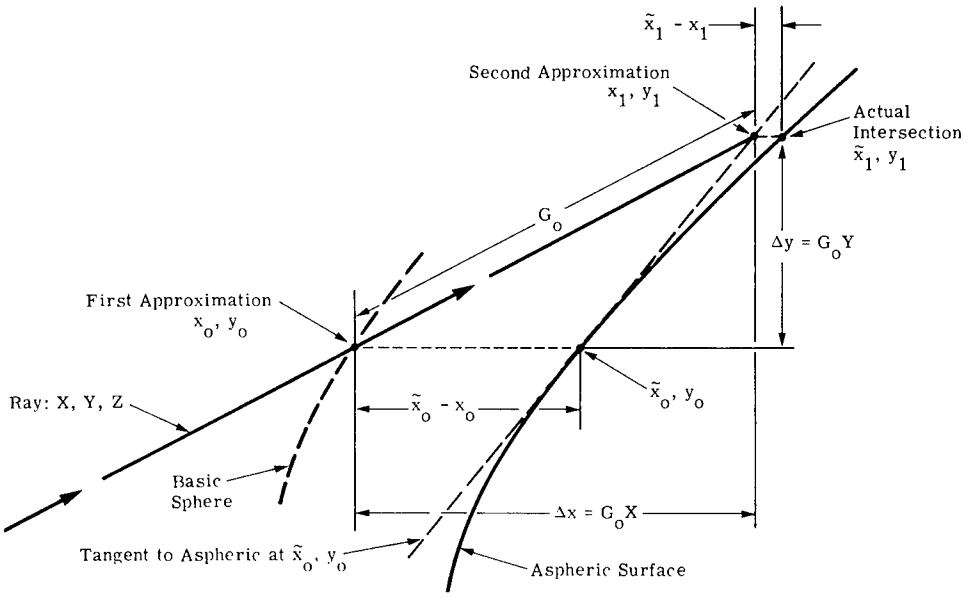


Fig. 2.12 Determination of ray intersection with an aspheric surface. The intersection of the ray with aspheric surface is found by a convergent series of approximations. Shown above are the relationships in finding the first approximation after the intersection with the basic sphere has been found.

$$P \cos I = F = Xl_k + Ym_k + Zn_k \ , \tag{2.57}$$

$$P \cos I' = F' = \left\{ P^2 \left[1 - \left(\frac{n}{n_1} \right)^2 \right] + \left(\frac{n}{n_1} \right)^2 F^2 \right\}^{1/2} \ , \tag{2.58}$$

$$g = \frac{\left[F' - \left(\frac{n}{n_1} \right) F \right]}{P^2} \ , \tag{2.59}$$

$$X_1 = X\left(\frac{n}{n1} \right) + gl_k \ , \tag{2.60}$$

$$Y_1 = Y\left(\frac{n}{n_1} \right) + gm_k \ , \tag{2.61}$$

$$Z_1 = Z\left(\frac{n}{n_1} \right) + gn_k \ . \tag{2.62}$$

2.4.2 Graphical Ray Tracing

Meridional rays can be traced using only a scale, straight edge, and compass (see Fig. 2.13). The ray is drawn to the surface, and the normal to the surface is erected at the ray-surface intersection. Two circles are drawn about the point of intersection with their radii proportional to n and n', the refractive indices

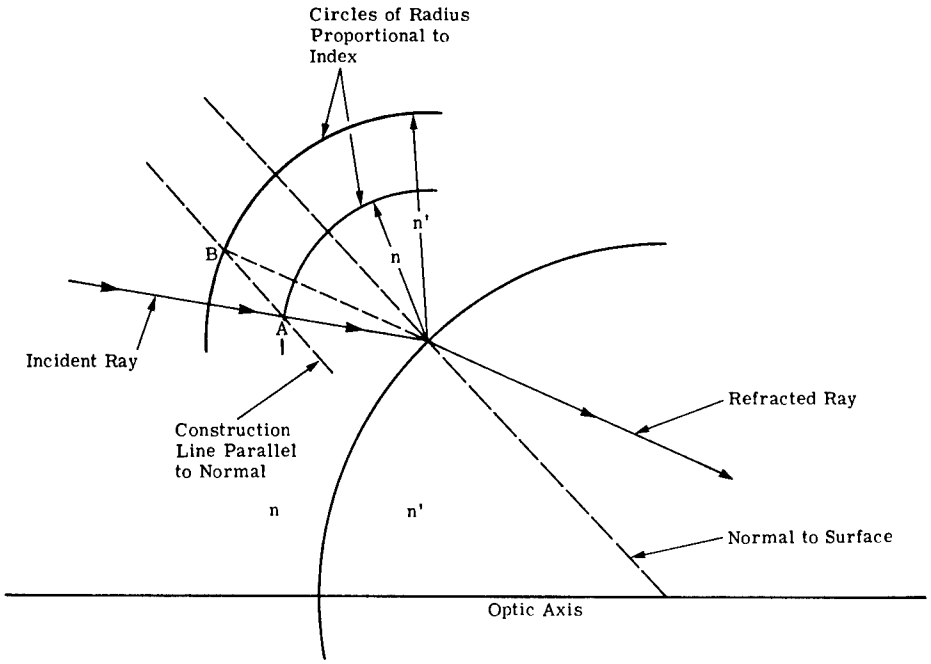


Fig. 2.13 Graphical ray tracing. Start with the construction of circles (with radii proportional to the indices on either side of the surface) about the point of intersection of the ray and surface and the development of the refracted ray.

before and after the surface, respectively. From the intersection of the ray with circle n at point A, a line is drawn parallel with the normal until it intersects circle n' at point B. The refracted ray is then drawn through point B and the ray-surface intersection. For reflection, $n' = -n$, and a single circle is drawn. Point B is located at the intersection of the parallel and the index circle on the opposite side of the surface. If desired, the index circle construction can be carried out off to one side of the drawing (to avoid cluttering the diagram) and the angles transferred to the drawing. An alternative is to measure the angle of incidence and compute the angle of refraction using Snell's law (Sec. 2.2). The accuracy of this technique is poor and the process is laborious. Thus, it is rarely used except for crude condenser-type design. It is usually preferable to use a computer and draw the rays from the computed data.

2.5 ABERRATIONS

2.5.1 Wave-Aberration Polynomial[6]

If the image aberrations of an optical system are expressed as an optical path difference (OPD) it can be shown (by reasons of symmetry) that the OPD can be expressed as a series expansion of the following form:

$$\mathrm{OPD}(h,\rho,\phi) = \sum_{l,m,n} W_{2l+n,\,2m+n,n} h^{2l+n} \rho^{2m+n} (\cos\phi)^n \ , \tag{2.63}$$

where

$W_{2l+n,2m+n,n}$ = wave-aberration numerical coefficient
h = image height
ρ,ϕ = polar coordinates of ray intersection with the system entrance pupil
l,m,n = running indices, all positive integers, 0, 1, 2, 3, ...

The term $W_{2l+n,0,0}$ is zero by definition; the term $W_{020}\rho^2$ is equivalent to a simple longitudinal shift of the image (or reference) plane; the term $W_{111}h\rho\cos\phi$ is equivalent to a vertical shift of the reference point (i.e., a change in image height). These latter two terms are called the first-order terms and become zero when the reference point is chosen at the paraxial focus.The "order" is given by $[(2l + n) + (2m + n) - 1]$.

The next five terms in the series are the third-order, or Seidel, aberrations:

- Spherical aberration $W_{040}\rho^4$
- Coma $W_{131}\rho^3\cos\phi$
- Astigmatism and Petzval $W_{220}h^2\rho^2 + W_{222}h^2\rho^2\cos^2\phi$
- Distortion $W_{311}h^3\rho\cos\phi$.

The fifth-order aberrations are

- Spherical aberration $W_{060}\rho^6$
- Linear coma $W_{151}h\rho^5\cos\phi$
- Elliptical coma $W_{331}h^3\rho^3\cos\phi + W_{333}h^3\rho^3\cos^3\phi$
- Oblique spherical $W_{240}h^2\rho^4 + W_{242}h^2\rho^4\cos\phi$
- Astigmatism and Petzval $W_{420}h^4\rho^2 + W_{422}h^4\rho^2\cos^2\phi$
- Distortion $W_{511}h^5\rho\cos\phi$.

There are 2 first-order terms, 5 third-order terms, 9 fifth-order terms, 14 seventh-order terms, 20 ninth-order terms, and $[2 + 3 \cdots + 0.5\ (N + 3)]$ N'th-order terms.

The wave aberration polynomial of Eq. (2.63) represents the departure of the actual wavefront from a perfect spherical reference surface, which passes through the axial intercept of the exit pupil and is centered on the ideal (or reference) image point. Thus, TA_y and TA_z, the y and z components of the transverse ray aberration, can be expressed in terms of the wave aberration as

$$TA_y = -\left(\frac{l}{n}\right)\frac{\partial OPD}{\partial y}\ , \tag{2.64}$$

$$TA_z = -\left(\frac{l}{n}\right)\frac{\partial OPD}{\partial z}\ , \tag{2.65}$$

where

l = distance from the exit pupil to the image point
n = index of the final medium
OPD = optical path difference [Eq. (2.63)].

2.5.2 Ray-Aberration Polynomial

The result of the operations indicated in Eqs. (2.64) and (2.65) is a pair of polynomial expressions for the ray aberrations:

$$\begin{aligned}
\mathrm{TA}_y = -\left(\frac{l}{n}\right)[&W_{020}\cdot 2\rho\cos\phi + W_{111}\cdot h + W_{040}\cdot 4\rho^3\cos\phi \\
&+ W_{131}\cdot h\rho^2(2+\cos 2\phi) + (W_{220}+W_{222})\cdot 2h^2\rho\cos\phi \\
&+ W_{311}\cdot h^3 + W_{060}\cdot 6\rho^5\cos\phi + W_{151}\cdot h\rho^4(3+2\cos 2\phi) \\
&+ W_{331}\cdot h^3\rho^2(2+\cos 2\phi) + W_{333}\cdot\frac{3}{2}h^3\rho^2(1+\cos 2\phi) \\
&+ W_{240}\cdot 4h^2\rho^3\cos\phi + W_{242}\cdot h^2\rho^3\cos\phi(3+\cos 2\phi) \\
&+ (W_{420}+W_{422})\cdot 2h^4\rho\cos\phi + W_{511}\cdot h^5 \\
&+ (\text{seventh and higher order terms})] \ , \qquad (2.66)
\end{aligned}$$

$$\begin{aligned}
\mathrm{TA}_z = -\left(\frac{l}{n}\right)[&W_{020}\cdot 2\rho\sin\phi + W_{040}\cdot 4\rho^3\sin\phi + W_{131}\cdot h\rho^2\sin 2\phi \\
&+ W_{220}\cdot 2h^2\rho\sin\phi + W_{060}\cdot 6\rho^5\sin\phi + W_{151}\cdot 2h\rho^4\sin 2\phi \\
&+ W_{331}\cdot h^3\rho^2\sin 2\phi + W_{240}\cdot 4h^2\rho^3\sin\phi \\
&+ W_{242}\cdot h^2\rho^3\sin\phi(1+\cos 2\phi) \\
&+ W_{420}\cdot 2h^4\rho\sin\phi + (\text{seventh and higher order terms})] \ . \qquad (2.67)
\end{aligned}$$

In these expressions, the sum of the exponents of the aperture and field terms ρ and h indicates the order of the aberration represented by that term. Thus, the first-order terms contain either ρ or h; the third-order aberration terms have ρ^3, $\rho^2 h$, ρh^2, or h^3; the fifth-order terms are in ρ^5, $\rho^4 h$, $\rho^3 h^2$, $\rho^2 h^3$, ρh^4, or h^5; and so forth.

2.5.3 Aberration Descriptions

Optical aberrations are faults or defects of the image. They are described in terms of the amount by which a geometrically traced ray misses a desired location in the image formed by the optical system. Ordinarily, the desired location for a ray in the image is that indicated by the first-order laws of image formation, as set forth in Sec. 2.3.

Spherical aberration can be defined as the longitudinal variation of focus with aperture (see Fig. 2.14). Longitudinal spherical aberration is the distance from the paraxial focus to the axial intersection of the ray. Lateral (or transverse) spherical aberration is the vertical distance from the axis to the intersection of the ray with the paraxial image plane.

Coma is the variation of magnification (i.e., image size) with aperture (see Fig. 2.15). Tangential coma is the vertical distance from the chief ray to the

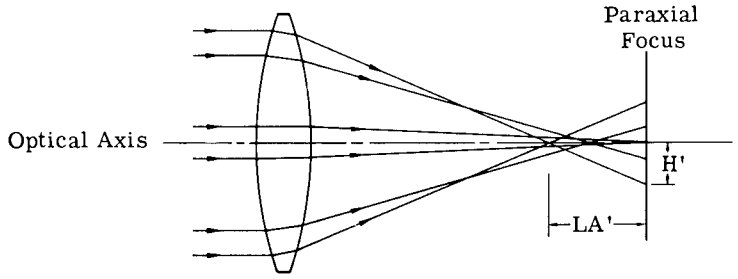


Fig. 2.14 A simple converging lens with undercorrected spherical aberration. The rays further from the axis are brought to a focus nearer the lens.

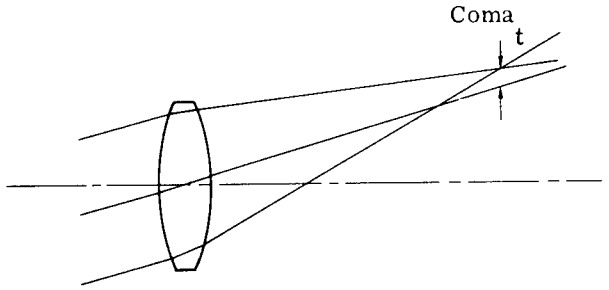


Fig. 2.15 Coma occurs in off-axis images when the rays through the outer zones of the lens form an image that is larger (as shown) or smaller than the rays through the center of the lens.

intersection of the upper and lower rim rays. (The appearance of a comatic point image is simulated in Fig. 2.28.)

Field curvature describes the amount by which the off-axis image departs longitudinally from the surface (usually flat) in which it should be located. Curvature of field may differ for rays in different meridians. The focus of a fan of rays in the meridional plane is called the tangential focus; the focus of rays lying in a plane normal to the meridional plane is called the sagittal focus. The distance from the sagittal to the tangential focus is the astigmatism, and the longitudinal distance from the paraxial image plane to the focii are the tangential and sagittal field curvatures, x_t and x_s, respectively. The *Petzval curvature* is the basic field curvature of an optical system. The Petzval surface lies three times as far from the tangential focal surface as from the sagittal focal surface, as indicated in Fig. 2.16. Field curvatures may be described as inward, undercorrected, and negative (as shown in Fig. 2.16) or as backward, overcorrected, and positive.

Distortion is the amount by which an image is closer to or further from the axis than its position as given by first-order optics. The linear amount of simple distortion varies with h^3. Thus, nonradial straight lines are imaged as curved lines (as shown in Fig. 2.17).

Axial chromatic aberration is the longitudinal variation of focal position with wavelength. The longitudinal chromatic aberration is the distance from the long-wavelength focus to the short-wavelength focus (see Fig. 2.18).

Lateral chromatic aberration is the variation of image size with wavelength. It is the vertical distance from the off-axis image of a point in long-wavelength

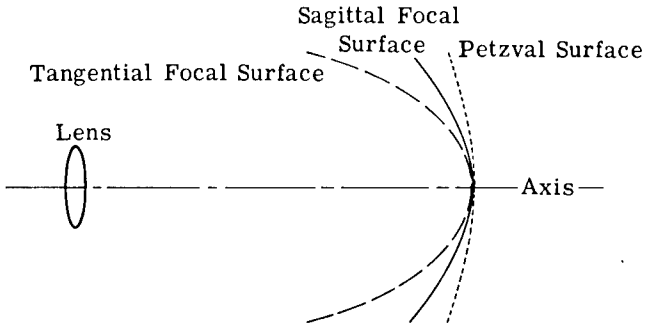


Fig. 2.16 The primary astigmatism of a single lens. The tangential image is three times as far from the Petzval surface as the sagittal image.[5]

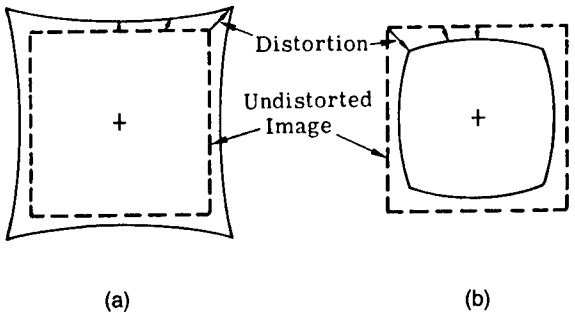


Fig. 2.17 Distortion[2]: (a) positive, or pin-cushion distortion, and (b) negative, or barrel distortion.

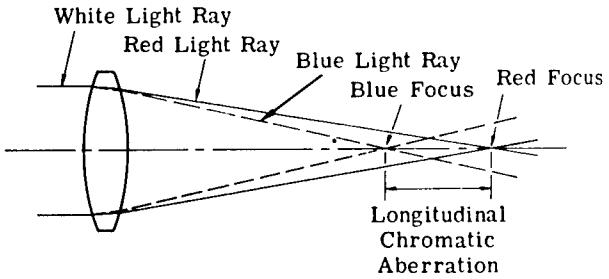


Fig. 2.18 The undercorrected, longitudinal, chromatic aberration of a simple lens. This is due to the blue rays undergoing a greater refraction than the red rays.[2]

light to the corresponding image point in short-wavelength light. It is also known as chromatic difference of magnification (CDM).

Monochromatic aberrations vary with wavelength. Chromatic variation of spherical aberration, or spherochromatism, is the most commonly encountered. Usually spherical aberration for the shorter wavelength is more overcorrected (or less undercorrected) than for the longer wavelength. Thus, a typical system might have the spherical aberration overcorrected in short-wavelength light, corrected in the center of the bandpass, and undercorrected in long-wavelength light. Chromatic variations of the other aberrations are less frequently encountered and are usually less serious.

2.5.4 Third-Order Aberrations

The third-order portion of the aberration polynomials [Eqs. (2.63), (2.66), and (2.67)], may be computed from the data of two paraxial rays traced through the optical system. The equations given here indicate directly the amount of transverse aberration at the image. Two paraxial rays are traced through the system. The axial or aperture ray is traced from the axial intercept of the object and passes through the rim of the system's pupil. The *principal, chief,* or *field ray* is traced from an appropriate off-axis point in the object and passes through the center of the pupil. The axial ray data are indicated by plain symbols. The principal ray data are indicated by the subscript p. The rays are traced using the equations of Sec. 2.3.5. The data i, y, u and i_p, y_p, u_p are thus available for each surface of the system. The optical invariant $\mathcal{J}$ is evaluated from the data of the two rays at the first (or any other convenient) surface: $\mathcal{J} = n(y_p u - y u_p)$. The final image height can be determined from the intersection height of the principal ray with the image plane, or from $h = \mathcal{J}/(n'_k u'_k)$, where n'_k and u'_k are the index and the slope, respectively, of the axial ray after it passes through the last surface of the system.

The third-order aberration and the chromatic contribution of each surface can be evaluated from the following equations:

Spherical:
$$\text{TSC} = Bi^2h + Wy^4 , \tag{2.68}$$

Sag coma:
$$\text{CC} = Bii_ph + Wy^3y_p , \tag{2.69}$$

Astigmatism:
$$\text{TAC} = Bi_p^2h + Wy^2y_p^2 , \tag{2.70}$$

Petzval:
$$\text{TPC} = \frac{(n - n')\mathcal{J}hC_e}{2nn'} , \tag{2.71}$$

Distortion:
$$\text{DC} = h\left[B_pii_p + \frac{1}{2}(u_p'^2 - u_p^2)\right] + Wyy_p^3 , \tag{2.72}$$

Axial chromatic:
$$\text{TLchC} = \frac{yi\left(\Delta n - \dfrac{\Delta n'n}{n'}\right)}{u'_k} , \tag{2.73}$$

Lateral chromatic:
$$\text{TchC} = \frac{yi_p\left(\Delta n - \dfrac{\Delta n'n}{n'}\right)}{u'_k} , \tag{2.74}$$

where
$$B = \frac{n(n' - n)y(u' - i)}{2n'\mathcal{J}} \tag{2.75}$$

$$B_p = \frac{n(n' - n)y_p(u'_p - i_p)}{2n'\mathcal{J}} \tag{2.76}$$

Δn = dispersion of the medium
$= n_{\text{short}} - n_{\text{long}}$

$$W = 4Kh(n - n')/\mathcal{J}$$

$$K = A_4 - \frac{1}{4}A_2(4A_2^2 + 6CA_2 + 3C^2)$$

$$C_e = C + 2A_2 \ .$$

The surface equation is in the form of Eq. (2.46). Equivalent curvature C_e is used in the paraxial ray tracing of any aspheric surfaces. The longitudinal values of the spherical aberration, astigmatism, Petzval curvature, and axial chromatic contributions may be obtained by dividing the transverse contributions (TSC, TAC, TPC, or TLchC) by u'_k, the final slope of the axial ray.

The third-order aberration at the image of the complete optical system is the sum of the contributions from each of the surfaces.

2.5.5 Stop-Shift Equations

When the stop (or pupil) of an optical system is shifted, the changes produced in the third-order aberrations can be computed from the following equations, in which the starred (*) terms are the aberration contributions after the stop is shifted and the unstarred terms are the contributions before the shift:

$$\text{TSC}^* = \text{TSC} \ , \tag{2.77}$$

$$\text{CC}^* = \text{CC} + \text{TSC} \cdot Q \ , \tag{2.78}$$

$$\text{TAC}^* = \text{TAC} + \text{CC} \cdot 2Q + \text{TSC} \cdot Q^2 \ , \tag{2.79}$$

$$\text{TPC}^* = \text{TPC} \ , \tag{2.80}$$

$$\text{DC}^* = \text{DC} + (\text{TPC} + 3\text{TAC}) \cdot Q + \text{CC} \cdot 3Q^2 + \text{TSC} \cdot Q^3 \ , \tag{2.81}$$

$$\text{TLchC}^* = \text{TLchC} \ , \tag{2.82}$$

$$\text{TchC}^* = \text{TchC} + \text{TLchC} \cdot Q \ . \tag{2.83}$$

The Q term represents the amount of the pupil shift:

$$Q = \frac{(y_p^* - y_p)}{y} \ , \tag{2.84}$$

where

y_p^* = ray height of the principal ray after the stop is shifted
y_p = original principal ray height, before the stop shift
y = height of the axial ray,

and Q is an invariant and thus may be evaluated at any convenient surface of the system. A further consequence of the invariance of Q is that the starred

and unstarred terms of Eqs. (2.77) through (2.83) may represent the contributions of an entire optical system or any part of it (e.g., a single surface or element).

2.5.6 Thin-Lens Aberrations

If the elements of an optical system can be regarded as thin lenses (i.e., of zero thickness) surrounded by air, the following equations give the third-order and chromatic aberration contributions of each element, assuming that the stop (pupil) is in contact with the element:

$$\text{TSC} = \frac{-y^4(G_1C^3 - G_2C^2C_1 + G_3C^2v + G_4CC_1^2 - G_5CC_1v + G_6Cv^2)}{u_k'}$$

$$= \frac{-y^4(G_1C^3 + G_2C^2C_2 - G_3C^2v' + G_4CC_2^2 - G_5CC_2v' + G_6Cv'^2)}{u_k'}, \tag{2.85}$$

$$\text{CC} = -hy^2\left(\frac{1}{4}G_5CC_1 - G_7Cv - G_8C^2\right)$$

$$= -hy^2\left(\frac{1}{4}G_5CC_2 - G_7Cv' + G_8C^2\right), \tag{2.86}$$

$$\text{TAC} = -\frac{1}{2}h^2\phi u_k', \tag{2.87}$$

$$\text{TPC} = \frac{-\frac{1}{2}h^2\phi u_k'}{n} = \frac{\text{TAC}}{n}, \tag{2.88}$$

$$\text{DC} = 0, \tag{2.89}$$

$$\text{TLchC} = \frac{-y^2\phi}{Vu_k'}, \tag{2.90}$$

$$\text{TchC} = 0. \tag{2.91}$$

Paraxial rays are traced through the system using Eqs. (2.29) and (2.30); a principal and an axial ray (as defined in Sec. 2.5.4) are traced. Equations (2.77) through (2.83) are used with $Q = y_p/y$ to determine the contributions of each element for the actual stop position. The contributions from the elements are summed to determine the aberrations of the whole system. The aspheric terms of Eqs. (2.68) to (2.81) can be added to the thin-lens aberrations obtained from Eqs. (2.77) to (2.83) when the thin lenses have aspheric surfaces.

In Eqs. (2.85) through (2.91), C represents the total curvature of the element, and C_1 and C_2 are the curvatures of the left and right surfaces, respectively, so that $C = C_1 - C_2 = \phi/(n - 1)$. The symbol v is the reciprocal object-distance (for the element), and $v = u/y$. Similarly, $v' = u'/y$. The reciprocal relative

dispersion of the lens material, V, is conventionally equal to $(n_D - 1)/(n_F - n_C)$ in the visible. In the infrared, the reciprocal relative dispersion is $(n_M - 1)/(n_S - n_L)$, where the subscripts M, S, and L identify the refractive index at middle, short, and long wavelengths, respectively.

G_1 through G_8 are functions of the index, as follows[8]:

$$G_1 = \frac{n^2(n-1)}{2} \qquad G_5 = \frac{2(n+1)(n-1)}{n}$$

$$G_2 = \frac{(2n+1)(n-1)}{2} \qquad G_6 = \frac{(3n+2)(n-1)}{2n}$$

$$G_3 = \frac{(3n+1)(n-1)}{2} \qquad G_7 = \frac{(2n+1)(n-1)}{2n}$$

$$G_4 = \frac{(n+2)(n-1)}{2n} \qquad G_8 = \frac{n(n-1)}{2} \tag{2.92}$$

2.5.7 Interpretation of Third-Order Aberration Contributions

To the extent that third-order aberrations approximate the complete aberration polynomial, aberration values computed from the equations of the preceding three sections will correspond to the aberration values as determined by actual ray tracing. Thus ΣTSC (the sum of the transverse, third-order, spherical aberration contributions) will approximate the intersection height (in the paraxial image plane) of a trigonometrically traced axial ray (i.e., the transverse spherical aberration). Similarly, 3ΣCC will approximate the tangential coma and ΣDC the linear distortion, with $100\Sigma DC/h$ the percentage distortion. The longitudinal curvature (or sag) of the field is approximated by

$$x_s \approx \frac{(\Sigma TAC + \Sigma TPC)}{u'_k}$$

and

$$x_t \approx \frac{(3\Sigma TAC + \Sigma TPC)}{u'_k} \tag{2.93}$$

for the sagittal and tangential fields, respectively. The chromatic aberration equations are

$$\frac{\Sigma TLchC}{u'_k} = LchC \approx l_F - l_C$$

and

$$\Sigma TchC \approx h_F - h_C \ , \tag{2.94}$$

where $(l_F - l_C)$ is the paraxial longitudinal axial-chromatic aberration and $(h_F - h_C)$ is the paraxial lateral chromatic aberration.

2.6 DEPTH OF FIELD AND FOCUS

2.6.1 Photographic Depth of Focus

Based on the concept that some arbitrarily selected level of blur (of diameter B) caused by lack of focus of the optical system can be tolerated, the depth of focus of the optical system is calculated to be $\pm B/2(\text{NA})$—assuming purely geometrical optics and no aberrations. The corresponding depth of field at the object ranges from S_{near} to S_{far}:

$$S_{\text{far}} = \frac{fS(D - B)}{(fD + SB)} \tag{2.95}$$

and

$$S_{\text{near}} = \frac{fS(D + B)}{(fD - SB)} , \tag{2.96}$$

where

S = nominal distance at which system is focused
D = diameter of entrance pupil
B = "tolerable" blur diameter in image.

The hyperfocal distance is the distance at which the optical system must be focused so that S_{far} is infinitely large (S_{near} is one-half the hyperfocal distance):

$$S_{\text{hyp}} = -f\frac{D}{B} . \tag{2.97}$$

2.6.2 Physical Depth of Focus

Actually, there is no sharp demarcation between being in focus and out of focus; the image worsens gradually as the amount of defocus is increased. The wavefront aberration due to defocusing is given by

$$\text{OPD} = \frac{1}{2}(\delta S)n \sin^2 U_m , \tag{2.98}$$

where

δS = longitudinal distance from position of best focus
n = index of the medium
U_m = slope angle of the marginal ray at the image.

Thus, a depth-of-focus tolerance corresponding to the Rayleigh quarter-wave criterion is given by

$$\delta S = \pm\frac{\lambda}{2n \sin^2 U_m} . \tag{2.99}$$

Note that the same equation applies to both depth of field or focus if U_m is taken as the marginal ray slope at the object or image, respectively.

2.7 VIGNETTING AND BAFFLING

2.7.1 Vignetting

Vignetting is the blocking of an oblique bundle of light rays by the limiting diameters of the optical system. Figure 2.19 shows an example of vignetting. To eliminate vignetting, one must make the lens elements toward the ends of the optical system large enough to pass all the rays of the oblique bundle. A long system is likely to require large-diameter elements to prevent vignetting; conversely, a short system can be quite compact.

2.7.2 Baffles

Baffles are opaque diaphragms that prevent the propagation of light through the system by reflection or by scattering from the mechanical (or nonoptical) elements. Figure 2.20 shows a system of baffles designed to prevent stray light from being directly reflected onto the detector from the walls of its mount.

2.7.3 Glare Stop

A glare stop is an opaque diaphragm located at the image of the optical system aperture stop. The diameter of the glare stop is exactly the same size as the

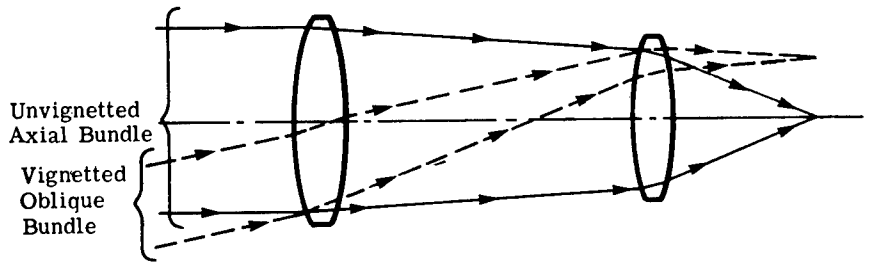


Fig. 2.19 Vignetting. The passage of the full diameter of the oblique ray bundle (dashed lines) is prevented by the lower edge of the front lens and the upper edge of the near lens.

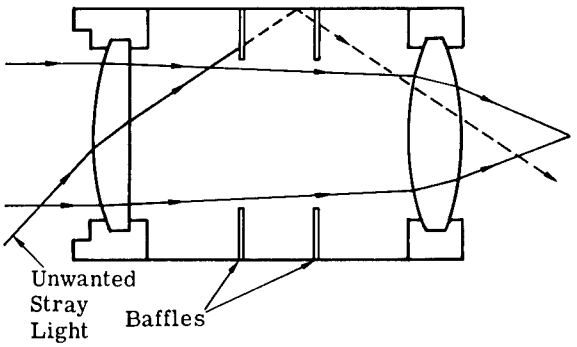


Fig. 2.20 Baffles.

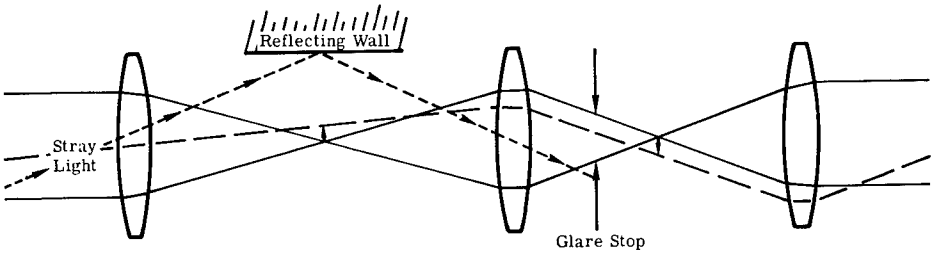


Fig. 2.21 Glare stop. An internal glare stop prevents the passage of unwanted stray light reflected from outside the field of view, but permits passage of all the useful light.

image of the aperture stop (which is a pupil of the system); thus, it will pass all the rays passing directly through the system from the aperture stop, but it will intercept those scattered from walls, etc., as shown in Fig. 2.21.

2.8 MEASURES OF OPTICAL PERFORMANCE

2.8.1 Diffraction Integral

The exact point-spread function, $h(y,z)$, of an image (as contrasted to the approximate geometrical function derived from a ray-traced spot diagram) is given by

$$h(y,z) = |A(y,z)|^2 \ , \tag{2.100}$$

where $A(y,z)$ is the complex amplitude of the wavefront emerging from the optical systems, and is given by

$$A(y,z) = \int\int_{-\infty}^{\infty} B(u,v) \exp[-i\Delta\phi(u,v)] \exp\left[i\frac{2\pi}{\lambda R}(uy + vz)\right] dudv \ , \tag{2.101}$$

where

y,z and u,v = spatial coordinates in the image plane and exit pupil, respectively, normal to the principal (or chief) ray
$B(u,v)$ = an amplitude factor proportional to the square root of the flux density (i.e., transmission) at point (u,v) in the pupil; $B(u,v) = 0$ outside the pupil
$\Delta\phi(u,v)$ = wave-aberration phase function difference, which equals the $(2\pi/\lambda)$ OPD of the ray through point (u,v)
R = reference sphere radius
i = the imaginary $\sqrt{-1}$.

The terms $B(u,v)$ and $\exp[-i\Delta\phi(u,v)]$ are often combined and referred to as the pupil function of the system. Equation (2.101) is sometimes written to include a focusing term, $\exp[i\pi x(u^2 + v^2)/\lambda R^2]$, where x is the coordinate along the chief ray.

2.8.2 Diffraction Image

If the transmission of the system is uniform over a circular aperture and the system is aberration free, the illuminance or irradiance distribution in the image is

$$E_v(y,z) = \pi\left(\frac{\mathrm{NA}}{\lambda}\right)^2 \Phi_t\left[\frac{2J_1(m)}{m}\right]^2 = E_0\left[\frac{2J_1(m)}{m}\right]^2 , \tag{2.102}$$

where

Φ_t = the total power in the point image
$J_1(m)$ = first-order Bessel function[a]
E_0 = peak illuminance
NA = $n'\sin U'$, the numerical aperture
m = normalized, radial coordinate.

Thus,

$$m = \frac{2\pi}{\lambda}\mathrm{NA}(y^2 + z^2)^{1/2} = \frac{2\pi}{\lambda}(\mathrm{NA})r . \tag{2.103}$$

The fraction of the total power falling within a radial distance, r_0, of the center of the pattern is given[7] by $[1 - J_0^2(m_0) - J_1^2(m_0)]$, where $J_0(m)$ is the zero-order Bessel function.[b]

Equation (2.102) is plotted in Fig. 2.22. The pattern consists of a circular patch of light (the Airy disk) surrounded by rings of rapidly decreasing intensity. Table 2.2 indicates the size and distribution of energy in the pattern for both a circular aperture and a slit aperture. When the aperture is rectangular and uniformly illuminated, then

$$E_v(y,z) = E_0\left\{\frac{\sin\left[\frac{2\pi(\mathrm{NA})y}{\lambda}\right]}{\frac{2\pi(\mathrm{NA})y}{\lambda}}\right\}^2\left\{\frac{\sin\left[\frac{2\pi(\mathrm{NA})z}{\lambda}\right]}{\frac{2\pi(\mathrm{NA})z}{\lambda}}\right\}^2 . \tag{2.104}$$

2.8.3 Gaussian (Laser) Beams

When the amplitude factor $B(u,v)$ in Eq. (2.101) is such that the beam cross section has a Gaussian flux-density distribution,

$$E(r) = E_0 \exp[-2(r/w)^2] , \tag{2.105}$$

then the diffraction pattern in an unaberrated image also has a Gaussian

[a] $J_1(m) = \frac{m}{2} - \frac{(m/2)^3}{1^2 2} + \frac{(m/2)^5}{1^2 2^2 3} - \cdots .$

[b] $J_0(m) = 1 - \left(\frac{m}{2}\right)^2 + \frac{(m/2)^4}{1^2 2^2} - \frac{(m/2)^6}{1^2 2^2 3^2} + \cdots .$

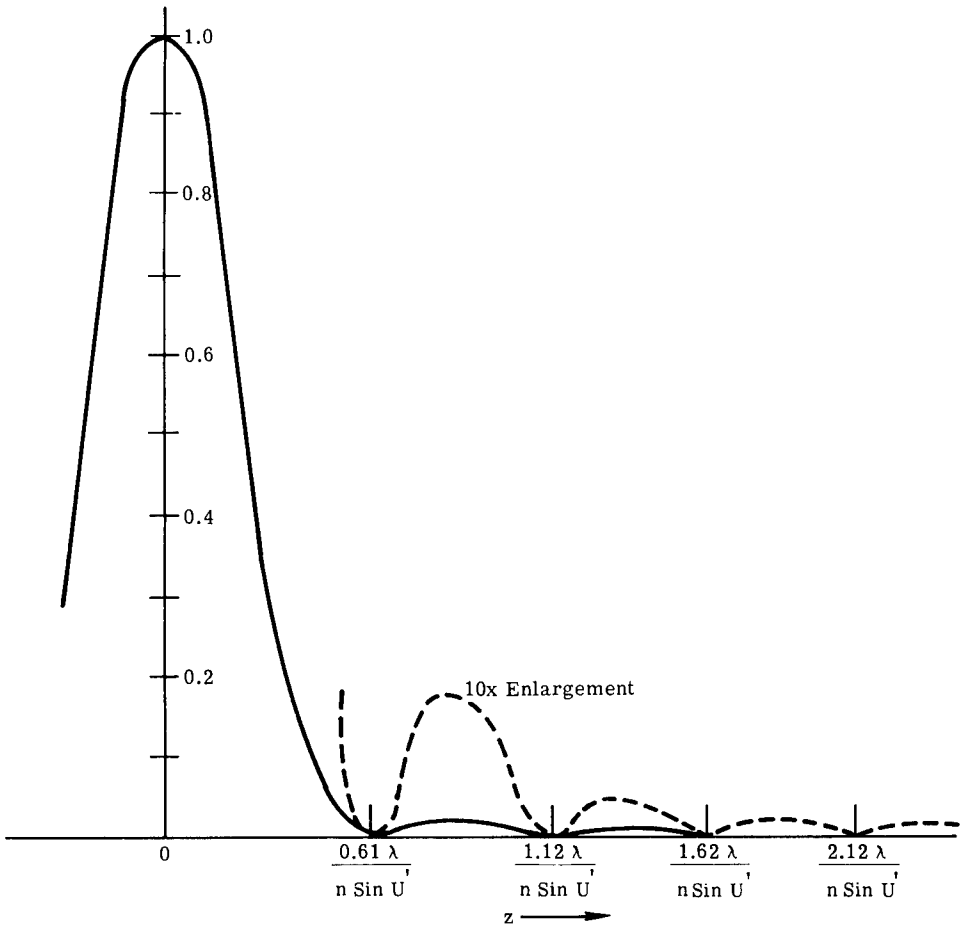


Fig. 2.22 The distribution of illumination in the diffraction pattern.[5]

Table 2.2 Tabulation of the Size of and Distribution of Energy in the Diffraction Pattern at the Focus of a Perfect Lens

Ring (or band)	*Circular Aperture*			*Slit Aperture*	
	z	*Peak Illumination*	*Energy in Ring*	z	*Peak Illumination*
Central maximum	0	1.0	83.9%	0	1.0
1st dark ring	$0.61\lambda/(n' \sin U')$	0.0	–	$0.5\lambda/(n' \sin U')$	0.0
1st bright ring	$0.82\lambda/(n' \sin U')$	0.017	7.1%	$0.72\lambda/(n' \sin U')$	0.047
2nd dark ring	$1.12\lambda/(n' \sin U')$	0.0	–	$1.0\lambda/(n' \sin U')$	0.0
2nd bright ring	$1.33\lambda/(n' \sin U')$	0.0041	2.8%	$1.23\lambda/(n' \sin U')$	0.017
3rd dark ring	$1.62\lambda/(n' \sin U')$	0.0	–	$1.5\lambda/(n' \sin U')$	0.0
3rd bright ring	$1.85\lambda/(n' \sin U')$	0.0016	1.5%	$1.74\lambda/(n' \sin U')$	0.0083
4th dark ring	$2.12\lambda/(n' \sin U')$	0.0	–	$2.0\lambda/(n' \sin U')$	0.0
4th bright ring	$2.36\lambda/(n' \sin U')$	0.00078	1.0%	$2.24\lambda/(n' \sin U')$	0.0050
5th dark ring	$2.62\lambda/(n' \sin U')$	0.0	–	$2.5\lambda/(n' \sin U')$	0.0

distribution. The size of a Gaussian beam or diffraction pattern is usually given in terms of the semidiameter w, at which the radiation falls to e^{-2}, or about 0.135 of its central value. At large distances, the angular spread of a Gaussian beam is $4\lambda/\pi D$ between e^{-2} points, where D is the e^{-2} beam diameter at the optical system. [Compare this with Eq. (2.108), which gives a half-beam spread angle.] If a Gaussian beam is truncated, or stopped down, the Gaussian distribution gradually disappears, approaching the distribution as a result of a uniformly illuminated aperture (see Sec. 2.8.2). If the clear aperture of the optical system is equal to at least twice the e^{-2} beam diameter, the flux density distribution of the beam is within a few percent of a true Gaussian shape.

2.9 RESOLUTION CRITERIA

2.9.1 Point Resolution: The Rayleigh and Sparrow Criteria.

The Rayleigh criterion is that two adjacent, equal-intensity, point sources can be considered resolved if the first dark ring of the diffraction pattern of one point image coincides with the center of the other pattern. This represents an arbitrary, but very useful, resolution limit for an optical system. Thus, in an aberration-free system with a uniformly illuminated pupil, the separation of the two points is equal to the radius of the first dark ring:

$$\text{separation} = \frac{0.61\lambda}{n' \sin U'} = \frac{0.61\lambda}{\text{NA}} \ . \tag{2.106}$$

The Sparrow criterion is that two adjacent point sources (not necessarily equal) can be considered resolved if the combined diffraction pattern has no minimum between the two point images. This occurs when

$$\text{separation} = \frac{0.5\lambda}{n' \sin U'} = \frac{0.5\lambda}{\text{NA}} \ , \tag{2.107}$$

where U' is the angle to the axis of the marginal rays of the image-forming cone. The separations can be converted to object separations either by using the object-space numerical aperture in the equations or by dividing the image separation as given above by the system magnification. In object space, the angular separation of distant object points (for axially symmetric systems) is given by

$$\text{Rayleigh angular resolution} = \frac{1.22\lambda}{D} \text{ rad} \ , \tag{2.108}$$

$$\text{Sparrow angular resolution} = \frac{\lambda}{D} \text{ rad} \ , \tag{2.109}$$

where D is the effective aperture (entrance pupil diameter) of the optical system.

Since these resolution criteria are based on an aberration- and defect-free optical system, they are frequently used as a standard for excellence of design and construction, as well as an indication of the limiting performance for a given size system.

2.9.2 The Aerial Image Modulation Curve

The aerial image modulation (AIM) curve is a plot of the minimum image modulation required to produce a response in a responsive element as a function of spatial frequency. AIM curves are commonly used for such detectors as photographic film, image tubes, and the human eye. A typical AIM curve rises with frequency, indicating that a higher modulation in the image is necessary to produce a response at higher spatial frequencies. For example, if the AIM curve and the MTF curve for a film and camera-lens combination are plotted on the same graph, the frequency at which the two curves intersect is the limiting frequency, or resolution of the combined system. More complicated measures than this are used for "system performance" as opposed to resolution.

2.10 IMAGE QUALITY CRITERIA

2.10.1 The Rayleigh Quarter-Wave Limit

The Rayleigh limit for image quality states that if the wavefront aberration, OPD, varies no more than one quarter-wavelength over the aperture of an optical system, the image will be sensibly perfect. When the wavefront is relatively smooth and free of high-order ripples, this is the reliable criterion.

The amounts of certain aberrations that correspond to a maximum wavefront deformation of one quarter-wave are as follows[8]:

Out of focus:
$$\Delta l = \frac{\pm\lambda}{2n \sin^2 U_m}, \tag{2.110}$$

Spherical aberration (third-order):
$$\mathrm{LA}_m = \frac{\pm 4\lambda}{n \sin^2 U_m}, \tag{2.111}$$

Zonal spherical aberration:
$$\mathrm{LA}_z = \frac{\pm 6\lambda}{n \sin^2 U_m}, \tag{2.112}$$

Axial chromatic aberration:
$$\mathrm{LchA} = \frac{\pm\lambda}{n \sin^2 U_m}, \tag{2.113}$$

Sagittal coma:
$$\mathrm{coma}_s = \frac{\pm\lambda}{2n \sin U_m}. \tag{2.114}$$

These values assume that the reference point is chosen to minimize the OPD.

The effect of a wavefront deformation on the diffraction pattern (Sec. 2.8.2) is to shift some radiation from the central disk into the rings. In a perfect system, 84% of the energy is in the central disk and 16% is in the rings. A

quarter-wave of defocusing produces a pattern with 68% in the central disk and 32% in the rings. This is a detectable change. In practice, however, the change is difficult to measure, and a system with less than a quarter-wave of aberration is an excellent one for most applications.

Because the Rayleigh quarter-wave limit assumes a smooth wavefront, it is less reliable when the wavefront has large, high spatial-frequency components or abrupt deformations. The rms OPD is a somewhat more widely applicable measure of the quality of a system. A rms OPD of between one-fourteenth and one-twentieth of a wave is approximately the equivalent of the classical quarter-wave Rayleigh limit.

2.10.2 Strehl Definition

The Strehl definition[7] is the ratio between the illuminance at the peak of the diffraction pattern of an aberrated point-image and the illuminance at the center of an aberration-free image. A Strehl ratio of 0.8 is equal to the Rayleigh quarter-wave limit and has a much broader applicability.

The Strehl definition of a system can be evaluated by calculating the normalized illuminance at the center of the diffraction pattern:

$$\text{Strehl definition} = \left(\frac{1}{A}\int\int e^{i\Delta\phi}dA\right)^2 , \tag{2.115}$$

where the integration is over the pupil area A, and $\Delta\phi$ is the optical phase difference of the wavefront (Sec. 2.8.1). The Strehl resolution is equal to the ratio of the integral of the three-dimensional MTF for the system divided by the integral of the MTF for an unaberrated system.

2.11 TRANSFER FUNCTIONS

The performance of any linear, shift-invariant system (see Ref. 9, particularly Chapter 6) can be described by its impulse function, or the ratio of the output spectrum to that of the corresponding input spectrum. The impulse (or transfer) function of an optical system is generally complex, is a function of spatial frequencies in two dimensions, and is not limited by causality (in contrast to an electrical system, which depends on time and cannot provide an output before there is an input).

2.11.1 Optical Transfer Function (OTF)[c]

The point spread function $h(y,z)$ is defined as the response of an optical system to a point source of light (a two-dimensional delta function). The response of the system can be written directly as the convolution of the spatial distribution and the point spread function. It can also be written in terms of the spectra of these quantities:

$$I(\omega_y,\omega_z) = H(\omega_y,\omega_z)O(\omega_y,\omega_z) , \tag{2.116}$$

[c]Sections 2.11.1 and 2.11.2 contributed by William L. Wolfe, The University of Arizona, Tucson, Arizona.

where

$I(\omega_y,\omega_z)$ = spatial frequency spectrum of the image
$O(\omega_y,\omega_z)$ = spatial frequency spectrum of the object
$H(\omega_y,\omega_z)$ = transfer function of the optical system
ω_y = radian spatial frequency in the y direction
ω_z = radian spatial frequency in the z direction.

The transfer function can be written as the Fourier transform of the point spread function:

$$H(\omega_y,\omega_z) = \int_{-\infty}^{\infty} h(y,z)\,\exp[-i(y\omega_y + z\omega_z)]\,dy\,dz$$

$$= \int_{-\infty}^{\infty} h(y,z)\,\exp[-2\pi i(yf_y + zf_z)]\,dy\,dz\ . \qquad (2.117)$$

2.11.2 Modulation and Phase Transfer Functions

In general, the optical transfer function is complex. It can be written in Cartesian or polar form as follows:

$$H(\omega_y,\omega_z) = \mathrm{Re}\{H\} + i\,\mathrm{Im}\{H\}$$

$$= |H|e^{i\psi}\ , \qquad (2.118)$$

where

ψ = arctan [Im{H}/Re{H}]
Re{H} = real part of the complex OTF
Im{H} = imaginary part of the complex OTF
$|H|$ = absolute magnitude of the OTF.

Many authors use the following abbreviations:

$$H(\omega_y,\omega_z) = \text{OTF (optical transfer function)}\ ,$$

$$|H(\omega_y,\omega_z)| = \text{MTF (modulation transfer function)}\ ,$$

$$\psi(\omega_y,\omega_z) = \text{PTF (phase transfer function)}\ . \qquad (2.119)$$

Modulation is a measure of the relation between the dimmest and brightest portions of the scene and the average level. It is one measure of what is commonly called *contrast*. Modulation of radiance is defined as follows:

$$\text{modulation} = \frac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}}\ . \qquad (2.120)$$

The modulation transfer is the ratio of the modulation in the image to that in the object:

$$\mathrm{MT} = \frac{\left(\dfrac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}}\right)\text{image}}{\left(\dfrac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}}\right)\text{object}} . \tag{2.121}$$

In the visible spectral region, the scene is usually illuminated by the sun, so that the modulation results only from reflectivity differences:

$$\text{modulation} = \frac{\rho_{\max} - \rho_{\min}}{\rho_{\max} + \rho_{\min}} . \tag{2.122}$$

The maximum value of modulation in this case is "one" and the minimum zero; it is never negative. In the infrared, the modulation can be caused by differences in emissivity ε and temperature T so that it is given by

$$\text{modulation} = \frac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}} \tag{2.123}$$

and since $L = \varepsilon L^{\mathrm{BB}}(T)$, where L^{BB} is the radiance from a blackbody,

$$\text{modulation} = \frac{\varepsilon_1 L^{\mathrm{BB}}(T_1) - \varepsilon_2 L^{\mathrm{BB}}(T_2)}{\varepsilon_1 L^{\mathrm{BB}}(T_1) + \varepsilon_2 L^{\mathrm{BB}}(T_2)} = \frac{\varepsilon_1 - \varepsilon_2 \dfrac{L^{\mathrm{BB}}(T)}{L^{\mathrm{BB}}(T_1)}}{\varepsilon_1 + \varepsilon_2 \dfrac{L^{\mathrm{BB}}(T_2)}{L^{\mathrm{BB}}(T_1)}} \tag{2.124}$$

$$= \frac{1 - \dfrac{\varepsilon_2[\exp(c_2/\lambda T_2) - 1]^{-1}}{\varepsilon_1[\exp(c_2/\lambda T_1) - 1]^{-1}}}{1 + \dfrac{\varepsilon_2[\exp(c_2/\lambda T_2) - 1]^{-1}}{\varepsilon_1[\exp(c_2/\lambda T_1) - 1]^{-1}}} . \tag{2.125}$$

The modulation can be any value between 0 to +1. The expression can be written

$$\text{modulation} = \frac{1 - a}{1 + a} \tag{2.126}$$

when

$$a = \frac{\varepsilon_2[\exp(c_2/\lambda T_2) - 1]^{-1}}{\varepsilon_1[\exp(c_2/\lambda T_1) - 1]^{-1}}$$

$$= \frac{\varepsilon_2[\exp(c_2/\lambda T_1) - 1]}{\varepsilon_1[\exp(c_2/\lambda T_2) - 1]} . \tag{2.127}$$

Figure 2.23 is a curve of modulation as a function of a.

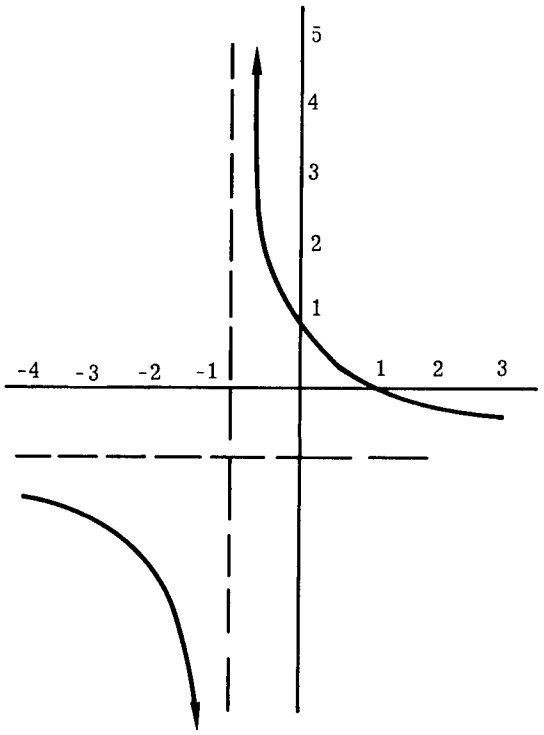


Fig. 2.23 Graph of $(1 - a)(1 + a)^{-1}$. Asymptotes are $a = -1$ and $(1 - a)(1 + a)^{-1} = -1$.

The modulation transfer of an infrared system is sometimes described in terms of the minimum variations in temperature it can sense. This description is based on the assumption that changes in flux sensed by the system are due only to changes in temperature from one part of the infrared source to the other. This minimum resolvable temperature (MRT) is proportional to the MTF.

The MTF is a more complete expression of the resolution performance than Rayleigh, Strehl, or other single measures, but it is still incomplete when presented as only a single curve. The MTF may be different in one direction of the field of view than it is in the other. This is especially true for scanning devices in which the MTF is one value in the direction of scan, but quite a different value perpendicular to the scanning direction. The MTF can be different for different field angles; this is especially true for wide-angle systems. It can also be different for different object and image positions and different focal positions. It will almost always vary with wavelength, because the diffraction limit is dependent on wavelength. To a much smaller degree, the MTF depends on the location of baffle and glare stops in the scanning optics as well as the general level of background radiation.

The geometrical MTF can be calculated from ray-trace data; both this and diffraction MTF calculations are available as subroutines on some lens-design computer programs. The usual procedure is to calculate the point spread function (sometimes with the source point at a variety of different field angles)

and then to calculate its Fourier transform. Some programs also include a diffraction contribution as part of the MTF calculation.

2.11.3 Specific Modulation Transfer Functions[10–13]

An optical system is incapable of transmitting spatial frequencies higher than f_c, which can be determined from

$$f_c = \frac{2\text{NA}}{\lambda} = \frac{1}{\lambda(F/\#)} \ . \tag{2.128}$$

For many applications, it is convenient to express the limit as an angular frequency at the object by $f_c = D/\lambda$ cycles per radian, where D is the effective clear aperture of the system.

The MTF of an optical system without aberrations and with a uniformly transmitting circular aperture is given by

$$\text{MTF}(f_y) = \frac{2}{\pi}\left(\arccos\left[\frac{\lambda f_y}{2(\text{NA})}\right] - \frac{\lambda f_y}{2(\text{NA})}\sin\left\{\arccos\left[\frac{\lambda f_y}{2(\text{NA})}\right]\right\}\right) \ . \tag{2.129}$$

Equation (2.129) is plotted in Fig. 2.24. For a slit or rectangular aperture, the MTF is

$$\text{MTF}(f_y) = 1 - \frac{f_y}{f_c} = 1 - \frac{f_y\lambda}{2\text{NA}} \ . \tag{2.130}$$

When the center of the pupil is obscured, as in a Cassegrain mirror system, the MTF of an aberration-free system is reduced at low frequencies and increased slightly at high ones. This is shown in Fig. 2.25.

Figure 2.26 shows the effect of various amounts of defocus. The amount of the focus shift is given as a function of $\sin U$, so that the graph can be applied to systems of any aperture. When the defocusing is relatively large, to the

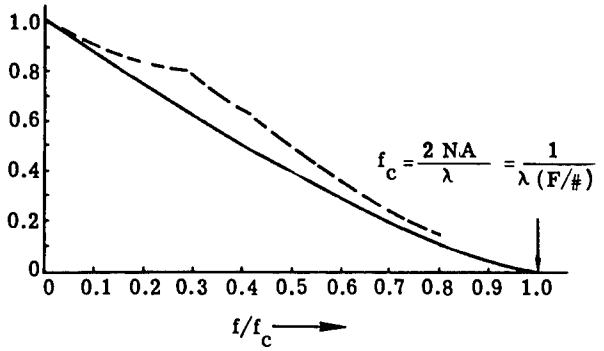
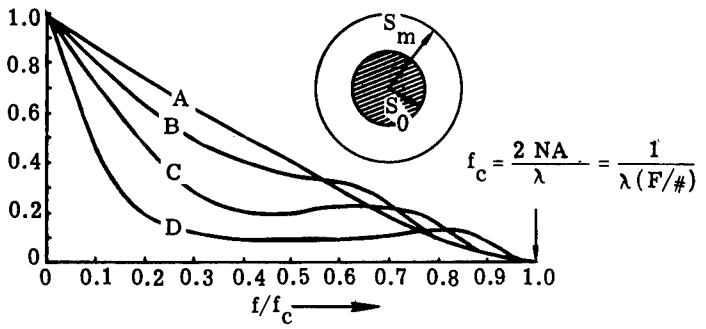


Fig. 2.24 The modulation transfer function of an aberration-free system. Note that the frequency is given in terms of the limiting resolution frequency f_c. This curve is based on diffraction effects. The dashed line is the MTF for a square-wave target; the solid line is for a sine-wave target.[2]



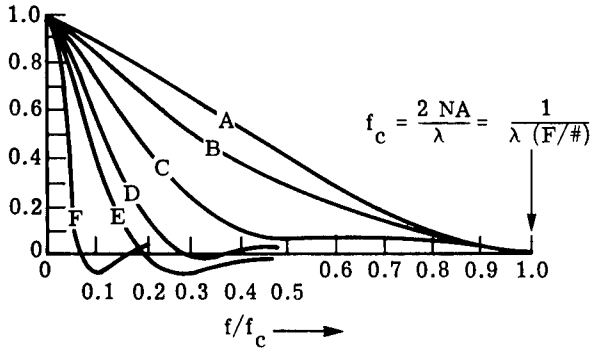
(A) $S_0/S_m = 0.0$

(B) $S_0/S_m = 0.25$

(C) $S_0/S_m = 0.5$

(D) $S_0/S_m = 0.75$

Fig. 2.25 The effect of a central obscuration on the modulation transfer function of an aberration-free system.[5]



(A) In Focus OPD =0.0

(B) Defocus = $\lambda/2n \sin^2 U$ OPD = $\lambda/4$

(C) Defocus = $\lambda/n \sin^2 U$ OPD = $\lambda/2$

(D) Defocus = $3\lambda/2n \sin^2 U$ OPD = $3\lambda/4$

(E) Defocus = $2\lambda/n \sin^2 U$ OPD = λ

Fig. 2.26 The effect of defocusing on the modulation transfer function of an aberration-free system.[5]

order of $4\lambda/n \sin^2 U$ or more, the diffraction effects can be neglected and the MTF is well approximated by

$$\text{MTF}(f_y) = \frac{2J_1(\pi B f_y)}{\pi B f_y} = \frac{J_1(2\pi\delta d\text{NA} f_y)}{\pi\delta d\text{NA} f_y}\ , \tag{2.131}$$

where

$J_1(\ )$ = first-order Bessel function (Sec. 2.8.2)

B = diameter of blur spot produced by defocusing

δd = longitudinal defocusing
NA = $n \sin U$ = numerical aperture
f_y = spatial frequency in cycles per unit length.

A system whose image is a uniformly illuminated slit, or band, of light has an MTF given by

$$\mathrm{MTF}(f_y) = \frac{\sin(\pi W f_y)}{\pi W f_y} = \mathrm{sinc}\,(W f_y)\ , \tag{2.132}$$

where W is the width of the slit or band. The equivalent cutoff frequency at which the MTF drops to zero is equal to $1/W$. This is applicable in the study of blurs caused by image motion.

2.11.4 Square Waves and Sine Waves[14]

The OTF and the MTF apply (by definition) to the imagery of an object or target whose radiance can be described by a sine function. A convenient target for the testing of optical systems is the square-wave target, which is a pattern of alternate bright and dark bars of equal width. When the MTF (i.e., the sine-wave response) is known, the modulation transfer of a square wave, $S(f_y)$, can be calculated:

$$S(f_y) = \frac{4}{\pi}\left[M(f_y) - \frac{M(3f_y)}{3} + \frac{M(5f_y)}{5} - \frac{M(7f_y)}{7} + \cdots\right]\ , \tag{2.133}$$

where $M(f_y)$ is the MTF. Conversely, when $S(f_y)$ is known,

$$M(f_y) = \frac{\pi}{4}\left[S(f_y) + \frac{S(3f_y)}{3} - \frac{S(5f_y)}{5} + \frac{S(7f_y)}{7} - \cdots\right]\ . \tag{2.134}$$

In general, the modulation transfer factor is higher for a square-wave target than for a sine-wave target. For example, the factor for a perfect optical system at frequencies between $0.25\,f_c$ and $0.5\,f_c$ is about 0.1 greater for a square-wave target than for a sinusoidal target (see Fig. 2.24).

The most common form of bar (i.e., square-wave) target is the USAF 1951 target, which consists of only three bright bars on an extended dark background (or the reverse) for each frequency. If the frequency of the target is taken as the reciprocal of the center line spacing of the bars, the modulation transfer factor is higher than that indicated by the sine-wave MTF for this frequency. This is because the frequency content of a three-bar pattern is heavily concentrated in frequencies lower than the basic frequency of the target, i.e., a spectral breakdown (Fourier analysis) of a three-bar target shows lots of power at frequencies less than 1/(bar spacing).

2.11.5 Pupil Convolution

The computation of the OTF can be carried out by an autoconvolution of the system pupil function (Sec. 2.8.1). For aberration-free systems, the MTF can

be readily computed regardless of the shape of the aperture since the MTF is simply the normalized area common to the pupil and the pupil is displaced laterally by an amount proportional to the frequency. The displacement corresponding to f_c, the cutoff frequency, is, of course, equal to the maximum dimension of the pupil in the direction of the spatial frequency (i.e., the displacement beyond which there is no common area). Note that Eq. (2.129) in Sec. 2.11.3 can be derived using this principle; it is simply the area common to a circle (the aperture) and that same circle displaced, normalized by dividing by the area of the circle, and with f_c corresponding to a displacement equal to the circle diameter.

2.12 RAY-INTERCEPT PLOTS AND SPOT DIAGRAMS

2.12.1 Ray-Intercept Plot

To prepare a ray-intercept plot (H-tan U curve), one traces a fan of rays (either meridional or sagittal) from an object point through the optical system, and the coordinates of the ray intersection with the image surface H are plotted against the position of the ray in the aperture, which is often represented by the slope of the ray, tan$U$, at the image. The spread of radiation in the image can thus be read directly from the plot, and an estimate of image blur can easily be made. When the plot coordinates are H and tan$U$, the effects of refocusing on the size of the image blur are readily evaluated by rotating the tan$U$ axis of the plot.

Figure 2.27 illustrates ray-intercept plots for several common aberrations. Ray-intercept plots are often (incorrectly) called rim-ray curves.

2.12.2 Spot Diagrams and Spread Functions

If the aperture of an optical system is divided into a large number of equal, small areas, and if a ray from a selected point is traced through the center of each small area, then a plot of the intersection points (spots) of the rays with the image surface is an approximate representation of the (geometrical) irradiance distribution at the image (see Fig. 2.28). The more rays traced, the better the approximation. Such a representation is called a *spot diagram.* Assuming the geometrical spot diagram to be a reasonable representation of the actual image irradiance distribution, several other representations can be derived from it. The radial energy distribution is obtained by arbitrarily selecting a center point and plotting the percentage of the energy (i.e., the number of spots) encircled within a radius R as a function of R. If the irradiance (i.e., the spot density) is represented as a function of the y and z coordinates of the image plane, this is the point spread function of the system. It can be compared to the measured value or transformed to obtain a geometrical approximation to MTF.

2.13 RELATIONSHIP BETWEEN SURFACE IMPERFECTIONS AND IMAGE QUALITY

The effect of a manufacturing defect on the image quality of an optical system can be estimated by converting it into a wavefront deformation or OPD. For

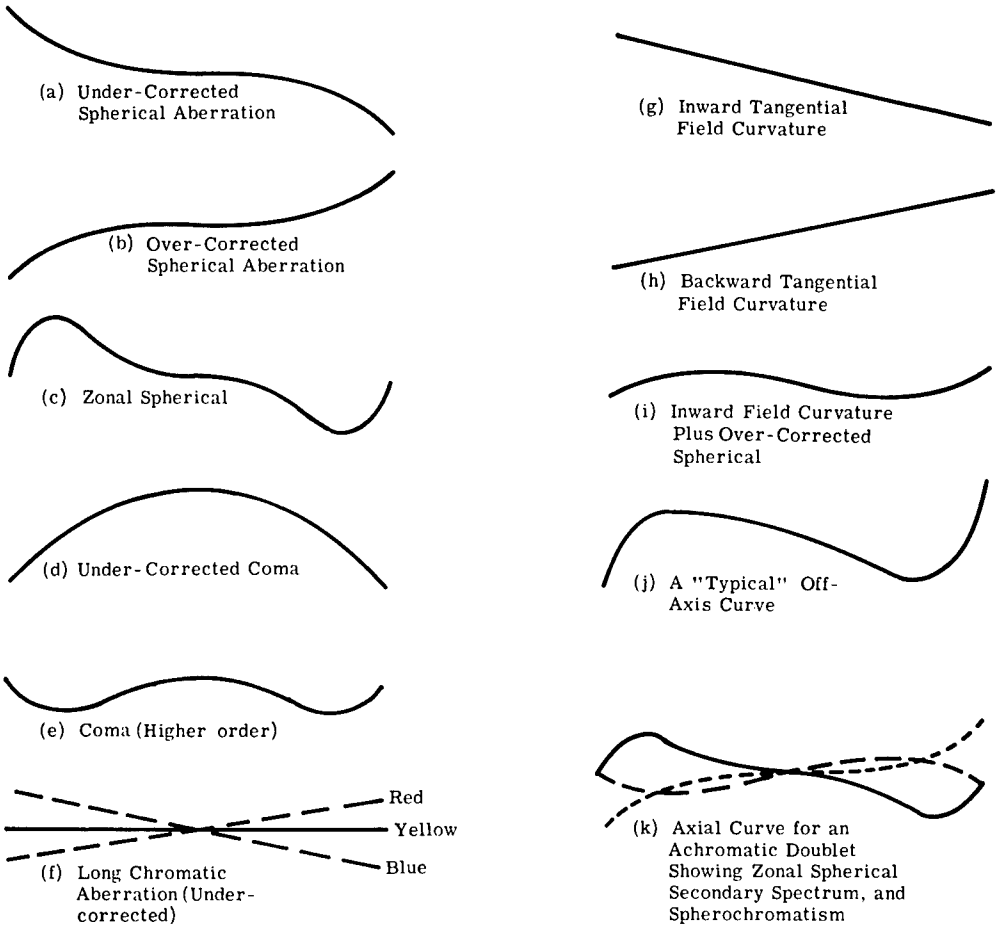


Fig. 2.27 Ray-intercept curves for various aberrations. The ordinate for each curve is H, the height at which the ray intersects the (paraxial) image plane; the abscissa is tan U', the final slope of the ray with respect to the optical axis.[5]

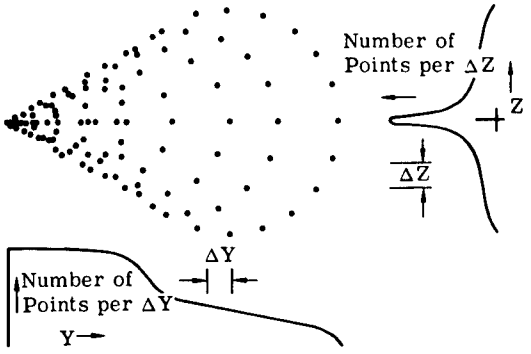


Fig. 2.28 Spot diagram (for a system with pure coma) and the two line spread functions (below and on the right) obtained by counting the number of points between parallel lines separated by a small distance, ΔY or ΔZ (Ref. 5).

example, the irregularity of figure (i.e., the departure of a surface from its ideal geometrical configuration) is usually measured in interference fringes, each of which represents a departure from the nominal surface of one half-wavelength of visible light. This can be converted to a wavefront deformation:

$$\text{OPD} = \frac{1}{2}Fr(n' - n) \text{ wavelengths} ,$$

where Fr is the number of fringes of irregularity (or asphericity in the case of a spherical surface) and $(n' - n)$ is the change in index across the surface.

When the OPD is summed for the entire system, its effect on the image can be estimated using Fig. 2.26 as a guide, since OPD indicates the reduction in MTF caused by a low-order distortion of the wavefront. The defect can also be evaluated as an rms defect and related to the peak-to-peak measure, as indicated in Sec. 2.10.1.

References

1. D. P. Feder, "Optical calculations with automatic computing machinery," *Journal of the Optical Society of America,* **41**(4), 630–635 (April 1951).
2. L. Montagnino, "Ray tracing in inhomogeneous media," *Journal of the Optical Society of America* **58**(12), 1667–1668 (Dec. 1968).
3. G. H. Spencer and M. V. R. K. Murty, "Generalized ray-tracing procedure," *Journal of the Optical Society of America* **52**(6), 672–678 (June 1962).
4. *Handbook of Optical Design,* MIL-HDBK-141, U.S. Government Printing Office, Washington, DC (1962).
5. Warren Smith, *Modern Optical Engineering: The Design of Optical Systems,* 2nd ed., pp. 66–69, 81, 82, 152, 288, 340, 357, 360, McGraw-Hill, New York (1990).
6. H. H. Hopkins, *Wave Theory of Aberrations,* Oxford University, London (1950); University Microfilms, Ann Arbor, MI, No. OP17185.
7. M. Born and E. Wolf, *Principles of Optics,* Macmillan, New York (1964).
8. A. E. Conrady, *Applied Optics and Optical Design,* Dover, New York, two volumes (1957 and 1960).
9. J. W. Goodman, *Introduction to Fourier Optics,* McGraw-Hill, New York (1968).
10. R. Barakat, "Numerical results concerning the transfer functions and total illuminance for optimum balanced fifth-order spherical aberration," *Journal of the Optical Society of America* **54**(1), 38–44 (Jan. 1964).
11. R. Barakat and A. Houston, "Diffraction effects of coma," *Journal of the Optical Society of America* **54**(9), 1084–1088 (Sep. 1964).
12. R. Barakat and A. Houston, "The effect of a sinusoidal wavefront on the transfer function of a circular aperture," *Applied Optics* **5**(11), 1850–1852 (Nov. 1966).
13. E. L. O'Neill, "Transfer function of an annual aperture," *Journal of the Optical Society of America* **46**(4), 285–288 (April 1956).
14. J. W. Coltman, "The specification of imaging properties by response to a sine wave input," *Journal of the Optical Society of America* **44**(6), 468–471 (June 1954).

Bibliography

Barakat, R., "Numerical results concerning the transfer functions and total illuminance for optimum balanced fifth-order spherical aberration," *Journal of the Optical Society of America* **54**(1), 38–44 (Jan. 1964).

Barakat, R., and A. Houston, "Diffraction effects of coma," *Journal of the Optical Society of America* **54**(9), 1084–1088 (Sep. 1964).

Barakat, R., and A. Houston, "The effect of a sinusoidal wavefront on the transfer function of a circular aperture," *Applied Optics* **5**(11), 1850–1852 (Nov. 1966).

Born, M., and E. Wolf, *Principles of Optics,* Macmillan, New York (1964).

Buchdahl, H. A., *Optical Aberration Coefficients,* Dover, New York (1968).

Coltman, J. W., "The specification of imaging properties by response to a sine wave input," *Journal of the Optical Society of America* **44**(6), 468–471 (June 1954).

Conrady, A. E., *Applied Optics and Optical Design,* Dover, New York, two volumes (1957 and 1960).

Cox, Arthur, *A System of Optical Design,* Focal, London (1965).

Driscoll, Walter, Ed., *Handbook of Optics,* McGraw-Hill, New York (1978).

Feder, D. P., "Optical calculations with automatic computing machinery," *Journal of the Optical Society of America* **41**(4), 630–635 (April 1951).

Fischer, Robert E., Ed., *Proceedings of the 1980 International Lens Design Conference, Proc. SPIE* **237** (1980).

Goodman, J. W., *Introduction to Fourier Optics,* McGraw-Hill, New York (1968).

Handbook of Optical Design, MIL-HDBK-141, U.S. Government Printing Office, Washington, DC (1962).

Hardy, A. C., and F. H. Perrin, *The Principles of Optics,* McGraw-Hill, New York (1932).

Hertzberger, M., *Modern Geometrical Optics,* Interscience, New York (1958).

Hopkins, H. H., *Wave Theory of Aberrations,* Oxford University, London (1950); University Microfilms, Ann Arbor, MI, No. OP17185.

Kingslake, Rudolf, Ed., *Applied Optics and Optical Engineering,* Academic Press, New York, five volumes (1965–1969); Kingslake, Rudolf, and B. J. Thompson, Eds., one volume (1980); Shannon, R. R., and J. C. Wyant, Eds., five volumes (1979–1992).

Kingslake, Rudolf, *A History of the Photographic Lens,* Academic Press, San Diego (1989).

Kingslake, Rudolf, *Lens Design Fundamentals,* Academic Press, New York (1978).

Kingslake, Rudolf, *Optical System Design,* Academic Press, New York (1983).

Kingslake, Rudolf, *Optics in Photography,* SPIE, Bellingham, WA (1992).

Korsch, Deitrich, *Reflective Optics,* Academic Press, San Diego (1991).

Laikin, Milton, *Lens Design,* Marcel Dekker, New York (1991).

Lawrence, George, Ed., *Proceedings of the 1990 International Lens Design Conference, Proc. SPIE* **1354** (1990).

Linfoot, E. H., *Fourier Methods in Optical Design,* Focal Press, New York (1964).

Martin, L. C., *Technical Optics,* Pitman, London, two volumes (1960).

Montagnino, Lucian, "Ray tracing in inhomogeneous media," *Journal of the Optical Society of America* **58**(12), 1667–1668 (Dec. 1968).

O'Neill, E. L., *Introduction to Statistical Optics,* Addison-Wesley, Reading, MA (1963).

O'Neill, E. L., "Transfer function of an annular aperture," *Journal of the Optical Society of America* **46**(4), 285–288 (April 1956).

Perrin, F. H., "Methods of appraising photographic systems," *Journal of the Society of Motion Picture and Television Engineers* **69,** 151–156 (March 1960); see also **69,** 239–249 (April 1960).

Rutten, Harrie G. J., and Martin A. M. Van Venrooij, *Telescope Optics,* Willmann-Bell, Richmond, VA (1988).

Schroeder, D., *Astronomical Optics,* Academic Press, San Diego (1987).

Smith, Warren J., Ed., *Lens Design,* Volume CR-41, SPIE, Bellingham, WA (1992).

Smith, Warren J., *Modern Lens Design,* McGraw-Hill, New York (1992).

Smith, W., *Modern Optical Engineering: The Design of Optical Systems,* 2nd ed., 319, 320, 322, McGraw-Hill, New York (1990).

Southall, J. P. C., *Mirrors, Prisms and Lenses,* Dover, New York (1964).

Spencer, G. H., and M. R. V. K. Murty, "Generalized ray-tracing procedure," *Journal of the Optical Society of America* **52**(6), 672–678 (June 1962).

Stephens, R. E., and L. Sutton, "Diffraction images of a point in the focal plane and several out-of-focus planes," *Journal of the Optical Society of America* **58**(7), 1001–1002 (July 1968).

W. H. Taylor and D. T. Moore, Eds., *Proceedings of International Lens Design Conference, Proc. SPIE* **554** (1985).

CHAPTER 3

Optomechanical Scanning Applications, Techniques, and Devices

Jean Montagu
Herman DeWeerd
General Scanning Inc.
Watertown, Massachusetts

CONTENTS

3.1 INTRODUCTION

The field of scanning is extremely broad. Any scanning system is a complex arrangement of optical, mechanical, electrical, and electronic subsystems. This chapter is limited to scanning techniques that apply to reasonably fast and accurate systems as they are encountered in infrared applications. This chapter focuses on the subsystem that specifically provides the scanning capability. It does not cover the pre- or post-scanning optics nor the associated detectors, electronics, controls, logics, or software.

The second section discusses scanning applications, and the third section derives the appropriate parameters to select the scanner best suited for the application. The fourth section describes the actual types of scanners available to the designer. The main features and merits of each type are reviewed with respect to appropriate applications. The fifth section gives examples of multi-scanner configurations and shows their applicability to the task on hand. The final sections offer references on the performances of scanners and materials to guide further pursuits in the field. A list of symbols and nomenclature is given in Table 3.1.

3.2 SCANNING APPLICATIONS IN THE INFRARED

At the most basic level, all the scanners and scanning systems described in this chapter are used in one or more of the following applications: warning systems, tracking, pointing/designating, communications, imaging, radiometry, and scanning microscopes.

3.2.1 Warning Systems

Warning systems, for example, can detect a missile launch, hostile fire, hostile aircraft, and terrain and weather conditions.

To detect high-flying aircraft from space, a scanner sweeps in a raster pattern over an area of several hundred square miles making periodic observations. A computer compares these observations to background radiation levels. The system detects a target even on days when scattered solar radiation exceeds the target level.

Another important type of infrared warning system detects enemy laser beams from weapon guidance systems in time to permit evasion or preemptive counterfire.

3.2.2 Tracking

When tracking, a system follows a moving infrared emitter or reflector, as in a missile launch. In one system the scanner locks on to the rocket flame behind the missile, then moves to the heat center of the exhaust plume. This leads the tracking to point approximately 3 mrad behind the missile depending on the optical magnification.

Active warning and tracking systems irradiate a target with energy from an infrared laser and then gather part of the scattered energy. In an active heterodyne system, the collected energy is compared either to a second signal from the same laser or to another laser that is phase-locked to the first.

Table 3.1 Symbols, Nomenclature, and Units

Symbols	Nomenclature	Units
A	Area	m^2
a	Acceleration	ms^{-2}
B	Magnetic induction	T
b	Bandwidth	Hz
d	Thickness	m
D	Diameter	m
D^*	Specific detectivity	$cm\ Hz^{1/2}\ W^{-1}$
e	Voltage	V
h	Distance	m
i	Current	A
J	Inertia	$g\ cm^2$
L	Axial length of magnetic field	m
$\mathscr{L}$	Inductance	H
N	Number of turns	—
n	Number of resolution elements	—
Q	Quality factor	—
r	Radius	m
R	Resistance	Ω
SD	Standard deviation	—
T	Torque	N m
t	Time	s
T_f	Time frame	s
Greek:		
α	Angles	rad
β	Angles	rad
ΔF	Change in focal length	m
η	Poisson's ratio	—
θ	Angles of rotation	rad
λ	Wavelength of light	m
ξ	Angles	rad
ρ	Density	$g\ cm^{-3}$
τ	Time constant $\mathscr{L}/R$	s

3.2.3 Pointing/Designating

In weapon guidance, pointing and designating indicate the existence and location of a source or reflector of infrared radiation.

Optical trackers are used extensively in the guidance of weapons such as air-to-air, air-to-surface, and surface-to-air missiles. Some aircraft fire control systems use passive infrared trackers for initial target acquisition. They also serve as a pointing reference to which infrared missiles are slaved.

Some battlefield missiles with command guidance systems contain an infrared beacon. The launcher tracks the beacon in order to generate steering commands to transmit to the missile. Frequently a second scanner also tracks the target. A computer compares the relative positions of the missile and target and produces steering commands.

Beam rider weapon guidance systems use a laser beam target designator to illuminate the target. A sensor on the weapon tracks the target by following the reflected beam.

In a homing guidance system the missile contains a sensor, or seeker, that corrects the course to follow the target.

3.2.4 Satellite Communication

An active scanning system in a satellite communication system maintains alignment between a communication transmitter and receiver satellites. One satellite sends a beam of infrared light to a second satellite, which reflects the beam back to the first. The system maintains the alignment of the communication signal by measuring the divergence of the return beam from the original line of sight and adjusting for the difference.

3.2.5 Imaging/Mapping

An imaging and mapping system acquires information about an area on the basis of its infrared characteristics. Thermal imagers provide information about physical objects such as buildings, people, and vehicles. They also indicate atmospheric conditions such as weather patterns, air pollution, and the presence of poison gas.

Active systems use the infrared sensor in conjunction with a beacon, or laser beam, that either marks the target or provides a basis for radiometric analysis. Some infrared scanners, however, are purely passive systems. They simply respond to the radiation that already exists.

Military reconnaissance and surveillance systems use passive infrared imaging scanners widely. They differ from designating and tracking scanners in a basic way. Rather than finding and following a central point of energy, they provide a picture of a field in terms of its spatial distribution of infrared energy.

Infrared, or thermal, imagers grew out of earlier thermal mappers used in aircraft to map the terrain below. The forward motion of the aircraft swept a single detector or a pushbroom—a linear array of optical detectors—across the target area. Initially, detector response was relatively poor, and map information was not available in real time.

Improved detectors and the application of television technology for rapid update and real-time imaging led to the development of high-speed high-accuracy infrared surveillance scanners. The use of moving mirrors and prisms to sweep out a two-dimensional raster pattern made possible the important class of systems known as forward-looking infrared (FLIR) sensors.

A scanning system in a FLIR sensor scans a HgCdTe detector to collect the infrared radiation in a rectangular field of view. It then creates a visual representation of the thermal scene of the field using an array of light-emitting diodes or conventional television equipment.

FLIR sensors are often used in conjunction with target designators and weapon guidance systems. The operator views the FLIR display, selects a target, illuminates it with the designator, and launches the laser-guided weapon.

FLIR sensors have been developed for use in many different kinds of platforms. There are hand-held, tripod-mounted, and ground-vehicle-based FLIRS, as well as FLIRS carried by aircraft and spacecraft. FLIRS have nonmilitary applications, such as heat loss surveys, forest fire control, and security systems.

3.2.6 Radiometers

Radiometers are a broad class of devices for measuring radiation. Radiometers incorporating scanners include systems for identifying weather patterns, atmospheric turbulence, air pollution, and poison gas. Both active and passive devices exist. A poison gas detection scanner, for example, uses a spectroradiometer. It detects changes in the spectral characteristics of a laser beam that has been transmitted through a gaseous atmosphere.

3.2.7 Scanning Microscopes

Infrared microimaging, an important application in the semiconductor industry, exemplifies the capabilities of this technology. Figure 3.1 is a schematic description of such a unit with a conventional microscope used for alignment. Other similar uses of infrared imaging scanning systems include semiconductor wafer and printed circuit board fault detection systems.

3.3 DERIVATION OF SCANNER PERFORMANCE

The application of the scanning system defines the parameters of performance of the scanning devices. The scanner frequently contributes a large fraction of the error budget. A typical design target is 10% but often it rises to 50%.

3.3.1 Target Designator for Aircraft

A typical target, a tank, is 3 × 3 m (see Fig. 3.2). The maximum missile release range is 3000 m. Consequently, the maximum tolerable absolute total system error for a direct hit is 1 mrad. The error budget share ascribed to the scanner is in the vicinity of 100 μrad within the environmental constraints.

Speed of response is determined by the display frame rate, which is itself conditioned by the speed of the aircraft. A 2-ms flyback time or less is currently the expected performance.

- *Aperture:* Typically 25 mm.
- *Frame rate:* TV compatible; 25, 30, 50, 55, 60 Hz.
- *Line rate:* TV compatible; full repetition rate, 15,750 Hz or 0.5, 0.25, or less as defined by the number of detectors and the needed refresh rate.
- *Jitter and wobble:* A factor of 3 smaller than the resolution of the system.
- *Environmental conditions:* Operating temperature may be between −68° and 78°C. Systems must meet applicable military standards and specifications for vibration, acceleration, shock, electromagnetic interference, and so forth.

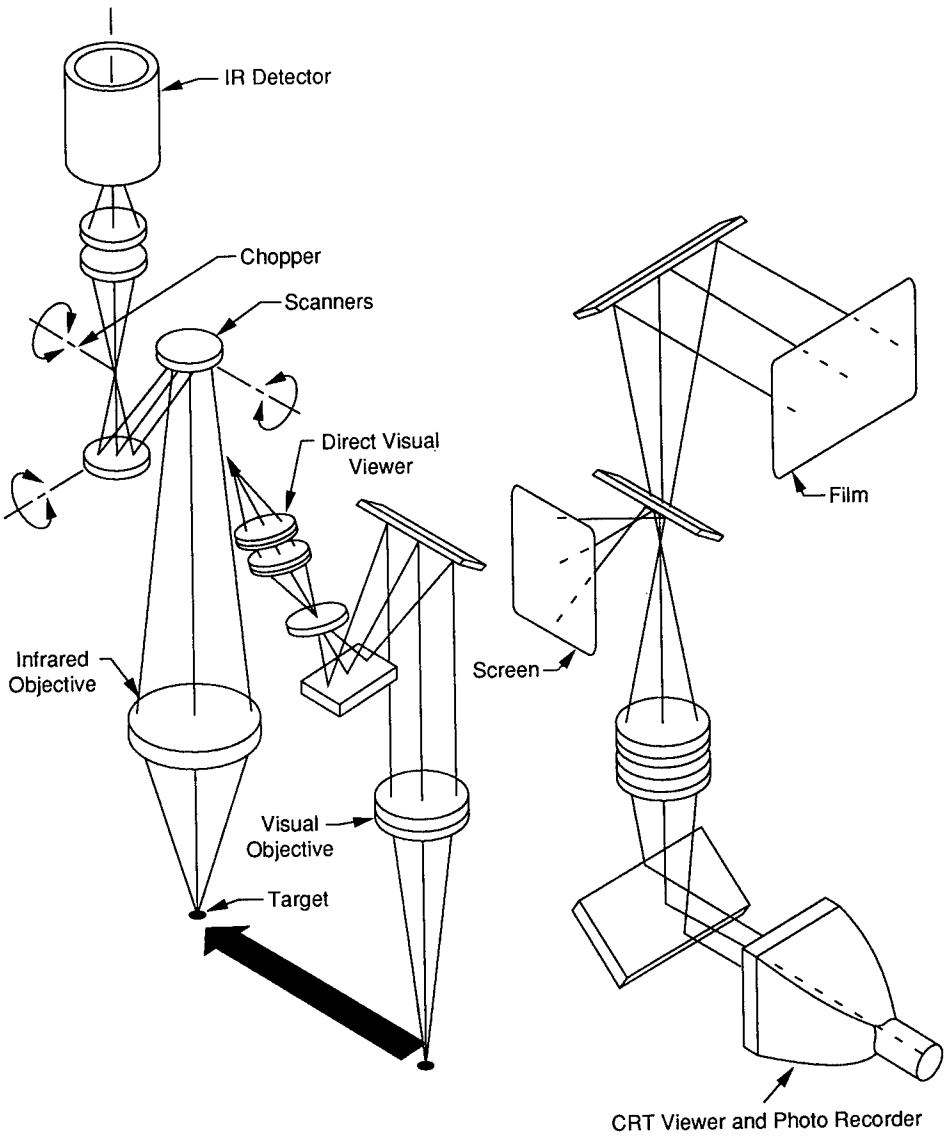


Fig. 3.1 Scanning infrared microscope. (Courtesy of EDO Corporation/Barnes Engineering Division)

- *Accuracy and temperature stability:* Because of the inherent drift of electrical devices and of materials employed in mechanical configurations, the transfer function of motors, as well as of position detectors, drifts with changes in temperature. Special measures may be needed to compensate either in the scanner itself (including the electronics) or at a higher system level.

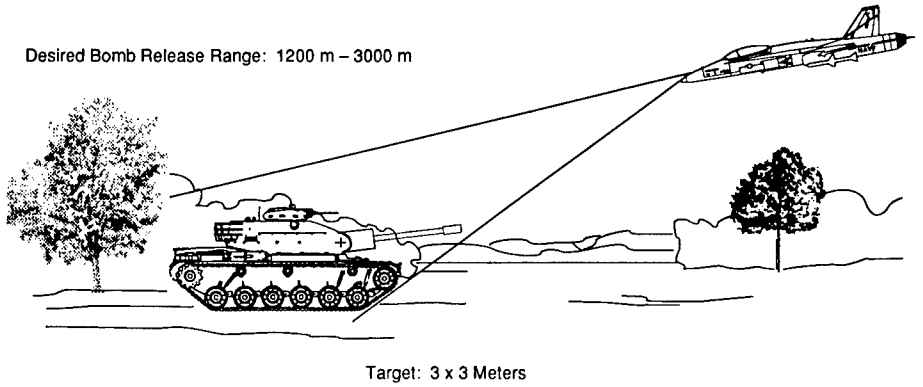


Fig. 3.2 Typical target.

3.3.2 Tracking: Missile Launch

For missile launch tracking, the target is, for example, a laser beam reflected off the missile nose cone. The target is tracked by a gimbal for coarse tracking and high-speed piezoelectric devices for fine tracking. Maximum angular velocity and acceleration are expected to be 4×10^5 arcsec/s and 14×10^5 arcsec/s^2. Accuracy is such that the laser beam is scanned to locate the missile within a field of 1 deg to an accuracy of 0.3 arcmin. The speed of response is measured by the dwell time. The dwell time on the object is 1.52 μs, which requires a receiver bandwidth of 330 kHz.

3.3.3 Tracking: Crosslink Satellite Communication

The targets are two geosynchronous satellites, each 40,000 km from earth and located 84,000 km from each other. The travel time of the signal from satellite 1 to satellite 2 and return, assuming no communication delay, is 0.55 s, during which time satellite 1 travels 1.67 km. The divergence of the communication beam is approximately 10 μrad, which converts to a 0.83-km spot at the receiving satellite (see Fig. 3.3).

Position accuracy must be in the microradian range over the operating environmental conditions. A 5-μrad system position error would cause a loss of contact between the two satellites. A typical accuracy requirement for open-loop pointing is 1 μrad. (A two-axis scanner is used.)

In the event of loss of contact, the expected acquisition time is typically 60 s.

3.3.4 Tracking: SDI Beam Steering

A tracking system (see Fig. 3.4) consists of separate surveillance and weapon platforms. Passive scanners perform target acquisition and coarse tracking. A laser illuminator is brought into play for fine tracking.

When located on a geosynchronous satellite, the surveillance scanner has an active fine tracking field of view of 0.25 μrad in two axes. The absolute accuracy should be 0.1 μrad or better in order to contact a typical target. The

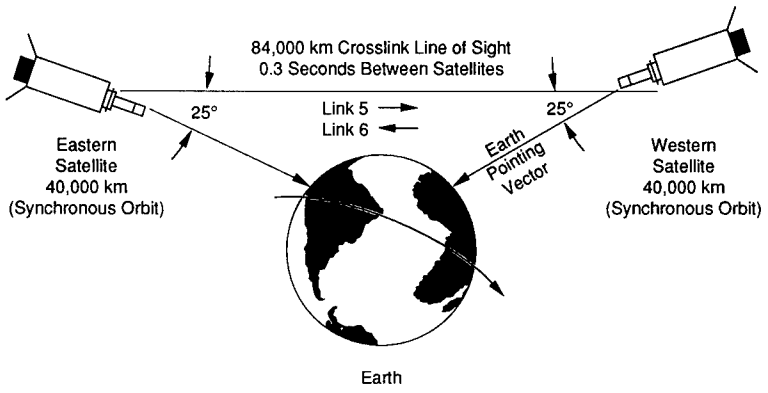


Fig. 3.3 Crosslink satellite communication.

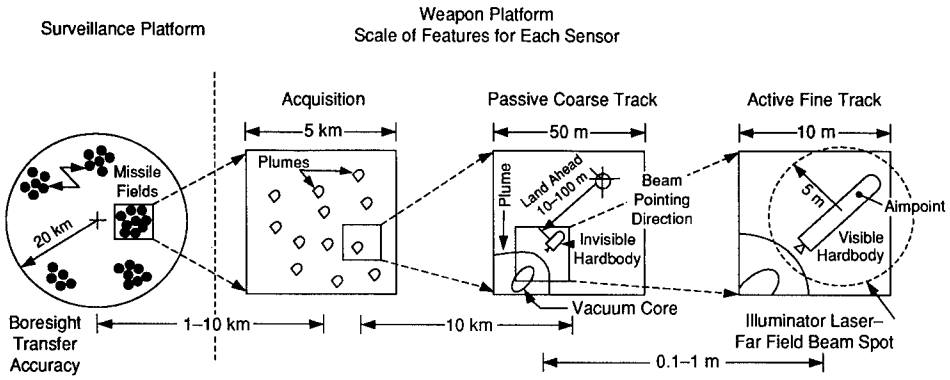


Fig. 3.4 SDI target acquisition.

bandwidth for an SDI scanner is derived from the speed of travel of a missile leaving the atmosphere. It requires millisecond response time.

3.3.5 Infrared Imaging: Two-Axis FLIRS

Typical performances of a two-axis FLIRS:

- *Field of view:* Horizontal view angle is 23 deg. The vertical viewing angle is 17 deg.
- *Scan format:* 30 frames per second, 60 fields per second (2:1 interlace).
- *Image quality:* Thermal resolution is 0.1°C. Image resolution is comparable to that of commercial TV (524 picture elements per scan line) (see Fig. 3.5).

3.4 SCANNING TECHNIQUES

This section describes the generic types of scanning devices and their performance characteristics when used in infrared scanning systems. Section 3.5

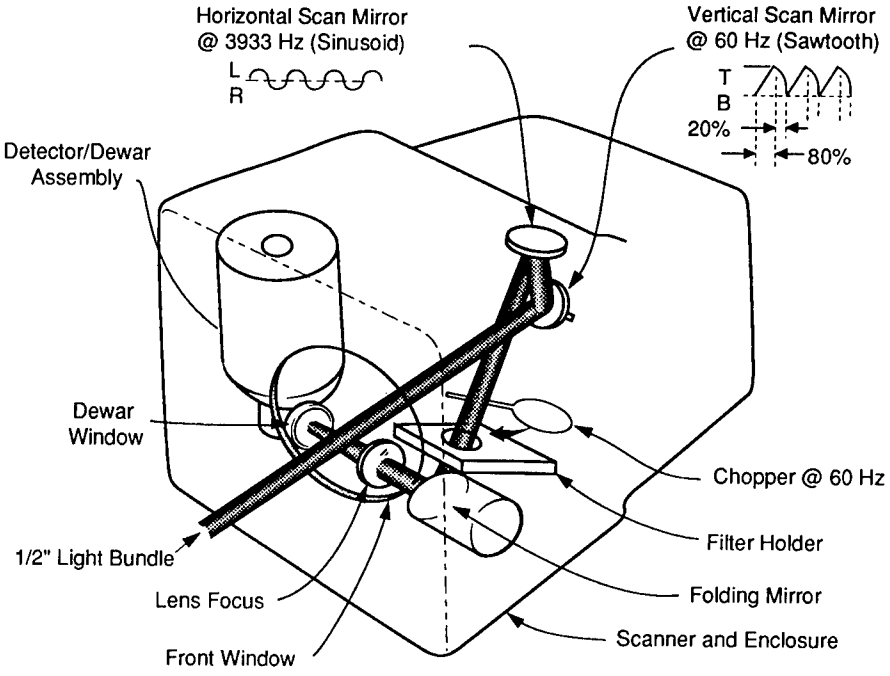


Fig. 3.5 Forward-looking infrared sensors. (Courtesy of Inframetrics, Inc.)

describes actual scanning systems employing one or a combination of the devices described here.

3.4.1 Review of Scanner Applications

Different scanning methods are suitable for different scanning applications. For the purpose of evaluating the major scanner types, infrared scanning applications may be categorized as follows:

Designating, imaging, and radiometric scanners generally scan in a raster pattern over a wide field. (An exception is some ground-based imagers that scan in a cycloid pattern.) Key scan characteristics are high precision, linearity, and, often, high speed. Thermal imagers are passive devices, but designators and radiometers can be either passive or active—combining an outgoing laser beam and a passive receiver.

Tracking scanners generally scan in a cycloid or rosette rather than a raster pattern. They require dynamic control over spot position with short access times. As discussed in Sec. 3.2, both active and passive trackers are in use.

Actual systems are often hybrid combinations of the above types: designator-imager-tracker systems, for example.

3.4.2 Major Classes of Scanner Types

A scanner by its nature directs a moving beam of energy—in this case, infrared energy. The direction of the beam is determined either by reflection with a

mirror or by refraction with a prism. (Another type of refractive scanner, a rotating hologram, has not been applied to the kinds of infrared uses discussed here.) Movement of the beam is effected either by an electromechanical or piezoelectric device that moves the mirror or prism in the energy path, or, in the case of electro- or acousto-optical scanners, by applying electric or sound waves to change the refractive index of a photoelastic material.

Scanners are sometimes categorized as rotary or oscillatory devices or as high-inertia or low-inertia devices. In rotary (high-inertia) scanners the optical element is a refractive prism, reflective polygon, or disk of lenses or mirrors that rotates continuously in one direction. In oscillatory (low-inertia) scanners the optical element is either an oscillating mirror or an electro-optic or acousto-optic deflector.

3.4.3 Rotating Scanners

In a polygon or disk scanner, a prism or a number of mirrors or lenses are arranged concentrically around an axis. The rotating device moves each facet, mirror, or lens through the energy path, creating a regular, repetitive scanning pattern.

With a polygon scanner, its faces are the reflective or refractive surfaces. A disk scanner has mirrors or lenses (or holographic lenses) arranged on its flat surface.

Limitations on Scan Rate.[a] Several factors are influential in limiting the rapidity of the rotation or oscillation of a scanning element. These include the forces generated, the strength of the materials, size, weight, lever arm, windage, and friction heating. For a polygonal scanner the maximum rate $\dot{\omega}$ is given by

$$\dot{\omega} = \frac{1}{2\pi r_o}\sqrt{\frac{8\text{UTS}}{\rho(3+\eta)}}, \tag{3.1}$$

where

r_o = the distance from the center to an edge
UTS = the ultimate tensile strength of the (solid) mirror
η = Poisson's ratio for the mirror material
ρ = its volumetric density.

The maximum would probably be reached before this because of mirror deformation.

Polygons. Polygon scanners nearly always rotate continuously in one direction at a constant speed. Thus, high scan rates are easily attained and relatively large optical elements can be used. Considerable variety in the designs is displayed with regard to the number of facets, their orientation—outward-facing or inward-facing—and the angle of each facet with respect to the axis

[a]This section is from W. L. Wolfe, "Optical-Mechanical Scanning Techniques and Devices," Chap. 10 in *The Infrared Handbook,* W. L. Wolfe, G. J. Zissis, Eds., Environmental Research Institute of Michigan, Ann Arbor (Revised 1985).

of rotation. A polygon whose facets make incrementally greater angles with the axis can produce an interlaced or overscanned pattern.

Types of Design. Inverted (inward-facing) polygons are conducive to compact scanner designs. Missile guidance and reconnaissance systems use such scanners, and when compactness is important, so do other systems.

Pyramidal polygon scanners, whose facets are at an angle to the axis of rotation, are more useful for producing smaller scan angles with fewer facets than polygons whose facets are parallel to the axis of rotation. Forty-five degrees is the most common pyramidal polygon facet angle.

Like mirror polygons, refractive polygon or prism scanners come in a variety of shapes. Figure 3.6 shows a few of them. Pyramidal prisms, or wedges, are often used in pairs to generate a variety of scan patterns, including circular, elliptical, rosette-shaped, and spiral. One of the earliest FLIRS was a ground-based scanner that used two rotating refractive prisms to create a spiral scan pattern.

Advantages and Disadvantages of Polygon Scanners. Polygons scan both at very slow and very fast rates (8 to 50 kHz), as well as small and large (approaching 180 deg) angles. The chief advantages of polygonal scanners are their high scanning speeds, potentially large scan angles, and velocity stability. Conversely, their high inertia makes them unsuited either for rapidly or frequently changing scan velocities. It also necessitates long startup times in addition to large starting and control power. The high speeds and high inertia

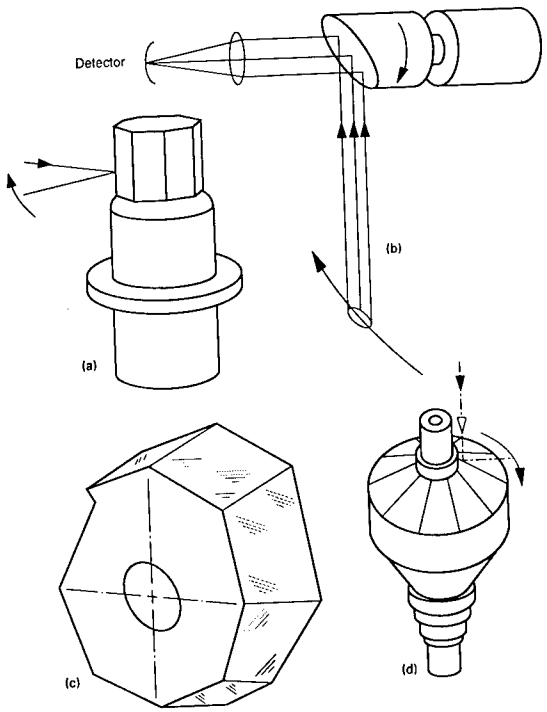


Fig. 3.6 Polygon scanners. [Parts (a), (b), and (d) from Ref. 1, © 1991 Marcel Dekker; part (c) from Ref. 2, © 1991 Marcel Dekker.]

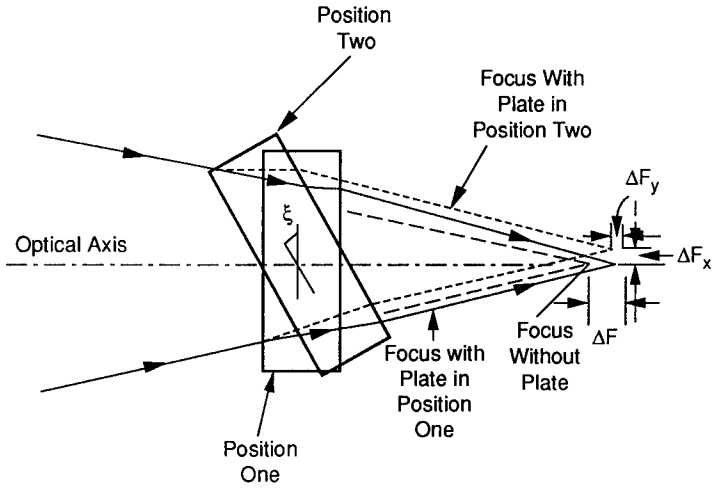


Fig. 3.7 Rotating parallel plate. Geometrical representation of change of focus due to insertion and rotation of a plane parallel plate.

of polygon scanners occasionally cause other problems too—mirror distortion, bearing wear, vibration, noise, and gyroscopic effects.

Scan efficiency, another issue with polygon scanners, is affected by the dead time as the light beam crosses from one facet to another. Obviously, the more facets, the greater the dead time. There are various ways to counter this effect—for example, by increasing the facet size (decreasing the dead time in proportion to the scan time) or, conversely, by making the beam diameter larger than a facet, so that a new scan begins before the previous one is completed.

Disc Scanners. One of the earliest scanners, the Nipkow disk scanner was used in early television systems and is described by Egger.[4] This disk contained a number of holes arranged in a spiral pattern. Its energy gathering efficiency is extremely low and subsequent infrared scanning systems have rarely incorporated Nipkow disks. Recently, however, a very compact infrared imaging scanner was developed that incorporates an oscillating flat mirror for the vertical scan and a disk containing concave spherical mirrors for the horizontal scan. This system is described in more detail in Sec. 3.5.

There is little history to go by regarding the performance characteristics of disk scanners. One can say, in any case, that they share with polygons the simplicity of unidirectional rotation at a steady speed, while their slimmer profile makes them less vulnerable to the dangers of high inertia.

Oscillating or Rotating Plane Parallel Plate.[b] A ray is refracted through a plane parallel plate with displacement but no deviation as shown in Fig. 3.7. If the plate is tilted at an angle ξ, and rotated about the optical axis, then a

[b]This section is from W. L. Wolfe, "Optical-Mechanical Scanning Techniques and Devices," Chap. 10 in *The Infrared Handbook*, W. L. Wolfe, G. J. Zissis, Eds., Environmental Research Institute of Michigan, Ann Arbor (Revised 1985).

circle will be generated. If the beam is collimated, it will only be displaced. A similar action occurs if the plate is oscillated or rotated about the axis perpendicular to the paper, but then the scan is vertical. These motions do little good unless coupled to an imaging system. This is one of the few scanning schemes that performs linear displacement of a beam rather than angular deviation. The change in focal length, ΔF, caused by the insertion of a plate with plane and parallel sides is given by

$$\Delta F = d\left(1 - \frac{\cos\theta_i}{\sqrt{n^2 - \sin^2\theta_i}}\right) , \tag{3.2}$$

where

d = the plate thickness
θ = the angle of incidence on the front face
n = the refractive index of the plate.

The focal shift is a function of the angle of incidence. Figure 3.7 shows how this changes with angle (from 0 to 30 deg). These values stop when the angle is sufficiently large that $n \sin\theta = 1$, the angle of total internal reflection.

The change in focus is a function of wavelength because of dependence of the refractive index on wavelength.

$$\frac{\partial}{\partial\lambda}\left(\frac{\Delta F}{d}\right) = \frac{1}{d}\frac{\partial(\Delta F)}{\partial\lambda} = \frac{n\cos\theta_i}{(n - \sin^2\theta_i)^{3/2}}\frac{dn}{d\lambda} . \tag{3.3}$$

The position of focus moves as the plate is rotated through an angle α. The angle of incidence changes as the rotation angle. Two rays at equal angles from the axis of the beam have angles with incidence of $\theta_i + \xi$ and $\theta_i - \xi$. If the cone angle is 2θ, then the location at which these rays cross (their focus) is given by

$$\Delta F_y = \frac{1}{2}(\Delta F_1 + \Delta F_2)\tan\theta , \tag{3.4}$$

$$\Delta F_x = \frac{1}{2}(\Delta F_1 + \Delta F_2) , \tag{3.5}$$

where

$$\Delta F_1 = d\left[1 - \frac{\cos(\theta + \xi)}{\sqrt{n^2 - \sin^2(\theta + \xi)}}\right] , \tag{3.6}$$

$$\Delta F_2 = d\left[1 - \frac{\cos(\theta - \xi)}{\sqrt{n^2 - \sin^2(\theta - \xi)}}\right] . \tag{3.7}$$

3.4.4 Oscillating Scanners

A second type of electromechanical scanning device moves a mirror in a limited angular direction, rather than rotating it continuously. Such devices use both small and large mirrors (0.1×0.15 to 2.0×3.5 in.) and are characterized as *low-inertia scanners*. The actuator may consist of either a direct-drive galvanometer, a cam drive, or a dc motor, or it may be configured as a resonant scanner. The mirror, which is attached to the rotor, can pivot either at a center axis or at one of its edges (a so-called "paddle" scanner).

An oscillating scanning system consists of three elements: a mirror, the scanner proper (a driver and position transducer, tachometer or spring, and a suspension), and driving electronics (amplifier, power supply, and logics). The two types of oscillatory optical scanners are galvanometric and resonant.

A galvanometric scanner can follow any shape signal with a fidelity limited by its transfer function and internally generated perturbation as well as its sensitivity to external perturbations.

Resonant scanners traditionally oscillate sinusoidally at one frequency only. More recently, resonant scanners have been made with dynamically tunable resonant frequency. Also, some resonant scanners are made to oscillate in a triangular motion. Again, the fidelity of motion of a resonant scanner has constraints similar to those of a galvanometric scanner plus others.

3.4.5 Galvanometric Scanners

The tables in Sec. 3.6 list a number of technologies of magnetic and inductive torque motors as well as a number of angular position transducers. Those are not the only critical elements of high-performance scanners, but they are often perceived as such under the assumption that the buyer supplies mirrors and electronics. This is not commonly the most economical solution nor the most expeditious because the entire scanning system should be designed as a system.

Electromagnetic Drivers. The torque transducer of a scanner should be selected for its ability to integrate with the other elements of the scanner, mirror, sensor, and electronics. The ideal galvanometric or resonant scanner driver would have the following properties:

- high torque-to-inertia ratio
- a linear relationship between torque, current, and angular position
- an armature rigid in bending and torsion
- an armature dynamically balanced
- some mechanical damping—constant and stable
- immunity to all external dynamic perturbations
- freedom from any self-induced dynamic perturbations
- immunity to external thermal or rf perturbations
- low power consumption and good heat dissipation capability
- demagnetization protection from temperature, current overload, or vibration causes
- infinite life with stable parameters.

The complete list of features of an ideal driver is long. Frequently, a compromise must be reached where some necessary properties are obtained through other means.

Figures of Merit of Galvanometric Drivers. The multiplicity of technologies and applications makes selection of a galvanometer driver difficult. A number of figures of merit have evolved:

1. the first uncontrollable resonance, normally the first cross-axis resonance
2. the acceleration, also expressed as the torque-to-inertia ratio (T/J)
3. the thermal dissipation capability.

The lowest uncontrollable resonance (on-axis or cross-axis) of the scanner's armature with its mirror limits the practical bandwidth of the servo system of the instrument. In a real product, it is difficult to escape exciting the resonance. Among the palliatives available are:

- added mechanical losses or damping
- critical balance of the mechanical system to avoid exciting the resonance
- soft mount for the scanner to isolate it from environmental perturbations
- inclusion of a notch filter in the feedback loop.

The T/J ratio determines the minimum stepping time t or step response of a second-order system with armature of inertia J and driven through a total angle θ by a motor capable of a bidirectional maximum torque T:

$$t = 2(\theta J/T)^{0.5} \quad . \tag{3.8}$$

The electronics must be able to carry the necessary current to deliver the torque. For moving magnet and moving coil devices, the power dissipation is the most meaningful measure of capability because it is possible to adjust windings to show nearly any torque constant. Catastrophic failure is shackled to the peak temperature experienced by the drive coil. That temperature is determined as much by the thermal resistance of the path to the mount as the necessary power input.

As usual, the selection of applicable parameters is a compromise where the available power, the load inertia, the necessary bandwidth, the duty factor, and the cost (not only of the magnetic device, but also of the driver amplifier and its power supply as well as supporting chassis) come into play with all of the features listed earlier.

Moving-Coil Drivers. In most cases moving-coil devices are the first to be considered. As demand for higher performance systems develops, the other designs compete because they offer higher rigidity, immunity from centrifugal deformation, and a solution to the need to heat sink the coils (see Ref. 3).

The need for low inertia limits the rigidity of the moving-coil structure. To circumvent the accuracy consequences of this deficiency, the position detector should be located on the output shaft. This leads to an extra long output shaft. A large-diameter shaft achieves the needed torsional rigidity. This extra length lowers the cross-axis resonance, and dedicated efforts to achieve well-balanced mirror and rotor construction minimize the dangers of self-excitation of this mode. The enlarged shaft facilitates heat conduction to the adjacent position sensor that needs high-performance electronic temperature compensation. The conventional environmental temperature controls only exacerbate tempera-

ture drifts because the moving element of the position transducer is the major heat sink for the coil-induced heat.

Some applications are well suited to the features of moving-coil scanners. In addition to low-duty factor operations, these applications have vibration- and temperature-protected environments.

Moving-Iron Drivers. Moving-iron drivers offer the best choice of temperature torque stability because they are built with ALNICO magnets, which have a temperature coefficient 10 times lower than those of rare-earth magnets and a Curie temperature of more than 500°C. Their major flaw, nonlinearity, can be corrected. Compensation flux paths have been shown by Montagu[5] to yield devices with very linear torque/current/position relationships. Their blend of features makes them, to this date, practically the only choice for military and aerospace applications. Their major shortcomings are their manufacturing complexity and cost.

Moving-Magnet Drivers. The most recent contenders as drivers for high-performance optical scanners are moving-magnet drivers. The availability of high-energy material brought them into the competition. The coil of a moving magnet scanner is stationary and well heat sunk. The rotor is cylindrical with a high natural frequency. This technology is capable of beam excursions of 120 deg and offers the highest torque per unit weight.

This technology has all the desirable features of moving-coil devices and none of the shortcomings. For most scanner applications, it satisfies all the requirements listed earlier to acceptable degrees. It is currently an attractive choice for manufacturers of low- and medium-speed scanners when the magnet is protected from temperature extremes; 120°C is a maximum.

Position Transducers. Position transducers must accurately convert the angular position of the rotor into a convenient electrical signal to be processed by the control electronics. An ideal sensor would have the following characteristics:

- high resolution and a high signal-to-noise ratio (SNR)
- a bandwidth a few orders of magnitude beyond the desired system bandwidth
- a linear transfer function
- mechanically balanced with low inertia and free of resonances within the range of interest
- insensitive to wobble or radial motions
- insensitive to environmental variations such as pressure, temperature, vibrations, humidity, and time
- low cost and easy to interface.

Most actual transducers fall short of the ideal and compromises are reached. The simplest is the torsion bar of the open-loop scanner. Optical interferometric encoders are made with low inertia and extremely high performance. Their cost and complexity limits their appeal.

The most common position sensor is the capacitive detector. When higher gain and null precision are necessary, two points of detection in the field of view or within the scanner are used for correction. Figure 3.8 shows such a unit on a galvanometric scanner.

Capacitive Sensors. All single-axis galvanometric scanners in the following sections use the capacitive bridge developed by Abbe.[6] Three different capacity plate configurations are represented:

1. One configuration is a cylindrical plate arrangement with a moving conductive vane. The stator is encapsulated with the magnetic driver, and the driver's rotor forms the moving vane. This is the most compact and economical design to build and offers a high SNR with low inertia penalty. The temperature sensitivity of the dielectric encapsulant is difficult to control, so the scanners need good temperature controls.
2. The same general construction as above but the detector is built as an independent unit and offers the benefits of a high SNR without the thermal drift. Manufacturing costs are higher, and the practical accuracy limit of this design is around 25 μrad.
3. Moving dielectric capacitor construction offers the advantage of lesser sensitivity to all radial motions. In 1944 Ergen and Petran[7] explained: "An important advantage . . . over the conventional type of variable condenser in which the movable vane is of electrically conductive material lies in the fact that the vane need not be accurately parallel with the plate of the condenser." Ergen recommends a flat or butterfly construction that is more economical to build. He also shows how to compensate temperature drifts with auxiliary capacitors outside of the reach of the moving vane.

   To benefit from these advantages, a moving vane transducer has a comparatively high inertia. This design can be justified for applications where errors due to micromotions from radial and axial bearing random imperfections are not tolerable.

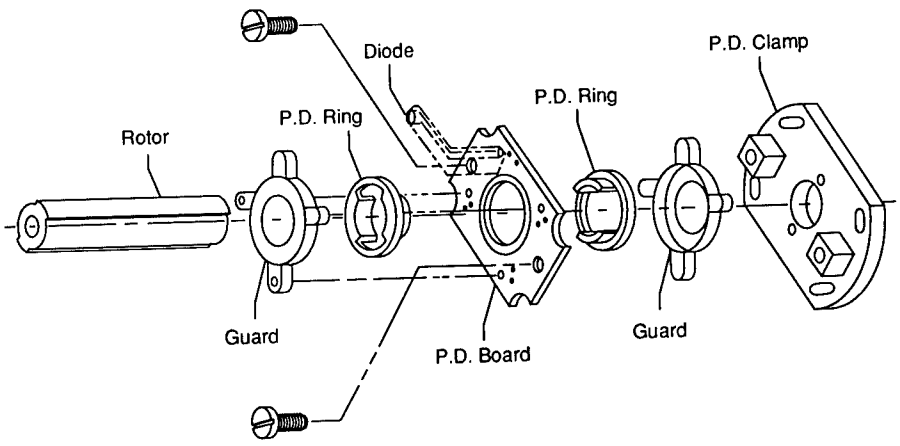


Fig. 3.8 Capacitive detector on galvanometric scanner.

3.4.6 Resonant Scanners

Resonant scanners can drive large mirrors at wide amplitudes and high repetition rates with a low power requirement and exhibit submicroradian wobble and jitter. They compete with rotating scanners, and for scanning angles under 0.5 rad they offer a much smaller package. They are also expected to satisfy a wish list similar to the ideal galvanometers discussed earlier in the subsections on "Electromagnetic Drivers" and "Position Transducers."

The attractive features of resonant scanners have been acknowledged for years, but these advantages have been overshadowed by their deficiencies. Recent developments call for a reassessment of the raster scanning technology.

Advances of the technology come from the availability of new materials and a better understanding of the problems.

- *Jitter:* Resonant scanners can be viewed as galvanometric scanners with extremely low damping or as high-Q filters. As such, they can reproduce the drive signal with all of its low- and high-frequency noises filtered but not eliminated. A high-resolution system requires a clean drive signal.

- *Mirror:* The mirror design is very complex because it is subjected to extreme high acceleration and deceleration forces. The choice of material is critical. Advances in ceramics and replication technologies have opened new possibilities. The major breakthrough has come with the development of techniques to fasten rigidly materials with dissimilar coefficients of temperature expansion and have the mirror remain flat to one-tenth wave over a large temperature range.

- *Perturbations:* Resonant scanners have been known to always find a hidden resonance somewhere in the instrument. Soft mounting is normally not possible because it destroys alignment. They also have been known to be susceptible to external shocks and vibration. Both of these shortcomings have been corrected: The first one by balancing the armature rather than increasing the mass of the stator; the second one with a judicious design where the armature is mounted with an anisotropic means. It is rigid in torsion but soft and lossy in compression. Understanding of air-buffeting perturbations has guided the mirror and environment designs to bypass the Reynold's number "sound barrier."

- *Timing correction:* Timing affects information processing or presentation in two ways: (1) the location in time of data bits with their location in space, knowing that transfer function, the mirror motion, is sinusoidal, and (2) timing drifts of the natural frequency of the resonant scanner versus the system's master clock. High-speed and high-precision/resolution imaging processing have data rate consequences that render the brute force approach impractical. Though each case is different, a general solution can be derived if the resonant scanner is so accurate as to be the master clock. (If this is not practical, tunable resonant scanners are viable.)

Triangular Wave and Sawtooth Resonant Scanners. A conventional resonant scanning system is inherently limited in that the scanning rate varies throughout the entire scan field. Nonlinear scan velocities due to sinusoidal drive may be compensated for by a variety of mechanical, optical, and/or electronic techniques. Digital encoding techniques are perhaps most widely used today to compensate or adjust for the sinusoidal scan rates.

An alternative method of improving scan linearity of resonant systems is through the addition of multiple phase-locked resonant harmonic frequencies to the fundamental scanning mechanism. Figure 3.9 illustrates the error of a triangular scan pattern that results from an arithmetic progression of sinusoidal components. Several methods of implementing cascaded optical systems of this type are reported in current literature. The linearity of the overall system continues to improve as the number of additive harmonic frequencies is increased. However, the complexity of the overall system increases rapidly with the number of harmonic frequencies.

The Fourier expansion of a triangular wave is of the form

$$f(t) = A\left(\cos\omega t + \frac{1}{9}\cos 3\omega t + \frac{1}{25}\cos 5\omega t + \ldots\right) . \tag{3.9}$$

It is apparent that a scanner system representing the fundamental and third harmonies can yield a pointing linearity of 1% best straight line with a high efficiency.

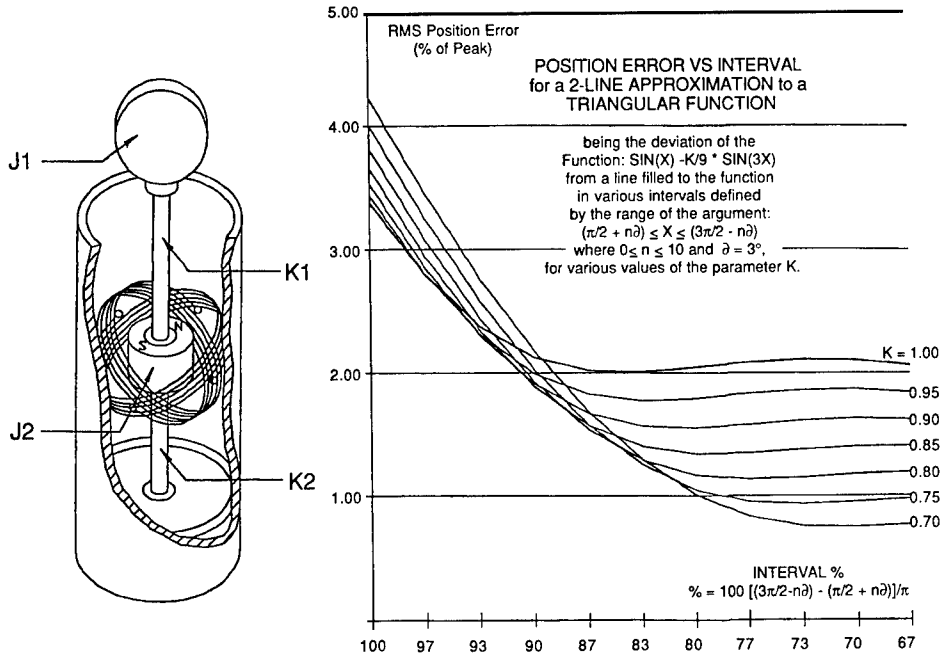


Fig. 3.9 Triangular wave resonant scanner.

The Fourier expansion of a sawtooth is more complex. Similar error for a sawtooth with three-quarters of the pivotal action requiring the presence of the second and fourth harmonies as its expansion is of the form:

$$f(t) = A\left(\cos\omega t - \frac{1}{4}\cos 2\omega t - \frac{1}{16}\cos 3\omega t - \frac{1}{25}\cos 5\omega t - \ldots\right) . \qquad (3.10)$$

Caveat: All frequencies and phases have to be controlled. This is possible with tunable resonant devices.

Tunable Resonant Scanners. A special kind of resonant scanner with torsion bar suspension or cross-flexure is designed to permit frequency tuning within a limited range. This is accomplished by altering the spring rate, either by incorporating a magnetic tuner or by changing the environmental temperature.

Table 3.2 shows the performances of the unit in Fig. 3.10, an ISX model from General Scanning with a magnetic tuner built on-axis. The frequency tuning section consists of a permanent magnet similar to that of the driving section.

The other method of controlling the spring rate of a scanner torsion bar is to control the environmental temperature of the unit. Tunability is much enhanced when the torsion bars are made of Nitinol, whose Young modulus can be changed by as much as 10% without affecting the performance life of the scanner. Figure 3.11 compares the performances of an IDS scanner fitted with a choice of torsion bars.

3.4.7 Piezoelectric Scanners

Another type of limited-rotation reflective scanner is the piezoelectric device, based on the deformation—shear or expansion—of certain ceramic materials when subjected to an applied electric field. Individual strips, disks, or devices of such materials are laminated, stacked, or ganged in order to multiply the effect to a useful degree. A mirror is affixed so as to tilt in response to the bending action and in proportion to the applied voltage. Some examples of piezoelectric scanning devices are shown in Fig. 3.12.

Piezoelectric scanners are capable of high frequencies (45 kHz or more), but only very small excursions (typically up to 0.05 deg). They require high voltages

Table 3.2 Tuning Performances of Mini-ISX Electromagnetically Tunable Resonant Scanner

Natural frequency	200 Hz
Tuning range, total	1.6 Hz
Bandwidth, minimum	20 Hz
Spring rate, mechanical	5.3×10 g cm rad^{-1}
Spring rate, magnetic	100 g cm rad^{-1} A^{-1}
Tuning coil, resistance	26.6 Ω
Tuning coil, inductance	1.2 mH
Tuning torque time rise	0.1 ms
Rotor inertia	3.31 g cm^2

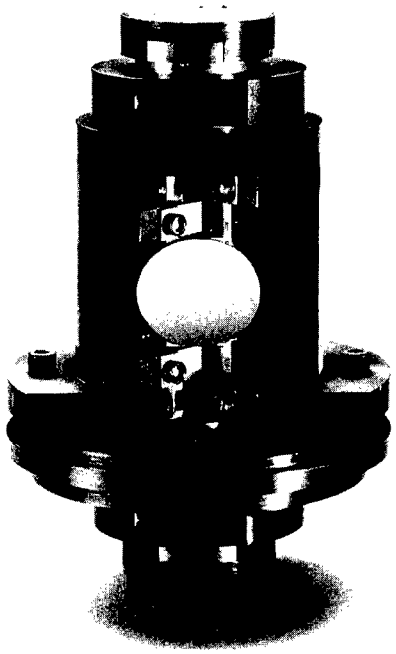


Fig. 3.10 Mini-ISX tunable resonant scanner.

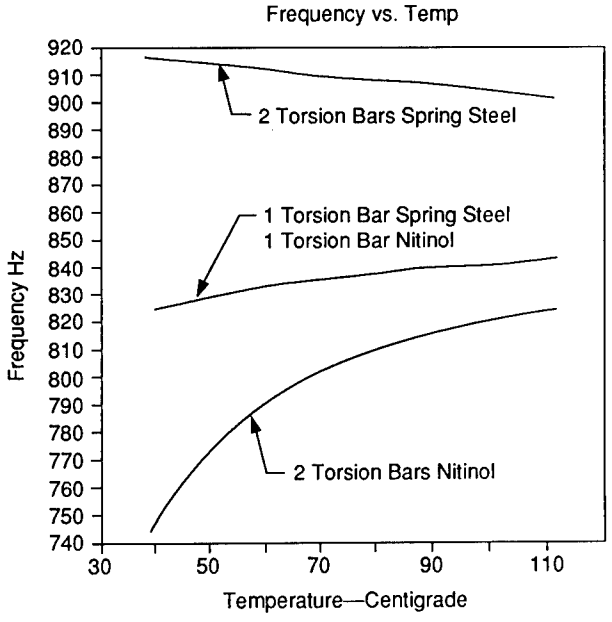


Fig. 3.11 Tuning performance of IDS resonant scanner.

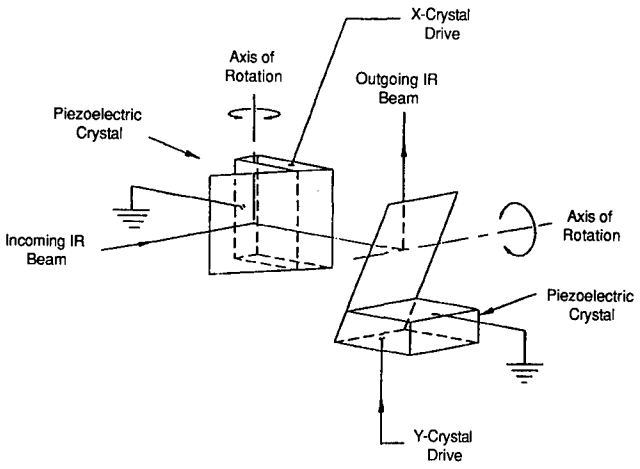


Fig. 3.12 Piezoelectric scanning devices.

to scan. A 1989 design (shown in Fig. 3.12), containing two cantilever-mounted bimorphs and a feedback control loop, reportedly is capable of + 1.7 deg angular deflection with a bandwidth of 80 Hz.

With their small excursions, piezoelectric scanners are found in such applications as missile launch trackers, communication satellite alignment, and the interlace positioner of single-axis "common module" thermal imagers.

3.4.8 Acousto-Optic and Electro-Optic Scanners

Electro-optic and acousto-optic scanners control beam movement by the application of voltages or ultrasonic sound waves to photoelastic materials. Varying this input changes the refractive index of the material or—in the case of digital electro-optic reflectors—by changing the polarization of the incident light beam and thus the deflection of the beam. Their angular deflection is proportional to the wavelength of light, which limits their application to infrared image acquisition.

The primary advantage of acousto- and electro-optic scanners is speed. They are capable of less than a microsecond of random access time, but they have aperture and scan-angle limitations (typically 1 to 2 deg, peak-to-peak). The aperture limitation results in a relatively lower solution compared to that of reflective scanners. It is also more difficult to control the optical and acoustic or dielectric qualities of acousto- and electro-optic materials than to control the quality of a mirror surface. For these reasons, acousto-optic and electro-optic scanners are much less prevalent in infrared scanning systems than electromechanically driven devices and are confined to laser beam deflection applications, for example, active rather than passive scanners.

Acousto-Optic Scanners. An acousto-optic scanning device is shown in Fig. 3.13. It is a glass prism wherein a sound wave is propagated approximately orthogonal to the incident laser beam. The sound wavelength determines the angle at which the beam is deflected. The scan angle is typically less than 3

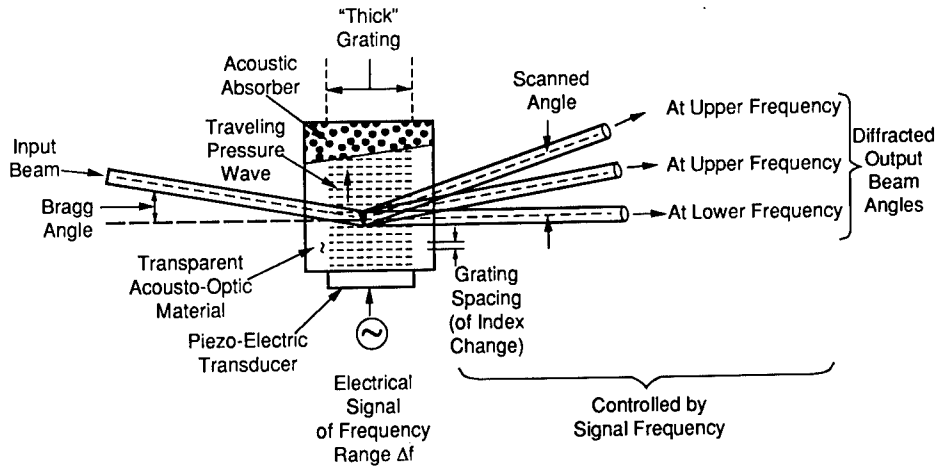


Fig. 3.13 Acousto-optic scanner.

deg; resolution is 1000 spots per line maximum. The merits and limitations of the technology in the IR are reviewed by Gottlieb.[8]

Electro-Optic Scanners. Electro-optic beam deflectors are of two major types: analog and digital. In analog deflectors, an applied voltage varies the phase delay across the incident beam; the phase front is tilted when the beam emerges, and the beam direction is changed by the same angle. In digital deflectors beam direction is controlled by modulation of the polarization direction of the incident light. To generate multiple exit beam directions a number of deflectors are cascaded $2n$ beam positions for n stages.

A digital electro-optic scanner can perform high-speed random access to any of its positions; however, power dissipation limits the deflection rate to about 10^6 bits per second.

Since analog electro-optic deflectors typically are capable of a change in index of refraction of only 1/10,000, they, too, are usually combined in series.

3.4.9 Two-Axis Beam Steering Scanner

Two-axis beam steerer is the term used for systems using only optical elements to scan in two axes. Two design practices are represented. The Ball Brothers design[9] is an example of a heavy support coupled to a counteracting inertia stabilizer. A light mirror with minimal inertial consequence such as General Scanning's flex pivot gimbal design is shown in Fig. 3.14. Both designs use electromechanical actuators for both the x and y directions. The angle of excursion is usually smaller than ± 5 deg. A common application of a two-axis beam steerer is alignment of satellite communication systems.

3.5 EXAMPLES OF INFRARED SCANNING SYSTEMS

Section 3.4 described the general types of scanners used in infrared scanning systems: disk, polygon, galvanometric, resonant, electro-optic, and acousto-

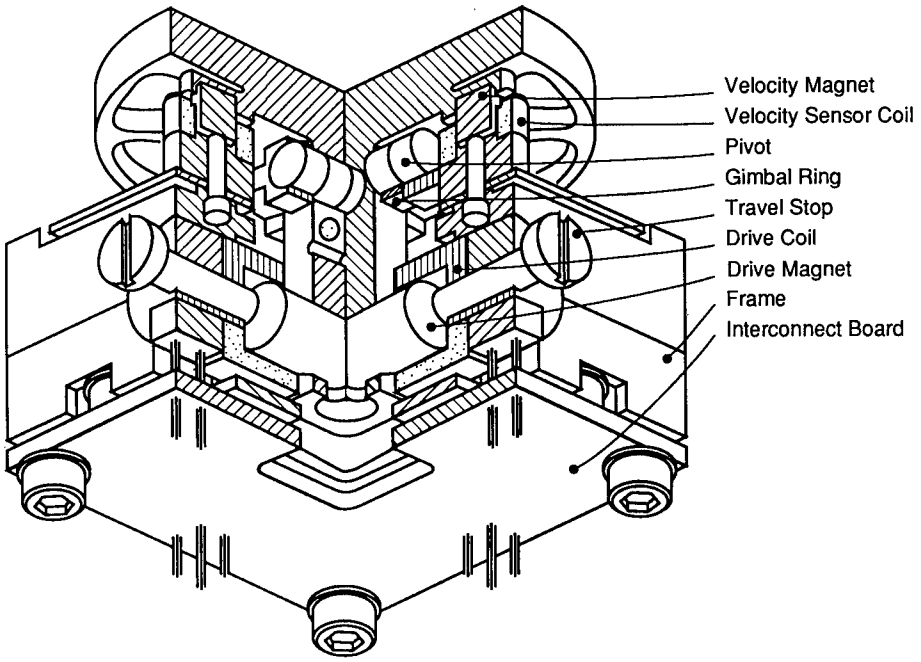


Fig. 3.14 Flex pivot gimbal TABS from General Scanning.

optic. This section reviews some examples of actual scanning systems employing these different types of devices.

3.5.1 Single-Axis Scanning

The most widely used infrared scanning systems are the *common module FLIRS* systems used in passive tactical infrared navigation, target acquisition, and surveillance—both ground based and airborne. There are approximately 40,000 to 50,000 common module scanning systems in use. They are produced by a handful of companies.

Common module scanners employ a brushless dc torque motor to move the scan mirror through an angle from 8 to 20 deg at a frequency up to 60 Hz. The common module scanner also uses a gimbal that allows the mirror to be moved incrementally in the cross-scan direction to provide interlace. An example is illustrated in Fig. 3.15.

A number of single-axis scanning systems use polygons. For example, the so-called "axe-blade" or "knife-edge" system originally developed by Haller-Raymond and Brown (now HRB-Singer) employs two 45-deg mirrors joined at one edge for airborne mapping (Fig. 3.16). As the scanner rotates, the two facets alternately scan the ground.

Another example of a polygon-based one-axis infrared scanner system is the RS 100 line of airborne reconnaissance line scanners built by Texas Instruments and Loral Defense Systems. The optical system shown in Fig. 3.17 has a 120-deg field of view, a four-sided rotating scan mirror, and several fixed mirrors as well as the field lens assembly.

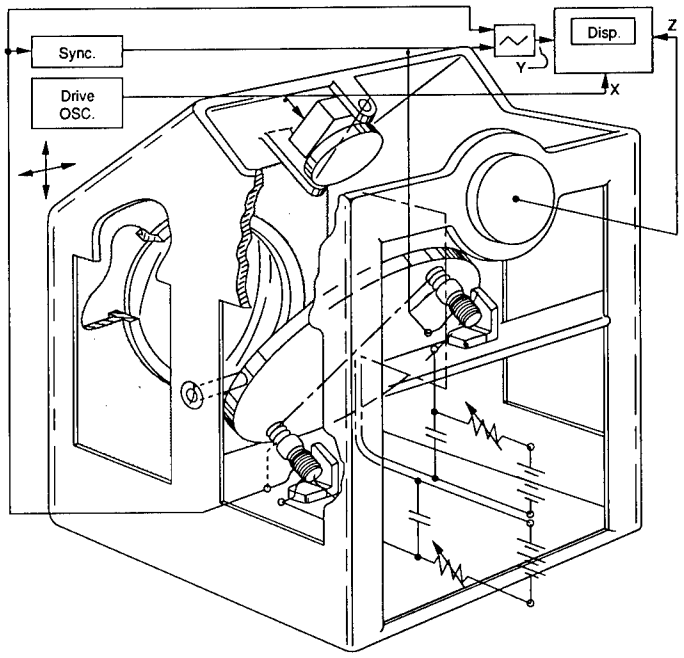


Fig. 3.15 Common module scanner.

An unusual variation on the rotating line scanner is the device shown in Fig. 3.18 (manufacturers Hughes, Honeywell) designed for small aircraft, helicopters, tanks, and periscopes. The use of a polygon and a paddle scanner offers a compact design. The axes of the two scanners are orthogonal and in a common plane. The polygon is diamond-machined and the paddle mirror is made of silicon or beryllium. The paddle must be balanced to be insensitive to vibrations and accelerations.

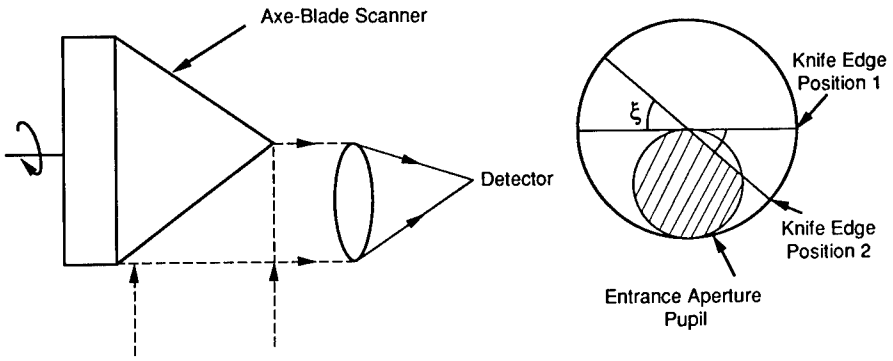


Fig. 3.16 Axe-blade or knife-edge scanning system.

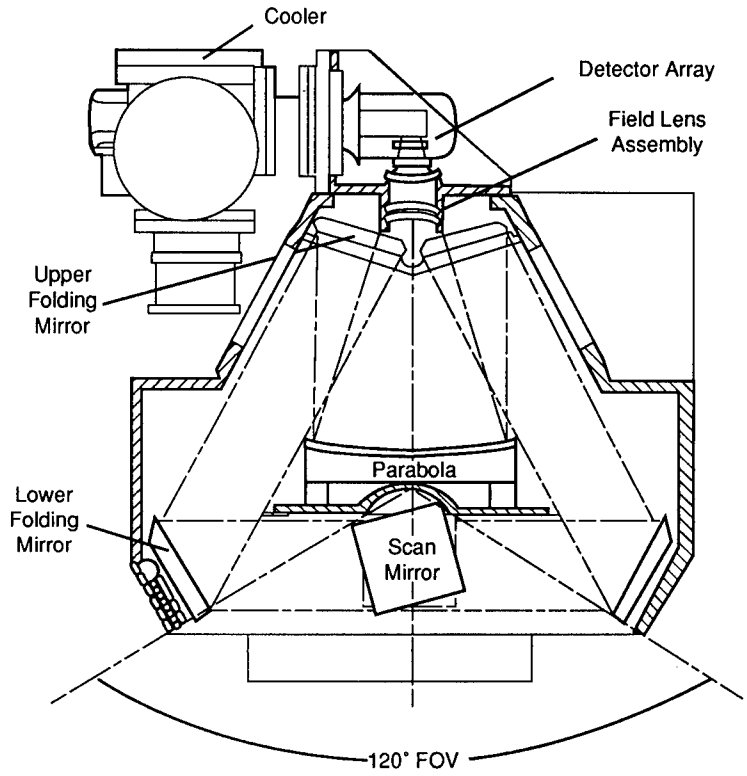


Fig. 3.17 RS-700 infrared line scanner. (Reprinted by permission of Texas Instruments)

3.5.2 Two-Axis Scanning

Scanning systems that must scan in two dimensions often employ a fixed-rate scanner (disk, polygon, or resonant) for one dimension and a variable-rate, usually galvanometric, scanner for the other. Several examples follow.

Polygon-Galvo Systems. Polygon-galvanometer scanners are a common type of infrared imaging system. Three different such designs, patented by Loral, Raytheon Company, and Barr & Stoud Limited, are shown in Figs. 3.19, 3.20, and 3.21, respectively. In each case, the polygon provides the line scan, while the galvanometer mirror provides the frame scan.

The Loral mini FLIRS is a compact TV-compatible, serial-scan FLIR sensor, using an octagonal mirror, approximately 70 mm in diameter and 10 mm thick, for the horizontal scan and a galvanometer scanner, 25 × 10 mm, for the vertical scan. The afocal telescope reduces the diameter of the incoming beam, which makes possible the small size of the polygon (each facet approximately 25 × 10 mm). The polygon rotation speed is 59,062 rpm.

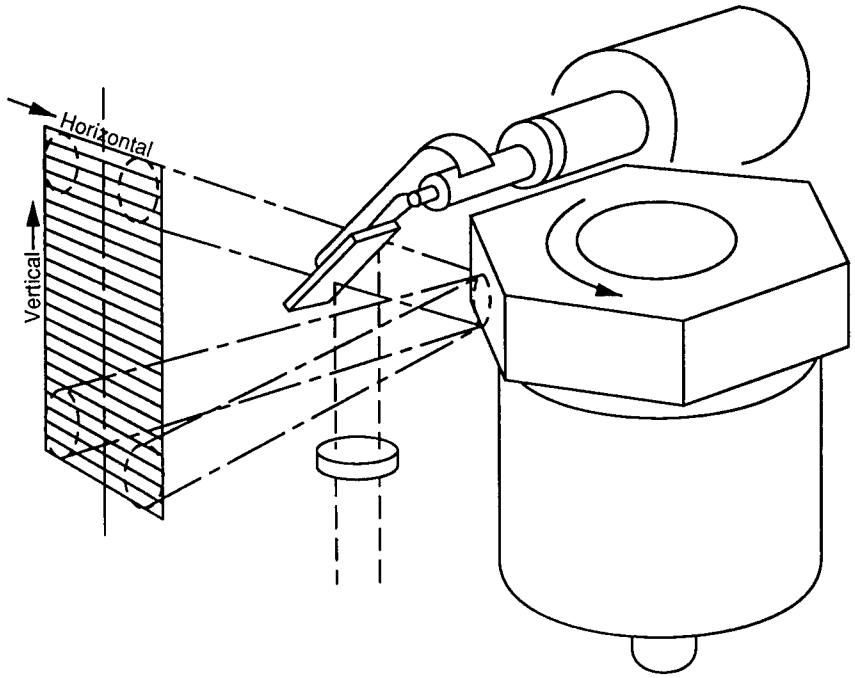


Fig. 3.18 Rotating polygon line scanner with paddle frame scanner.

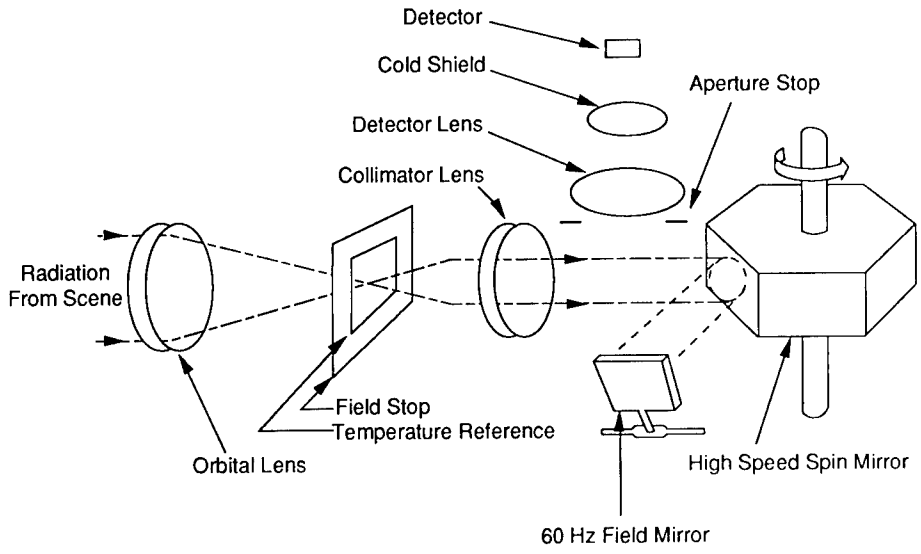


Fig. 3.19 Polygon scanner from Loral.

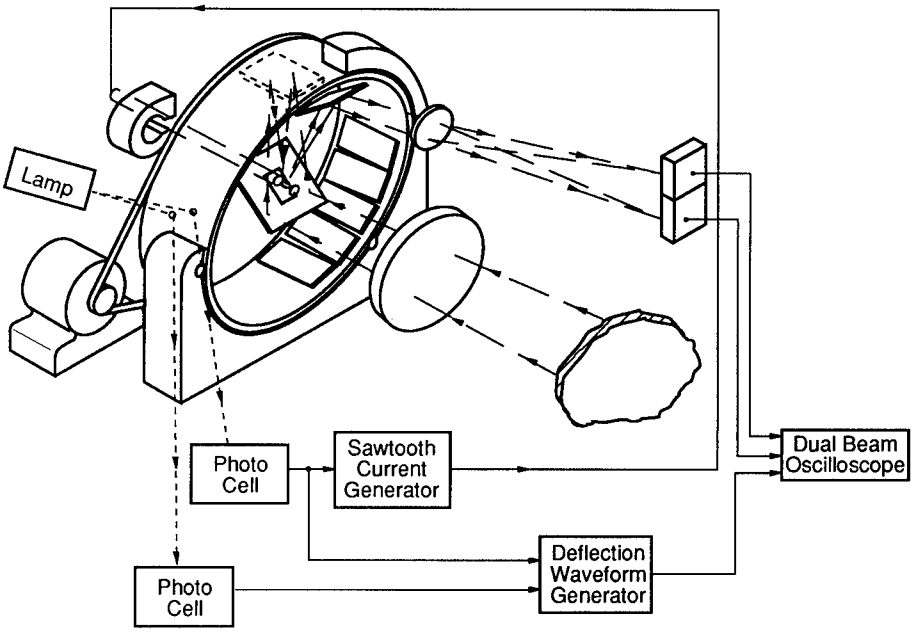


Fig. 3.20 Polygon-galvanometer scanner from Raytheon.

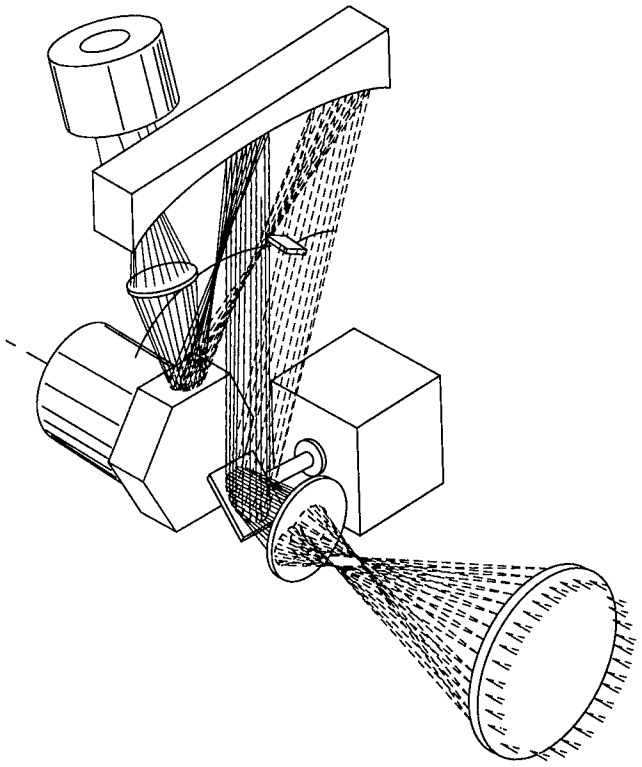


Fig. 3.21 Polygon-galvanometer scanner from Barr & Stoud Limited.

The Raytheon design includes an internal polygon with several mirrors tilted at successively greater angles; a fixed mirror, located on the axis of the polygon that receives the radiation and reflects it to each mirror as the polygon rotates; and a galvanometer-driven mirror that directs the scan to a remote detector. This design is notable for its compactness, which is due largely to the use of an internal polygon.

Disk-Galvo Systems. A FLIRS system combining a rotating multireflector disk for the horizontal scan and a galvanometer for the vertical scan is made by Kollmorgen Corporation (Fig. 3.22). The disk, containing 12 reflectors, is capable of 100% horizontal scan efficiency. This design was developed to minimize the size (dimensions are 10 × 5 in.) and weight (the unit weighs 2.5 lb). The rate of rotation is approximately half that required of a polygon scanner for the same scan rate. The manufacturing costs of the diamond-machined optical components are relatively low.

Resonant Scanner-Galvo System. Resonant scanners can be used instead of high-inertia scanners such as polygons in FLIRS scanning systems. The advantages of resonant scanners over polygons are reliability (no bearings), compactness, low power consumption, freedom from gyroscopic effects, low cost, and ruggedness over a broad set of environmental conditions.

A FLIR sensor using two low-inertia scanners is shown in Fig. 3.5. The horizontal trace is obtained by a resonant scanner generating 7,866 scan lines per second—by scanning 3,933 lines in each direction in each scan cycle—with a viewing angle of 28 deg. A dual-element detector obtains conventional TV raster density; a position sensor, closed-loop galvanometric scanner generates the frame scan. It is a compact, efficient design with low power consumption; the manufacturer is Inframetrics, Inc.

Galvo-Galvo Systems. Imaging systems using galvanometers for both scan dimensions are capable of highly versatile scan patterns. An example of a two-galvo FLIRS scanning system is the ultra linear scanner (ULS) developed by

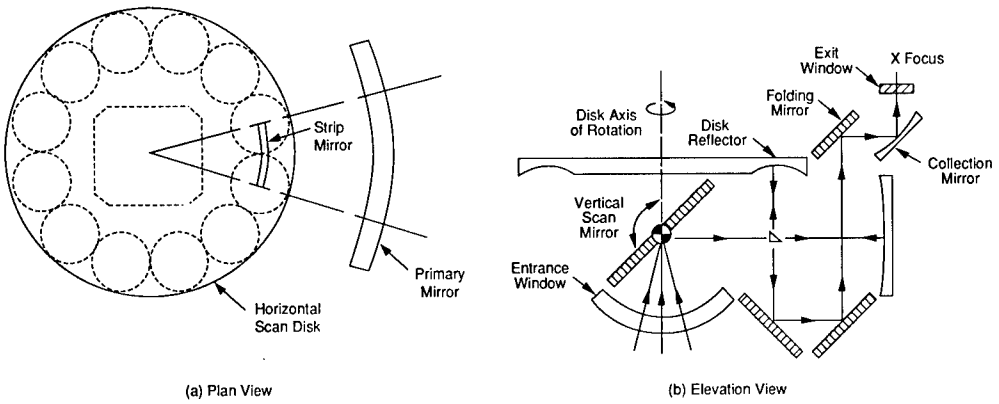


Fig. 3.22 Disk-galvo system. Beam-path folding with mirrors reduces the volume occupied by infrared imaging optics. Scene-scanning is accomplished by a small galvanometer scanner (the "vertical-scan mirror") and novel rotating-disk mechanism (the "horizontal-scan disk"). (Courtesy of Kollmorgen Corporation Electro-Optical Division)

Ford Aerospace Corporation. Figure 3.23 is a diagram of the ULS. According to the manufacturer, the system performance (scan) accuracies are 1/12 of an instantaneous field of view (IFOV) rms at the image plane while operating with a 15-deg IFOV and 80% scan efficiency; the mirror position linearity is 0.01% rms over the entire object FOV due to the use of an optical position sensor and three nested servocontrol loops.

Polygon-Polygon System. A "coaxial" double-polygon thermal imaging scanner has two polygons rotating along the orthogonal axis at slightly different speeds. In the example shown in Fig. 3.24, the lower polygon has seven facets and the upper has eight; the facets are inclined at different angles to the axis of rotation. The scanner has a pupil size of 10.4 mm and scans a field of view of 40 × 25.4 deg.

Cam Drive Scanner. A helicopter-based target acquisition infrared scanner is shown in Fig. 3.25. It uses a cam-driven oscillating mirror to scan 15 deg horizontally. The 3.4-deg vertical scan is achieved by a 48-element linear detector array. The active scan time is 255 ms, repeated at 340-ms periods.

3.5.3 Multiple-Axis Scanning Configuration

Two-axis scanners have a single mirror driven along two orthogonal axes. This approach has limitations and most multiple-axis systems are built with two independent scanners. They exhibit image distortions similar to those of relay lens scanners for large *f*/# systems.

Two-axis scanners face the age-old problem of mapping a sphere on a flat plane without distortions. An analysis of these distortions can be found in Marshall.[10] I address only uncommon configurations here. The following assumes that the work plane is flat.

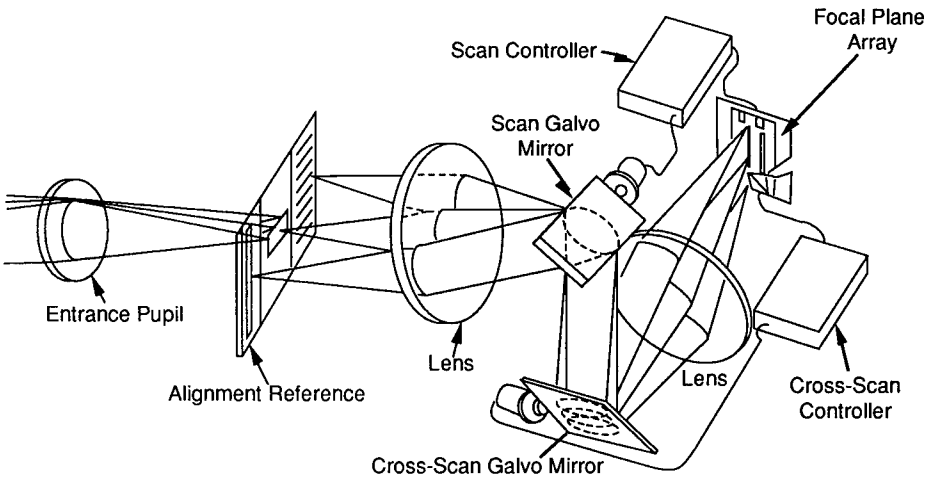


Fig. 3.23 Ultra linear scanner from Ford Aerospace.

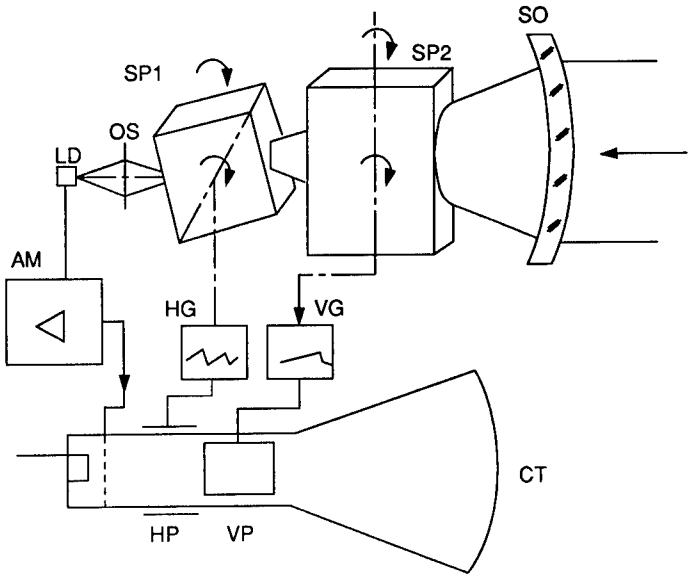


Fig. 3.24 Polygon-polygon system—AGA Thermovision.

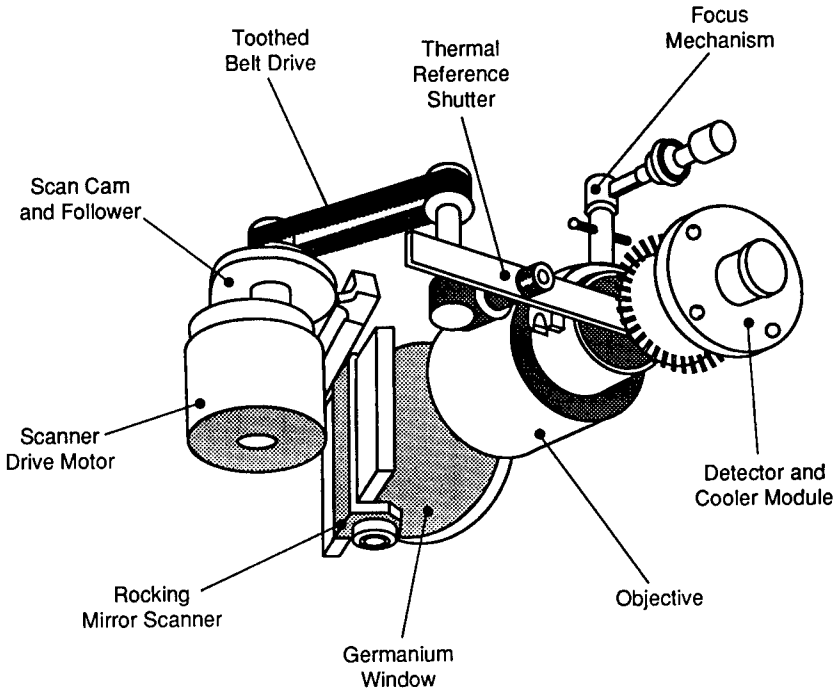


Fig. 3.25 Cam drive scanner.

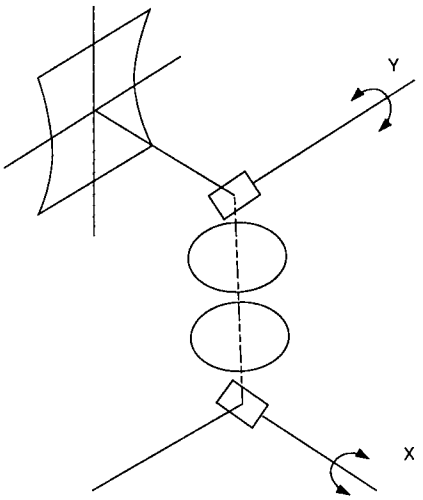


Fig. 3.26 Relay lens scanner.

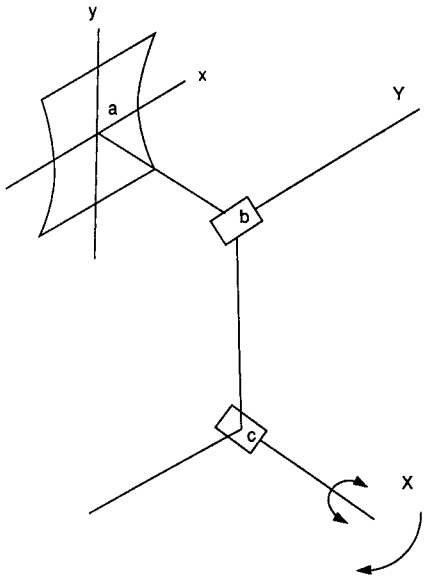


Fig. 3.27 Classic two-axis configuration.

Relay Lens Scanner. The relay lens configuration shown in Fig. 3.26 images the mirror of the X scanner onto the mirror of the Y scanner. This has the advantage that both axes can have identical dynamics. It also yields an unsymmetrical pincushion distortion of the image that can be corrected by a $F\theta$ lens and a look-up table to correct the scanner's position. This "perfect" system is to be considered only cautiously because perfect lenses could be extremely expensive.

Classic Two-Axis Configuration. The classic configuration of Fig. 3.27 causes the Y mirror to have a much higher inertia than the X mirror. This can be minimized if the X axis is tilted but remains within the *abc* plane. That plane has to be normal to the axis of rotation of the Y scanner. This also reduces the *bc* distance and its associated distortions. One should also note that the Y mirror need not be symmetrical, only mass balanced.

The distortions cannot be corrected by a spherical lens as we can see in Fig. 3.27. If both scanners are galvanometers, the distortions can be mapped in a look-up table and compensated electronically. If one scanner is a resonant device and a field flattening lens is used, it has to be the Y scanner. The excursion of a resonant scanner can be modulated in amplitude. The galvanometric scanner normally does not have the bandwidth to keep up with the resonant device. Unfortunately, this leads to a larger mirror for the fast resonant unit and a field flattening lens with a large pupil.

Paddle Scanner Arrangement. If a system calls for one resonant scanner, it should be in the Y position and the configuration of Fig. 3.28 is superior to the preceding one for two reasons:

1. The beam, to the first approximation, always hits the Y mirror in the same spot. Therefore, the mirror can be small, and the lens can have a small pupil.

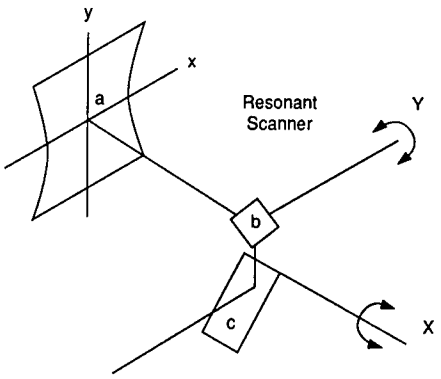


Fig. 3.28 Paddle scanner.

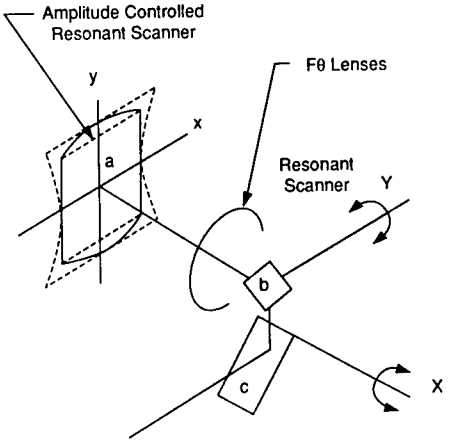


Fig. 3.29 Paddle scanner with error correction.

2. The field flattening lens converts the pincushion distortion of Fig. 3.28 into the barreling distortion of Fig. 3.29, which can be compensated as explained above.

3.6 SCANNER PERFORMANCE

A general-purpose specification of scanners is not practical. As the preceding sections have shown, an endless list of designs has been conceived to satisfy an even larger number of applications with a variety of technologies. The data in Tables 3.3 through 3.10 show the performance of some typical commercial polygon and oscillating scanners. These are the most common technologies and give a good indication of the state of the art.

The terminology of Sec. 3.7.1 indicates the design parameters that need to be quantified in order to specify a scanner. This list pertains to all types of scanners; however, each class has an additional number of its own class-specific parameters.

Table 3.3 Summary of Scanner Subsystem Requirements (from Ref. 1)

Number of facets	12
Facet angle	30 deg
Inscribed diameter	4.0 in.
Facet height	0.5 in.
Facet reflectance	89–95%
Facet flatness	λ/20
Facet quality	Scratch and digs (MIL-F-48616)
Scan rate	1200 scans/s
Rotational speed	6000 rpm

3.6.1 Rotating Scanners

Some features specific to polygons that need definition are:

Facets:

- facet length, polar location, and accuracy
- facet-to-facet angular tolerance
- facet cumulative angular tolerance
- facet-to-facet pyramidal tolerance
- facet cumulative pyramidal tolerance
- facet flatness
- facet surface quality
- facet coating uniformity
- facet durability
- facet scatter losses
- facet roll-off width.

Motor:

- speed stability
- power consumption
- bearing type, quality, and reliability
- delay time to come to speed
- dynamic tracking accuracy
- outgassing properties
- lubricant vapor pressure.

Tables 3.3, 3.4, and 3.5 describe high-performance polygon scanning subsystems. They are representative of precision scanners. They were prepared by Rynkowski[1] of Speedring Systems. Table 3.6 records the performance of a similar commerical polygon offered by Copal,[11] and Table 3.7 shows commercial polygon scanners from Lincoln Lasers, Inc.[12]

Table 3.4 High-Performance Polygonal Scanner Summary (from Ref. 1)

Characteristics	Reference System	State-of-Art System
Facet number	12	20
Facet tolerance	± 10 arcsec	± 1 arcsec
Apex angle error	± 0.4 arcsec	± 0.2 arcsec
Speed	6000 rpm	28,800 rpm
Scan rate	1200 scans/s	9600 scans/s
Regulation/revolution	< 10 ppm	< 1 ppm
Pixels/scan	10,000	50,000
Pixel/jitter/revolution	< ± 25 ns	< ± 2 ns
Pixel clock	12 MHz	480 MHz

Table 3.5 Subsystem Characteristics and Tolerances (from Ref. 1)

Characteristic	Tolerance	Comments
Number of facets	NA	Determined by scan angle
Facet angle	± 10 arcsec	One pixel-pixel angle
Diameter	NA	Controlled by facet width dimension
Facet width	1.035 in. min	0.020-in. roll-off
Facet height	0.5 in. min	0.020-in. roll-off
Flatness	λ/20 max	Spot control
Reflectance	± 3%	Tolerance
Apex angle	1.00 arcsec	Total variation − 10% of line-line angle
Speed regulation:		
1 revolution	± 10 ppm	± 1.08 arcsec/line
long-term	± 50 ppm	± 5.40 arcsec/line

Note: Scan error for any 12 scans ≤ ± 12.96 arcsec.

Table 3.6 Commercial Polygon Scanner

Characteristics	Copal PD 60
Facet number	12
Facet tolerance	± 10 arcsec
Speed	6000 rpm
Scan rate	1200 scans/s
Regulation/revolution	< 10 ppm
Pixels/scan	10,000
Pixel/jitter/revolution	< ± 125 ns
Pixel clock	12 MHz

Table 3.7 Typical Motor/Polygon Performance Parameters (from Ref. 13)

Application	Speed (rpm)	Number of Facets	Tracking Accuracy (Optional)	Velocity Stability (%)	Motor Type	Bearing Type
Low-end laser printer	4,200	8	1.5 arcmin	0.05	Hysteresis	Ball
Medium-end laser printer	6,100	14	1 arcmin	0.02	Brushless dc	Ball
High-end laser printer	23,000	18	45 arcsec	0.01	Brushless dc	Air (hydrodynamic)
Bar-code reader	2,000	6	5 arcmin	2.0	Brushless dc	Ball
Graphic arts printer	3,600	12	20 arcsec	0.01	Hysteresis	Ball
Graphic arts printer	12,000	1	2 arcmin	0.005	Hysteresis	Air (hydrodynamic)
Product inspection machine	28,000	12	5 arcsec	0.005	Brushless dc	Air (hydrostatic)

Table 3.8 Step Response of Moving Iron Galvanometric Scanners with Closed-Loop Integrating Amplifier—Shaft or Mirror Movements

Scanner Type	GSI G120D		GSI G400			GSI M1			GSI M3B			GSI G325	
Mirror													
Face, mm	7 × 7		15 × 15			9 × 16			24 × 38			26 × 26	
Thickness, mm	1		1.5			2			3.2			3.2	
Inertia, with mount, g cm^2	0.016		0.40			0.05			6.5			3.3	
Rotor inertia, g cm^2	0.023		0.20			0.35			6.5			3.7	
First resonance, kHz	4.0		7.0			12.5			5.0			1.0	
Stator													
Height, mm	33		75			74			98			79	
Cross section, mm	23× 33		30 diam			33 diam			51 diam			32 × 45	
Weight, g	107		240			150			300			435	
Life, cycles	10^{10}		10^{10}			10^{10}			10^{10}			10^{10}	
Bearing noise, arcsec	10		<2			<2			<2			5	
Wobble/jitter, μrad	25/50		8/15			6/12			4/12			12/25	
Step response, ms for angular step	a	b	a	b	c	a	b	c	a	b	c	a	b
1 deg	0.50	0.5	1.5	—	2.5	0.8	1.0	—	1.8	—	2.0	3.4	2.8
5 deg	0.55	0.57	1.6	1.7	—	—	—	—	1.9	2.2	3.2	3.5	3.4
10 deg	0.63	0.75	1.6	1.9	3.3	—	—	—	2.8	3.3	4.0	3.6	3.7
15 deg	0.70	0.95	1.8	2.0	—	1.7	1.9	2.0	2.8	—	—	4.1	4.4
20 deg	0.80	1.1	2.0	2.5	4.3	—	—	—	3.6	4.3	7.0	4.6	4.8

Conditions: Closed-loop integration amplifier
Power supply: ± 18 V, 3 A

a: Settled accuracy or error: 1% of step angle
b: Settled accuracy or error: 0.1% of step angle
c: Settled accuracy or error: 35 μrad

Table 3.9 Comparative Performances of Moving-Iron, Magnet, and Coil Galvanometric Scanners with Butterfly Moving-Dielectric Capacitive Encoders (Output shaft of 0.250 in. diameter for all units)

Model[a]	G3B	M3B	6650
Armature, moving	Iron	Magnet	Coil
Inertia, g cm^2	6.0	6.5	6.3
Thermal resistance coil to base, °C/W	—	1.4	2.8
Power at 80°C temperature rise, W	—	57.2	28.6
Torque at 10% duty factor, g cm	2500[b]	4750	3930
Acceleration, no load, rad/s^2 10^6	0.40	0.72	0.61
First resonance, kHz, 11 g cm^2 load	5	5	2.6
Resistance, Ω	2	4.8	4.5
Time constant, $\mathcal{L}/R$ ms	3	0.23	0.25
Weight, g	670	300	950

[a]G3B and M3B from General Scanning Inc. The 6650 from Cambridge Technology Inc.
[b]Torque limited by saturation.

3.6.2 Oscillating Scanners

In addition to the list of Sec. 3.7.1, the following parameters should be considered:

- sensitivity to external vibration
- mirror dynamic rigidity
- mirror mounting rigidity and alignment
- air turbulence effect on the mirror
- accuracy of the mounting surface
- coupling of the signal to the power ground
- sensitivity to environmental conditions
- outgassing properties
- heat dissipation and generation
- rf radiation
- perturbation to the environment.

Table 3.8 is an example of the performances obtainable with galvanometer scanners of various sizes and appropriate inertia. Table 3.9 lists the parameters of four similar commercial galvanometric scanners and compares the properties of the three torque transducers. The moving magnet and the moving iron devices have thermally controlled position transducers that yield low thermal drift. The moving-coil design does not offer a similarly elegant solution to drift control.

Table 3.10 compares the performances of nine commercial resonant scanners from three different vendors. Only those units with a balanced armature have very good vibration isolation and are free of vibration pollution.

Table 3.2 in Sec. 3.4.6 shows the tuning range and other parameters of a tunable resonant scanner.

Table 3.10 Typical Performance Characteristics of Resonant Scanners

Scanner Suspension	Taut Band	S-Flexure	Blade	Torsion Bar	IDS	Double Node	Torsion Bar	X-Flexure	S-Flexure
Performance range: resonant frequency/mirror diameter/angle, Hz/mm/deg	10/26/30 to 800/5/2	10/25/15 to 350/10/10	20/25/15 to 250/20/10	750/10/7 to 16,000/2/1.5	600/10/30 to 4000/7/10	4000/15/15 to 18,000/3/4	500/25/30 to 700/10/20	170/37/24 to 200/25/28	75/25/25 to 250/25/10
Model and number	SC-20	URS 100	IB20-100	SC-30	IDS 103010	IFS	CRS	IPX 4260	FLS 30
Typical performance									
Resonant frequency, Hz	100	100	100	1000	1000	4000	650	200	200
Temperature stability, ppm/°C	1000	200	300	500	200	200	−170	170	200
Mechanical scan angle, deg	15	15	15	6	30	15	21	28	15
Temperature stability, ppm/°C	NS[a]	400	NS	NS	100	NS	NA	10	1000
Mirror									
dimension, mm	15 × 16	20 × 20	20 × 20	10 × 10	10, diam	15, diam	13, diam	25, diam	25, diam
flatness, λ/cm	½	½	½	NS	½	½	¼	¼	¼
Wobble, μrad	NS	10	10	NS	35	NS	<2	NS	NS
Repeatability, μrad	NS	10	10	NS	5	5	10	<1	<5
Power, mW	40	15	18	300	100	100	2000	1000	1000
Operating temperature, °C	−20–40	0–60	0–60	−40–40	0–77	NS	10–32	15–60	0–60
Weight, g	21	20	18	56	100	50	260	320	150
Vibration sensitivity	Poor	Good	Good	Good	Excellent	Very good	Excellent	Excellent	Good

[a]Not specified.

Source: SC-20 and SC-30 data from Electro-Products Corporation data sheets; IB, IDS, IFS, CRS, and IPX data from General Scanning, Inc. data sheets; URS and FLS data from Laser Scanning product data sheets.

3.7 DEFINITIONS

With the raised expectations regarding the performance of scanning systems, the complexity of the technology of the components used has become visible. The following extended list of definitions and test methods has proven to be extremely useful. It was originally compiled as an effort to establish a common language to facilitate communications within the industry. In practice, it has been most valuable as a reminder to designers of all the factors they should address and how to measure the effectiveness of their solution.

3.7.1 Terminology

Accuracy: The maximum expected difference between the actual and commanded position. This includes any nonlinearities, hysteresis, noise, encountered drifts, resolution, and other factors.

Back EMF: The voltage produced by an inductor in opposition to a change in current or magnetic field. It is interesting to note that the back EMF produced in a galvanometer has the same units as the torque constant:

$$\frac{e}{d\theta/dt} = \frac{T}{i} = \text{BLND} \ . \tag{3.11}$$

Bandwidth: The maximum frequency at which a system can track a sinusoidal input with an output attenuated to no less than 0.707 (−3 dB point) of the command. For open-loop frequency responses with a phase margin of 90 deg, the open-loop crossover frequency will equal the closed-loop bandwidth. For other phase margins the relationship is not as straightforward. A complete treatment of these relationships can be found in any control theory text.

Current, demagnetization: The level that peak drive currents may not exceed without possible permanent damage to the scanner.

Damping, mechanical: Loss mechanisms that result in converting mechanical motion into heat. Typical mechanisms include bearing friction, aerodynamic losses (primarily in raster systems), mechanical hysteresis (due to material stress/strain behavior), and magnetic hysteresis.

Damping coefficient, mechanical: Used in linear system analysis, this parameter describes mechanical losses. For linear systems analysis, it is a constant, and as such can be used accurately only for certain kinds of loss mechanisms. At the microradian and millisecond levels, typical galvanometer losses (bearing drag, mechanical hysteresis, aerodynamic losses) do not fit the linear model and different mechanisms are dominant in different frequency ranges. Thus, for galvanometers the damping coefficient is actually a complex function of frequency and amplitude and understandably not specified.

Drift, mechanical null: The drift of the steady-state position of a nonpowered scanner when the armature is restrained with a torsion bar. This drift can occur with time and with temperature. Drift is usually specified in terms of the change in optical angle per amount of correlated influence, such as time or temperature.

Drift, position detector: The change in relationship between the output of the position detector and the position of the output shaft. It consists of the sum of the gain drift, null drift, and others.

Drift, position detector gain: The change in scale of the position detector. Since the absolute magnitude of this change is dependent on angle, it is specified in terms of the ratio of the change in output over the output per unit time or temperature (i.e., ppm/°C or %/1000 h). This takes into account that the effect is seen most at extreme angles where the output is greatest.

Drift, position detector null: The drift of the electrical zero of the position detector with time and temperature. Drift that occurs with temperature change is specified in units of angle per degree temperature (i.e., μrad/°C). Drift that occurs with time is specified in units of angle per units of time (i.e., μrad/1000 h).

Drift, uncorrelated: Drift that cannot be contributed to a change in a particular external condition, such as time or temperature. Often caused by mechanical ratcheting or noncatastrophic damage to the system due to overstress.

Hysteresis, magnetic: The property of a magnetic material exhibited by the lack of correspondence between a change in induction resulting from increasing and decreasing magnetomotive force. The result of this is often seen as a difference in settling times and positions when a given command angle is approached from different directions.

Jitter: Nonrepeatable position error fluctuations caused by velocity perturbation in a scanner (Fig. 3.30). Jitter is generally described in units of optical scan angle and often expressed as the standard deviation of the maximum jitter error observed in each scan line of a large number of consecutive scans. Some applications may require specification of the frequency as well as the magnitude of acceptable jitter.

Nonlinearity, best fit straight line: This method of quantifying nonlinearity involves finding a first-order linear function that is the closest approximation

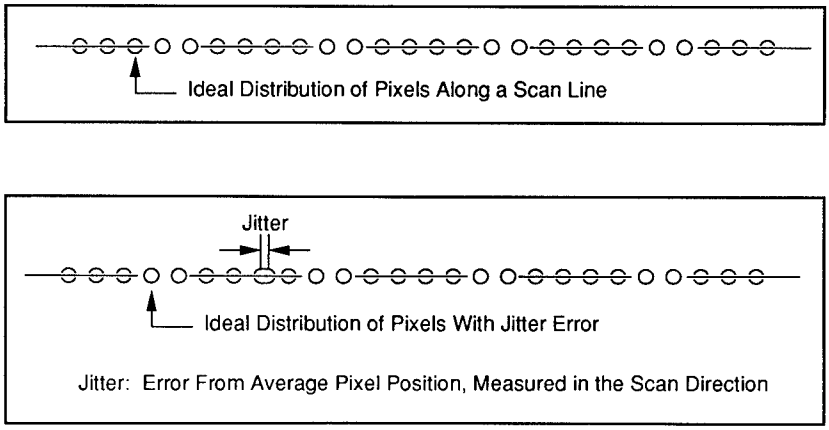


Fig. 3.30 Jitter.

to the measured data. The nonlinearity is then calculated as the maximum observed deviation from this line. This will result in the smallest measurement of nonlinearity.

Nonlinearity, pinned-center: Pinned-center nonlinearity uses a straight line that intersects a given datum point, such as the mechanical or electrical null of a scanner, and has a slope that best approximates that of the measured data. Sensor nonlinearity is then calculated from this reference.

Nonlinearity, pinned-endpoint: Pinned-endpoint nonlinearity uses a line drawn through the minimum and maximum measured values as the reference transfer function. The maximum nonlinearity is then measured from this reference.

Nonlinearity, torque constant: The inconsistency of the torque constant as a function of angle or current. Ideally, the torque constant would be constant with angle and a linear function of current. Often this parameter is specified as the maximum percentage error that may be seen over angle and up to a defined maximum current.

Nonlinearity, velocity: The inconsistency of velocity over the useful scan angle or other defined portion of the scan.

Null, electrical: The zero output point of the position transducer.

Null, mechanical: The steady-state position of a nonpowered scanner. This position is determined by the torsion spring, if any, and the magnetic spring of the scanner. In many scanners without a torsion spring, the magnetic spring is not of great enough magnitude to overcome frictional forces and make this an absolute position.

Overshoot: The amount of overscan that occurs prior to a scanner settling at a new location. Subject to the same considerations as settling time.

Repeatability: The inaccuracies in final position encountered while implementing a series of identical command inputs.

Repeatability, bidirectional: The inaccuracies in final position encountered while attempting to return to a position from different directions.

Resolution: The ability to discern individual spots in the target field. This is not to be confused with accuracy, which includes gain and offset drift, noise, resolution, and other factors. Dependent on system design, the limit to resolution may be due to optical considerations, digital resolution, or position detector signal-to-noise ratio and drift.

Resolution, optical: The optical resolution of a scanning system can be described as the number of separately resolvable spots that can be produced. For diffraction-limited optics, this is dependent on the aperture width in the scan direction, the aperture shape factor, the wavelength of the source, and the total scan angle. These factors are related through the scan equation, which can be found in many optical texts.

Resolution, scanner: Scanner resolution is limited by the noise and drift of the position detector. The rms signal-to-noise ratio will determine the statistical resolvability of a given level of command in a given frequency range. Filtering

can improve low-frequency resolution, but the drift factor will also come into effect.

Resonance, cross-axis: Structural resonances that cause motion perpendicular to the scan axis are referred to as cross-axis resonances. These resonances may be accentuated by poor mirror design and will cause periodic wobble, possible system instabilities, and may be a limit to achievable system bandwidth.

Resonance, torsional: An on-axis resonance that appears in scanners due to the distributed masses on a compliant rotor shaft or the flexible coil of a d'Arsonval system. These resonances can appear as periodic jitter and may cause controllability difficulties due to the resonant peak created in the scanner's transfer function. Mirror design and mounting will have a significant effect on torsional resonance.

Response time: The response time of a scanning system is defined as the tracking error divided by the slew rate. Due to characteristics of the controller, stage saturation, aerodynamics, and other nonlinearities, response time will not necessarily be a constant. Although not a constant, at least for vector-tuned controllers, it is approximately so if slew rates are neither driving the tracking error to near zero nor approaching the maximum slew rate. Response time is an indication of the relative speed of a scanner and the ultimate performance obtainable with a given load and tuning.

Scan angle, maximum: The maximum angle a scanner can achieve. Usually specified in optical angle.

Scan angle, mechanical: The angular excursion of a scanner's shaft. One-half the optical scan angle.

Scan angle, optical: The angular excursion of a scanned optical beam. Twice the mechanical scan angle for a reflective scanner.

Scan angle, peak: The maximum angle a scanner, or scanned beam, travels with a given raster or vector command. Usually specified in optical angle.

Scan angle, useful: The portion of scan over which useful work is done. Usually tied to another criteria (i.e., velocity linearity) that must be maintained during this period. Usually specified in optical angle.

Scan efficiency, angular: The ratio of useful scan angle to peak scan angle in raster scanned systems. Typically expressed as a percentage.

Scan efficiency, interval: The ratio of useful scan angle time interval to scan period in raster scanned systems. Typically expressed as a percentage.

Settling time: Settling time quantifies the ability of a scanning system to move to a given location and point to within a certain error band. This error band may be either in percentage of full field or step, typically 1 part in 1000 or 10,000, or a specified angle. The angle that must be traversed before settling at the new location has a major effect on settling time and must also be indicated when settling time is stated. Often this angle is the full-field angle. Settling time is sometimes specified with the error band referenced as a fraction of an executed step. Although physics dictates that settling to a given error

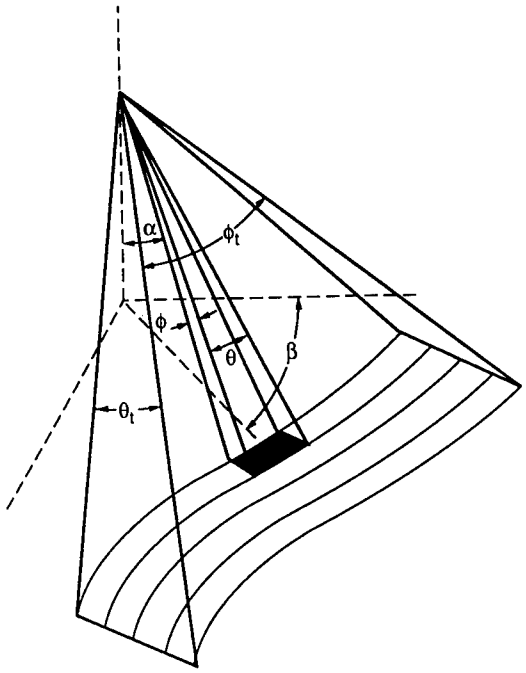


Fig. 3.31 A general scan pattern.

will occur quicker with smaller steps, care must be taken when specifying settling to a fraction of this step since, for a small step, the settling error band may be below the resolution of the system.

Signal-to-noise ratio, position detector: The signal-to-noise ratio (SNR) indicates the relative magnitudes of full-scale sensor output and sensor noise. This can be an rms, peak, or peak-to-peak measurement. The rms measurements have the most statistically complete data and are often considered the most convenient for calculations based on SNR. This parameter is also generally specified in decibels. The sensor noise is a function of frequency and the frequency band that it is measured over must be included in any specification of this parameter. The SNR of a position detector is the ultimate limit to a scanner's resolution, independent of system considerations.

Signal-to-noise ratio, scanning parameters[c]: In a general scanning system, the instantaneous field of view has angular dimensions θ and ϕ and points in the direction (α, β) (see Fig. 3.31). The total field angle is ϕ_t in one direction and θ_t in the other direction. The perpendicular distance from the scanner to the object at the nadir point is usually designated h. The SNR of a scanning infrared system can be written in the simplified form:

[c]This definition is from W. L. Wolfe, "Optical-Mechanical Scanning Techniques and Devices," Chap. 10 in *The Infrared Handbook,* W. L. Wolfe, G. J. Zissis, Eds., Environmental Research Institute of Michigan, Ann Arbor (Revised 1985).

$$\text{SNR} = D^* \phi_d A_d^{-1/2} b^{-1/2} ,$$

where

D^* = specific detectivity
ϕ_d = power or flux on the detector
A_d = effective area of the detector
b = noise bandwidth.

The noise bandwidth is often directly related to the information bandwidth and inversely proportional to the dwell time on a single resolution element. The electrical information bandwidth of an infrared system should be proportional to the reciprocal of the shortest dwell time on any resolution element in the total field of view. If scanning velocity is constant over the total field and only the field is converged (i.e., there is no overlap and no dead time in retrace or at the edges), then the dwell time t_d for a frame is

$$t_d = \frac{T_f}{n} ,$$

where n is the number of resolution elements and T_f is the frame time.

Slew rate: The angular velocity of a scanner. Generally specified in optical angle per unit time.

Slew rate, maximum: The maximum achievable slew rate for a given scanner, load, controller, and tuning combination. Typically measured during the response to a step input.

Spring, magnetic: The torque, per unit angle, applied to a scanner's rotor due to nonuniform magnetic fields or changes in field due to excursion of the rotor.

Spring, mechanical: The centering torque, per unit angle, created by mechanical attachment to the rotor of a scanner. This is often a torsion bar, but may be a wire attachment for moving-coil designs or grounding purposes.

Time constant, $\mathcal{L}/R$: The ratio of inductance to resistance, expressed in seconds. In response to a step change in voltage, the time required for the current through an inductor-resistor network to achieve 63% of its final value. This can be a performance limiting factor when driving galvanometers with voltage sources. When a current source output stage is used, this will not affect performance unless the limits of the voltage headroom are reached. This should not be confused with the back EMF constraints.

Torque constant: The measurement of torque output of a galvanometer, per unit current, is known as the torque constant. Torque constant is most often specified in Newton·meters or dyne·centimeters per ampere. The torque constant is important in calculating the relationships between current, load inertia, and scanner acceleration. Care must be taken in using this parameter since it is not constant over angle, nor is it linear with current, although it is approximately so. Careful controller design and analysis may be required to achieve the desired performance over the range of gains that will be encoun-

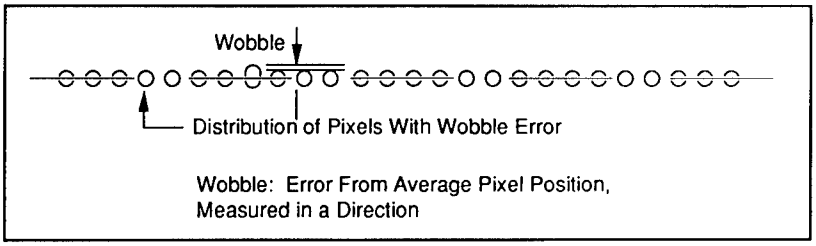


Fig. 3.32 Wobble.

tered due to these nonlinearities. Most often this will have the greatest effect on settling time performance, bandwidth, and stability.

Torque-to-inertia ratio: This key figure of merit defines the maximum achievable angular acceleration for a given scanner and load combination. The relationship between angular acceleration and torque is given by

$$a = T/J \ .$$

Tracking error: The difference in angle between actual and commanded position. Must be specified in conjunction with a slew rate and is typically specified in terms of the optical angle. Control system tuning will have a considerable effect on the value of this parameter.

Wobble: Cross-axis motion of a scanner during scan. Generally specified in terms of an optical angle representing the standard deviation of the maximum repeatable or nonrepeatable wobble error measured in each scan line of a large number of consecutive scans.

Wobble, nonrepeatable: Random or nonscan-synchronous periodic cross-axis motion of a scanner during a scanning motion (Fig. 3.32). This will appear as perturbation perpendicular to the scan motion in a line scan or as jitter in the perpendicular axis of an x-y system.

Wobble, repeatable: Consistent cross-axis motion of a scanner during scan. Often caused by bearing run out, this motion can appear as bow in a scanned line or as a nonlinearity in an x-y system.

3.7.2 Discussion and Test Methods

Scan angle: An important note is the difference between optical and mechanical rotation. A given mirror rotation results in twice the rotation of the optical path because of the change in both the incident and the reflected beam angles. Scanner specifications are usually given in terms of optical angles for ease of system design, but to avoid confusion, it is always wise to note which is being used.

Resolution: One should not confuse scanner resolution and the resolution of an optical instrument. Resolution in an optical system can be described as the ability to discern individual spots in the target field. This is not to be confused

with accuracy, which includes gain and offset drift, noise, resolution, and other factors. Depending on system design, there are three possible limits to achievable resolution: (1) Optical resolution is a function of wavelength, aperture, angle, and focal length, as described in the scan equation. (2) Digital system resolution is limited by the number of addressable bits, usually in the digital-to-analog converter. (3) There are limits to resolution particular to the scanner, the position detector, and SNR and drift.

Velocity linearity: Velocity linearity refers to the consistency of velocity over a portion of the scan. For raster scan applications the portion of the scan is the useful scan angle defined in the previous section. Velocity linearity is also important for vector applications in which work is being done while the beam is in motion (as opposed to point and shoot applications). Velocity linearity can be checked with an array of split cell photodetectors and a counter able to measure the time a scanned beam takes to traverse the detectors. An easier way is to differentiate a position signal to obtain velocity. If care is taken to calibrate this setup for scale and flatness, accurate results can be achieved. Velocity linearity is very dependent on load, drive electronics, and command signal. Therefore, it must be specified in conjunction with these considerations.

Jitter: In addition to the repeatable nonlinearity of scanner velocity, nonrepeatable disturbances to velocity create irregular position perturbations referred to as jitter. This is most easily pictured if we think about a raster application laying down evenly spaced pixels. Deviation from an expected location of a pixel is jitter (see Fig. 3.30). Jitter is described in units of scan angle and is often expressed as the standard deviation of the maximum jitter error observed in each scan line of a large number of consecutive scans. The frequency as well as the magnitude of allowable jitter varies with application, as does measurement technique.

Several methods are available for measuring jitter. The most straightforward is to lay down a series of pixels on film and measure the distance between each. Straightforward calculations then produce a jitter measurement. Care must be taken not to allow lens aberrations and flat field tangential errors to corrupt the measurement. Due to these problems, and the large amount of data that must be collected, other methods have been developed to measure jitter.

Using the scanner to deflect a laser across a split cell photodetector and observing the differential power from each side of the detector—noting when the power on both cells is equal—gives a very accurate trigger. If a precision timer is then used to measure the travel time between detectors, a good measurement of scan-to-scan velocity jitter can be obtained. If this measurement is made over a region of the scan where the velocity is relatively constant, an accurate estimate of position jitter can be inferred. This setup also lends itself well to computer data collection and analysis. By knowing the average travel time and the angle between the detectors, the average velocity can be calculated. The scan-to-scan time difference, multiplied by this average velocity, gives an indication of the position jitter of the scanner.

There is, however, one major concern with this method of jitter measurement: measuring only one small portion of the scan with each split cell. Due to control loop stiffness, a velocity perturbation may cause the position error

of a scanner to increase. The control system strives to correct this. The result is that a velocity perturbation that occurred in the middle of the scan may have little or no effect on the transit time between detectors situated at the beginning and the end of a scan.

One method to minimize this problem is to use an array of detectors. Although the complexity of the data collection system increases with the number of detectors, this is one of the best ways to implement an auto-test setup. When testing a small number of scanners, a simpler test setup with only two photodetectors obtains data that are just as accurate. If the velocity of a scanner is observed by electronically differentiating the position signal, any areas of the scan particularly susceptible to jitter can be observed. If one detector is positioned at a relatively stable portion of the scan and the other is placed at the most unstable portion of the scan, the worst case jitter can be measured.

To arrive at the true value of jitter,[13] a somewhat more rigorous approach involves making some statistical assumptions. With the two-sensor test, the sampled error is not necessarily the maximum encountered along the scan line, and the standard deviation obtained is therefore less than the real value. To calculate the true value of the standard deviation, the jitter error distribution is assumed to be Gaussian and the maximum peak-to-peak error observed over a large number of scans is assumed to be the same at all points of the scan. For the most part, both are very reasonable assumptions.

The maximum jitter error is half of the maximum peak-to-peak value. If the jitter distribution is Gaussian, the peak jitter observed is the following multiple of the standard deviation:

0.6745 SD occurs 50% of the time
1.0 SD occurs 32% of the time
2.0 SD occurs 4.5% of the time, or 1 scan in 22
3.0 SD occurs 0.27% of the time, or 1 scan in 370
3.5 SD occurs 0.0047% of the time, or 1 scan in 2150.

Thus, if the test is over 1000 scans, the peak jitter observed is about 3.4 times the true standard deviation.

This model estimates imprecision between two fixed points and contains the implicit assumption that the standard deviation of jitter is independent of scan angle. If the standard deviation varies with angle, determining jitter levels and imprecision at randomly selected angles gives a more accurate estimate of the standard deviation. The analysis of variance yields the within-scan standard deviation. These results determine tolerance intervals. When maximum jitter values are used, the Gaussian model is not correct. In this case, an extreme value distribution is the appropriate model.

Whichever method of measurement is used, a few basic techniques can be used to minimize system jitter. Care must be taken to minimize mechanical vibrations and beam wander that may be indistinguishable from jitter. Mirror imbalance and aerodynamic effects can be major causes of jitter. Scan amplitude and speed, as well as mirror weight, also affect observed jitter.

Each application defines the criteria of acceptable jitter. Good scan-to-scan jitter is important for applications such as line art, where displacement of a pixel is very noticeable. Comparing a particular scan time to the average time of a series of scans develops a feel for the distribution of jitter. Low-frequency

jitter can cause distortions that are seen over portions of a page rather than from one scan line to another.

Wobble: In addition to on-axis motion of scanners, undesired cross-axis motion is always present to some extent. This cross-axis motion is referred to as wobble (see Fig. 3.32). Most scanners exhibit wobble of two types, repeatable and nonrepeatable. Repeatable wobble is consistent deviation from a true straight line across the scan plane and often appears as a bow. Nonrepeatable wobble is a random variation in position perpendicular to the scan. Again different applications have varying tolerances of repeatable and nonrepeatable wobble.

Nonrepeatable wobble can be specified in terms of an optical angle representing the standard deviation of the maximum nonrepeatable wobble error measured in each scan line of a large number of consecutive scans. Repeatable wobble should be specified in terms of optical angle.

Wobble, especially repeatable wobble, is often a difficult parameter to measure. Cylindrical optics cancel out beam displacement in the scan direction, and a split cell photodetector used to detect variations perpendicular to the scan direction (see Fig. 3.33). The detector's output can then be observed with the aid of an oscilloscope. However, due to optical distortions the display does not show a horizontal line, even for a perfect scanner. As a result, unless care is taken to calibrate out the intrinsic errors of the test stand, this method accurately measures only nonrepeatable wobble.

Since nonrepeatable wobble has statistical characteristics similar to jitter, statistical analysis evaluates the maximum expected wobble from the standard deviation of a sampled measurement. This method eliminates the cylindrical optics and only a portion of the scan need be observed.

The excitation waveform used in a wobble test has a considerable effect on the measurement obtained. Duty cycle, slew rate, flyback time and control system tuning all have major effects on wobble, particularly repeatable wobble. These factors make a standardized wobble test both very difficult to define and quite necessary.

Structural resonance: Several sources of dynamic error in optical scanners involve the structural rigidity of the scanner and mirror assembly. A scanner's

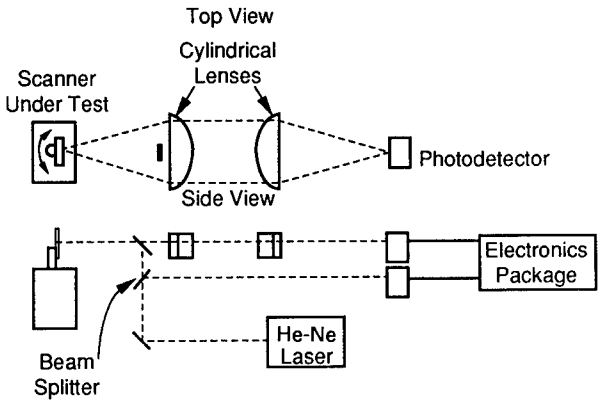


Fig. 3.33 Wobble measurement fixture.

shaft is subject to both on-axis and cross-axis resonances. They appear as jitter and wobble, respectively. The frequency and magnitude of these resonances depend on the design and composition of the shaft and the load. The design of a mirror and mounting clamp has major effects on the dynamic characteristics of a scanner. In addition to twisting and bending in the scanner's shaft, resonances also appear in a poorly designed or overstressed mirror assembly. The inherent oscillatory motion of galvanometers and resonant scanners causes dynamic loading and distortion in a mirror. Standard techniques[14] can be used to analyze this.

Structural resonances can be difficult to identify and characterize. Torsional resonances cause periodic jitter, but at a frequency almost impossible to measure due to the spatial resolution required. Torsional resonance effects are easily seen by generating a Bode plot of a galvanometer. It is readily apparent that pushing the bandwidth of a controller causes oscillations at the torsional resonant frequency. Cross-axis resonance is easily observed using a standard wobble test. Errors created by this cross-axis motion appear as periodic wobble.

Response parameters: The speed at which a galvanometer responds to commands is a concern for most any application. The rate at which a galvanometer scans across the field is referred to as the slew rate and is specified in units of optical angle per unit time (e.g., rad/s). An additional piece of information is required to characterize a scanner for an application. This parameter refers to the difference, in angle, between a scanner's actual and commanded position and is known as a tracking error. Tracking error and slew rate must be specified in conjunction with each other and both are dependent not only on scanner and load, but also on control system tuning. The parameters of tracking error and slew rate are interrelated through a third parameter, response time, which characterizes the delay between command and response. Response time is defined as the tracking error divided by the slew rate. It indicates the relative speed of a scanner and the ultimate performance obtainable with a given load and tuning.

Slew rate: Commanding a step input to the controller and observing the position output on an oscilloscope indicates the maximum achievable slew rate of the galvo-load-tuning combination (see Fig. 3.34). This is a figure of merit that may be used for system analysis and comparison; however, the maximum slew rate is often not the optimal for an application. Vector tuned systems tend to respond best to "structured inputs" that do not have the nonlinearities associated with step inputs. The galvo's slew rate is then the commanded slew rate, and the parameter of interest becomes tracking error.

Another practical limit to slew rate is the limitations of the galvanometer's bearings. At very high angular velocities bearings begin to skid rather than roll, causing degradation of performance and eventual scanner failure.

Tracking error: Tracking error is important for applications that require precise knowledge of a moving scanner's location at a particular moment in time. Often a scanner can be moved to a location, work done, and then the scanner moved to the next location of interest. When throughput or imaging quality requires continuous motion of a scanner during the work process, knowledge of the tracking error can be used to locate the galvanometer accurately relative

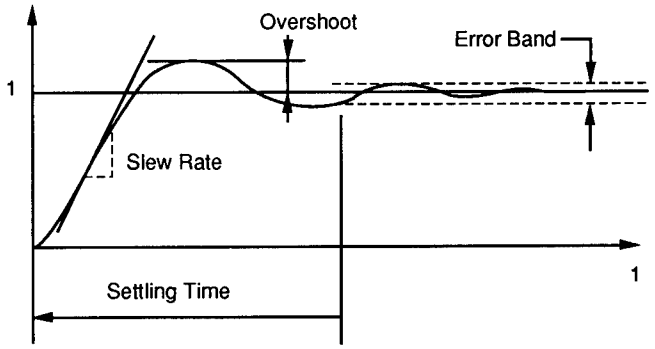


Fig. 3.34 Response parameters.

to the command. Certain types of photoplotting and laser marking use vector imaging. There, x and y galvanometers specify relationships of tracking error and response time to obtain straight diagonal lines.

Once the tracking error is determined for a given slew rate command, the response time is calculated. Due to characteristics of the galvanometer, controller, stage saturation, losses, and other nonlinearities, response time is not a constant.

Settling time: The ability of a galvanometer to settle depends on many factors. In addition to the galvanometer model, load, control system, and tuning, the "structure" of the move to a new location is important in the optimization of settling time. The general philosophy of structured steps is to eliminate nonlinearities in the command. Step inputs tend to saturate stages in the control system, causing changes in the tuning and degradation in performance. Structured moves are also used to minimize reaction torques and peak current draw, or to optimize other considerations. Settling time is measured on an oscilloscope by observing a calibrated position error signal. More accurate measurements are made with the aid of an autocollimator.

Step drift: In very high accuracy applications the subtlety of settling time must be considered. This involves the structural stability of the scanner. When a scanner changes field location, the amount of current supplied to the scanner changes. As the current level changes, so does the power dissipated. The heating and cooling of the scanner cause slight distortions of the scanner body. The distortions, in turn, change the mirror location relative to the mounting plane as well as gain and offset drift in the position detector. The result, a slow tail to the optical settling of the galvanometer, is known as step drift. Because the controller maintains a constant position detector output, this cannot be observed on the position error signal. In a well-designed scanner, observing this small error source requires a sensitive autocollimator.

References

1. G. A. Rynkowski, "High Performance Polygonal Scanners, Motors and Control Systems," Chap. 7 in *Optical Scanning,* G. F. Marshall, Ed., pp. 411–417, Marcel Dekker, New York (1991).
2. R. J. Sherman, Lincoln Laser, Inc., in *Optical Scanning,* Gerald F. Marshall, Ed., p. 358, Marcel Dekker, New York (1991).
3. D. Zook, "High beam deflector performance: a comparative analysis," *Applied Optics* **13**(4) (1974).
4. W. Egger, "Capacity pickup follow-up system," U.S. Patent No. 2,534,505 (1944).
5. J. Montagu, "Actuator with compensating flux path," U.S. Patent No. 4,528,533 (1985).
6. R. Abbe, "Apparatus and method for direct readout of capacitive gauge dimension," U.S. Patent No. 3,775,679 (1973).
7. M. D. Ergen and M. Petran, "New reflected-light microscope," *Science* **157,** 305–307 (1967).
8. M. Gottlieb in *Optical Scanning,* G. F. Marshall, Ed., pp. 650–652, Marcel Dekker, New York (1991).
9. L. Germann and J. Braccio, "Fine-steering mirror technology supports ten nanoradian systems," *Optical Engineering* **29**(11), 1351–1359 (November 1990).
10. G. Marshall, Ed., *Optical Scanning,* pp. 556–569, Marcel Dekker, New York (1991).
11. J. M. Lloyd, *Thermal Imaging Systems,* Plenum Press, New York (1975).
12. R. Sherman, Lincoln Laser Inc., private communication.
13. P. Brosens, "A guide to linear scanning," Section 5 in *Linear Scanning Set Users Manual,* PN12P-LSS, General Scanning Inc., Watertown, MA (August 1989).
14. P. Brosens, "Dynamic mirror distortion in optical scanning," *Applied Optics* **11**(12), 2987 (December 1972).

Bibliography

Berry, P. J. and H. M. Runciman, "Radiation scanning system," U.S. Patent No. 4,210,810.

Brosens, P. J., "Dynamic mirror distortion in optical scanning," *Applied Optics* **11**(12), 2987 (1972).

Buchroeder, R. A., C. R. Hayslelt, and W. H. Swantner, "Rotating prism compensators," *SMPTE Journal* **86** (June 1977).

Goodman, F. and A. Prantakis, "Light beam positional apparatus," U.S. Patent No. 4,685,775.

Gottlieb, M., C. Ireland, and J. M. Levy, *Electro-Optic and Acousto-Optic Scanning and Deflection,* Marcel Dekker, New York (1983).

Hibson, H. and J. M. Botermier, "Breadboarding of a high bandwidth acquisition and fine tracking system for satellite optical communications," *OPTO* (1989).

Marshall, G., Ed., *Optical Scanning,* Marcel Dekker, New York (1991).

Marshall, G., Ed., *Laser Beam Scanning,* Marcel Dekker, New York (1985).

Marshall, G. and L. Beiser, Eds., *Beam Deflection and Scanning Technologies, Proceedings of the SPIE* **1454** (February 1991).

Pawley, J. B., Ed., *Handbook of Biological Confocal Microscopy,* Plenum Press, New York (1990).

CHAPTER 4

Detectors

Devon G. Crowe
Georgia Institute of Technology
Atlanta, Georgia

Paul R. Norton
Santa Barbara Research Center
Santa Barbara, California

Thomas Limperis
Consultant

Joseph Mudar
Nichols Research Corporation
Vienna, Virginia

CONTENTS

4.1 INTRODUCTION

The purpose of this chapter is to provide the critical data, formulas, and written text for evaluation of infrared detectors for specific applications. The chapter is limited to single-element detectors and detector arrays that are sensitive to the 0.7- to 1000-μm spectral region. A detector may be defined as[1]:

> . . . a device that provides an electrical output that is a useful measure of the radiation incident on the device. It is intended to include not only the responsive element, but also the physical mounting of the responsive element, as well as any other elements—such as windows, area-limited apertures, Dewar flasks, internal reflectors, etc.—that form an integral part of the detector as it is received from the manufacturer.

Table 4.1 lists the symbols, nomenclature, and units used in this chapter.

4.1.1 Symbols and Descriptions for Detector Parameters

Tables 4.2 and 4.3 present the currently acceptable symbols and preferred units for the important detector parameters and noise equations. This nomenclature was assembled from the archival literature, government reports, and the standards report[1] prepared by Jones et al.

4.1.2 Responsive Elements

The responsive element is a radiation transducer. It changes the incoming radiation into electrical power that can be amplified by the accompanying electronics.

The methods of transduction can be separated into two groups: thermal detectors and photon detectors. The responsive element of thermal detectors is sensitive to changes in temperature brought about by changes in incident radiation. The responsive element of photon detectors is sensitive to changes in the number of free charge carriers, i.e., electrons and/or holes, that are brought about by changes in the number of incident infrared photons. Thermal detectors employ transduction processes including the bolometric, thermovoltaic, thermopneumatic, and pyroelectric effects. Photon detectors employ transduction processes including the photovoltaic, photoconductive, photoelectromagnetic, and photoemissive effects. Each process of transduction is described in this chapter.

4.1.3 Descriptions of the Processes of Transduction

Bolometric Process. Changes in the temperature of the responsive element, induced by the incident infrared radiation, cause changes in the electrical conductivity, monitored electrically.

Photoconductive Process. A change in the number of incident photons on a semiconductor causes a change in the average number of free charge carriers, or the mobility of the carriers, in the material. The electrical conductivity of the semiconductor is directly proportional to the average number of free charge carriers in the material. Therefore, the change in electrical conductivity is directly proportional to the change in the number of photons incident on the semiconductor.

Photoelectromagnetic Process. Photons absorbed at or near the surface of a semiconductor generate free charge carriers, which diffuse into the bulk and

Table 4.1 Symbols, Nomenclature, and Units

Symbols	Nomenclature	Units
A_c	Cross-sectional area	cm^2
A_d	Area of the detector	cm^2
A_e	Effective area of detector	cm^2
a	Absorption coefficient	cm^{-1}
B	Magnetic field	G
b	Ratio of electron to hole mobility	—
C	Electrical capacitance	F
$\mathscr{C}$	Heat capacitance	$J\ K^{-1}$
C_e	Equivalent capacitance	F
c_1	First radiation constant	$W\ cm^{-2}\ \mu m^4$
c_2	Second radiation constant	$\mu m\ K$
c	Velocity of light	$m\ s^{-1}$
D	Detectivity	W^{-1}
D^*	Detectivity, specific (normalized with regard to detector area and electrical bandwidth)	$cm\ Hz^{1/2}\ W^{-1}$ (Jones)
$D^*(\lambda)$	Spectral, specific (normalized) detectivity	$cm\ Hz^{1/2}\ W^{-1}$ (Jones)
D^{**}	Detectivity normalized with regard to detector area, electrical bandwidth, and effective, weighted angular field of view	
DQE	Detective quantum efficiency	—
D_e	Electron diffusion constant	$cm^2\ s^{-1}$
D_h	Hole diffusion constant	$cm^2\ s^{-1}$
d	Thickness of responsive element	cm
E	Irradiance	$W\ cm^{-2}$
$\mathscr{E}$	Electric field	$V\ cm^{-1}$
E_g	Photon energy	J
E_i	Impurity activation energy of photoconductor	J
E_k	Kinetic energy of the freed electron	J
E_q	Photon flux density, or photon irradiance	photons $cm^{-2}\ s^{-1}$
$E_{q,B}$	Photon flux density for background radiation	photons $cm^{-2}\ s^{-1}$
$E_{q,s}$	Photon flux density for signal radiation	photons $cm^{-2}\ s^{-1}$
e	Charge on an electron	C
FOV	Detector geometric field of view	sr or cone-angle degrees
f	Electrical frequency	Hz
f_c	Chopping frequency	Hz
f_o	Modulation frequency	Hz
$\mathscr{G}$	Thermal conductance	$W\ K^{-1}$
$\mathscr{G}_e$	Effective thermal conductance	$W\ K^{-1}$
$\mathscr{G}_0$	Combined effective and radiative thermal conductance	$W\ K^{-1}$
G_{gen}	Generation rate of free charge carriers	s^{-1}
G_p	Photoconductive gain	—

Table 4.1 *(continued)*

Symbols	Nomenclature	Units
G_{sh}	Effective shunt conductance	mho
g	Gain	—
h	Planck's constant	eV s
I	dc current	A
I_d	dc diffusion current	A
I_s	dc signal current	A
I_{sa}	dc saturation current	A
I_{sc}	dc short-circuit current	A
i,i_s,i_n	ac rms generalized, rms signal, or rms noise current	A
$\mathcal{I}(t),\mathcal{I}_s(t),\mathcal{I}_n(t)$	Instantaneous generalized, signal, or noise current	A
K	Constant	—
k	Boltzmann's constant	$J\ K^{-1}$
L	Diffusion length	cm
L_d	Effective diffusion length	cm
L_e	Electron diffusionlength	cm
L_h	Hole diffusion length	cm
l	Length; also electrode separation	cm
M	Free charge carrier multiplication factor	—
N	Total number of free charges	—
NEE	Noise equivalent irradiance	$W\ cm^{-2}$
NEP	Noise equivalent power	W
N_s	Signal photon rate	s^{-1}
N_λ	Average photon rate per unit wavelength and per unit area	$s^{-1}\ \mu m^{-1}\ cm^{-2}$
n	Density of free electrons	cm^{-3}
$\mathcal{P}$	Pyroelectric coefficient	$A\ cm^{-2}$
P_{ab}	Thermoelectric power	$V\ K^{-1}$
p	Density of free holes	cm^{-3}
R	Resistance	Ω
$\mathcal{R}$	Responsivity	$V\ W^{-1}$ or $A\ W^{-1}$
$\mathcal{R}_{BB}$	Blackbody responsivity	$V\ W^{-1}$ or $A\ W^{-1}$
$\mathcal{R}_{ref}(\lambda)$	Relative spectral responsivity of the reference	$V\ W^{-1}$ or $A\ W^{-1}$
R_d	Detector resistance	Ω
R_{dyn}	Dynamic resistance	Ω
R_e	Equivalent input resistance of the detector-preamplifier circuit	Ω
R_L	Load resistor	Ω
R_m	Resistance of the current meter	Ω
RQE	Responsive quantum efficiency	—
s	Surface recombination velocity	$m\ s^{-1}$
T	Temperature	K

(continued)

Table 4.1 *(continued)*

Symbols	Nomenclature	Units
T_B	Background temperature	K
T_d	Detector temperature	K
T_0	Sink temperature	K
t	Time	s
V	dc voltage	V
V_B	dc bias voltage	V
V_{bd}	dc breakdown voltage	V
V_0	dc open-circuit voltage	V
V_p	Peltier voltage	V
v,v_s,v_n	ac rms generalized, rms signal, or rms noise voltage	V
v_c	ac calibration signal voltage, rms	V
v_n^*	Root-noise-power-spectrum	V $Hz^{-1/2}$
v_0	ac open-circuit voltage	V
$v(t),v_s(t),v_n(t)$	Instantaneous generalized, signal, or noise voltage	V
v_T	Thermal noise voltage, rms	V
W_h	Heat generated in a detector due to I^2R_d heating	J
w	Width of the detector	cm
Z	Impedance	Ω
$\mathscr{Z}$	Thermal impedance	K W^{-1}
$\tilde{Z}$	Complex impedance	Ω
Greek:		
α	Temperature coefficient of resistance	K^{-1}
β	Efficiency factor for a photodiode	—
γ	Coherence factor	—
ΔT_d	Temperature change of the detector	K
δ	Phase angle	rad
ε	Emissivity	—
η	Quantum efficiency	—
λ	Optical wavelength	μm
λ_c	Cutoff wavelength	μm
λ_p	Peak wavelength	μm
λ_s	Signal wavelength	μm
μ	Carrier mobility	$cm^2\ s^{-1}\ V^{-1}$
μ_e	Electron mobility	$cm^2\ s^{-1}\ V^{-1}$
μ_h	Hole mobility	$cm^2\ s^{-1}\ V^{-1}$
ν	Optical frequency	s^{-1}
π_{ab}	Peltier coefficient	V
ρ	Surface reflectance	—
σ	Stefan-Boltzmann constant	$W\ m^{-2}\ K^{-4}$
σ_c	Capture cross section	cm^2

Table 4.1 *(continued)*

Symbols	Nomenclature	Units
σ_e	Electrical conductivity	mho cm^{-1}
τ	Time constant	s
τ_c	Average, free-charge carrier lifetime	s
τ_e	Electrical time constant	s
τ_{el}	Electron lifetime	s
τ_h	Hole lifetime	s
τ_T	Thermal time constant	s
Φ	Flux, or radiant power	W
$\Phi(t)$	Instantaneous radiant power	W
Φ_B	Background radiant power	W
$\Phi_{q,\lambda}(\lambda)$	Photon flux per unit wavelength	photons s^{-1} μm^{-1}
$\Phi_{q,\lambda,B}(\lambda)$	Photon flux per unit wavelength from the background	photons s^{-1} μm^{-1}
$\Phi_{q,\lambda,s}(\lambda)$	Photon flux per unit wavelength from the signal	photons s^{-1} μm^{-1}
Φ_s	rms signal radiant power	W
$\Phi_\lambda(\lambda)$	Spectral radiant power or flux	W μm^{-1}
ϕ	Surface work function of a material	J C^{-1}
χ_i	Phase shift between input flux and output voltage	rad
Ω	Solid angle (field of view)	sr
Ω_e	Effective, weighted detector solid angle	sr
ω	Angular frequency	rad s^{-1}

are separated en route by a magnetic field. This charge separation produces an output voltage that is directly proportional to the number of incident photons.

Photovoltaic Process. A change in the number of photons incident on a semiconductor *p-n* junction causes a change in the current generated by the junction.

Pyroelectric Process. The incident infrared radiation increases the temperature of the crystalline responsive element. This temperature change alters the dipole moment, which produces an observable, external, electric field.

Thermopneumatic Process. The radiation incident on a gas in a chamber increases the temperature (and therefore the pressure) of the gas, causing the chamber to expand, thus moving a mirror attached to an exterior wall. This movement can be detected optically.

Thermovoltaic Process. The temperature of a junction of dissimilar metals is varied by changes in the level of incident radiation absorbed at the junction and thus causes the voltage generated by the junction (due to the Seebeck effect) to fluctuate.

4.1.4 Windows

Windows are used to isolate the ambient environment from the special environment often required around the responsive element. In cooled detectors,

Table 4.2 Detector Parameters

Name, Symbol, and Preferred Units	Definition	Equation Definition	Functional Relationship
Responsive area A_d, cm^2	For responsive elements made of thin films or single crystals, the responsive area is usually the geometric area. For detectors using integrating chambers, the responsive area is the entrance aperture. An effective area A_e can be defined by integrating a normalized responsivity over the responsive area.	A_d = area of detector (geometric) $A_e = \int_{A_d}\int \frac{\mathcal{R}(x,y)\,dx\,dy}{\mathcal{R}_{max}}$ where $\mathcal{R}_{max}$ = maximum value of $\mathcal{R}(x,y)$ $\mathcal{R}$ = responsivity x,y = coordinates in plane of responsive area	
Impedance Z_d, ohms	The slope of the instantaneous voltage – instantaneous current curve at bias voltage V_B.	$Z_d = \left.\frac{dv(t)}{di(t)}\right\|_{V_B}$	Z_d is a function of the bias voltage, the interelectrode capacitance, and the level of irradiance.
Resistance R_d, ohms	The ratio of the dc voltage across the detector to the dc current through it.	$R_d = V/I$	R_d is a function of the detector temperature and the level of irradiance.
Background temperature T_B, K	The temperature of a uniform blackbody completely filling the detector field of view that would give the observed total flux on the detector.		
Detector solid angle Ω, sr	The solid angle (field of view) from which the detector receives radiation.		

Table 4.2 *(continued)*

Name, Symbol, and Preferred Units	Definition	Equation Definition	Functional Relationship
Effective, weighted detector solid angle Ω_e, sr	The solid angle (field of view), weighted by a cosine function, from which the detector receives radiation.	$\Omega_e = \int_{A_d}\int\left[\int_0^{\pi/2}\int_0^{2\pi}\frac{\cos\theta \sin\theta\, \mathcal{R}(x,y,\phi,\theta)}{A_e \mathcal{R}_{max}(0,0)}\, d\phi\, d\theta\right] dx\, dy$ where $\mathcal{R}_{max}(0,0)$ = the maximum value of $\mathcal{R}(x,y,0,0)$; measured with a small field of view $d\Omega$ ϕ and θ = spherical coordinates, with ϕ being the azimuthal angle $\cos\theta$ = the weighting function The z axis is normal to the plane of the responsive element. If the responsivity is not a function of ϕ, the element is said to have circular symmetry and $\Omega = \pi \sin^2(\Theta/2)$ where Θ is the total cone angle.	

(continued)

Table 4.2 *(continued)*

Name, Symbol, and Preferred Units	Definition	Equation Definition	Functional Relationship
Instantaneous signal voltage $v_s(t)$ or current $\mathcal{J}_s(t)$, V or A, respectively	That component of the electrical output voltage (or current) that is coherent with $\Phi_s(t)$, the instantaneous value of the input signal radiant power; $\Phi_s(t)$ can be monochromatic or have a blackbody character.	If the incident radiant power $\Phi_s(t)$ is periodic in time: $\Phi_s(t) = \Phi_0 + \Phi_1 \times \cos(2\pi ft + \delta_1) + \Phi_2 \times \cos(2 \cdot 2\pi ft + \delta_2) + \ldots$ then $v_s(t) = V_0 + V_1 \cos(2\pi ft + \psi_1) + V_2 \cos(2 \cdot 2\pi ft + \psi_2) + \ldots$	The signal voltage is a function of electrical frequency f. For a signal-time-constant, $v_s = \frac{v_{s,\max}}{(1 + 4\pi^2 f^2 \tau^2)^{1/2}}$
rms amplitude of the fundamental signal voltage or current component v_s or i_s, V or A, respectively	The rms amplitude of the fundamental signal component determined by taking the square root of the time average of the square of the first time-varying component in the series, i.e., the fundamental.	$v_s = (2)^{-1/2} V_1$	v_s is related to f, and bias voltage.
rms noise voltage v_n or current i_n, V or A	That component of the electrical output voltage (or current) that is incoherent with the signal radiant power. This value is determined with the signal power removed.	If the dc gain of the associated electronics is zero, $v_n = [\overline{v_n^2(t)}]^{1/2}$	v_n is related to the detector area, Δf, f, and in some cases to Ω and T_B.

Table 4.2 *(continued)*

Name, Symbol, and Preferred Units	Definition	Equation Definition	Functional Relationship
Spectral responsivity $\mathcal{R}(\lambda)$, $V\,W^{-1}$ or $A\,W^{-1}$	The ratio of the rms signal voltage (or current) to the rms value of the monochromatic incident signal power, referred to an infinite load impedance and to the terminals of the detector.	$\mathcal{R}(\lambda) = v_s/\Phi_{s,\lambda}\Delta\lambda$	Responsivity is a function of λ, f, T, and bias voltage or current.
Blackbody responsivity $\mathcal{R}_{BB}$, $V\,W^{-1}$ or $A\,W^{-1}$	Same as spectral responsivity except that the incident signal power is from a blackbody.	$\mathcal{R}_{BB} = v_s/\Phi_{s,BB}$	Responsivity is a function of f, T, and bias voltage or current.
Time constant τ, s	A measure of the detector's speed of response. The alternative equations for τ (next column) become identical if the noise has a flat power spectrum (see Sec. 4.4.3) and if the responsivity varies with frequency according to the relation $\mathcal{R}(f) = \dfrac{\mathcal{R}(0)}{(1 + 4\pi^2 f^2 \tau^2)^{1/2}}$	(a) The decay time constant is given as $\tau = 1/(2\pi f_c)$ where f_c is that chopping frequency at which the responsivity has fallen to $2^{-1/2}$ of its maximum value. (b) The rise time constant is the time required for the signal voltage (or current) to rise to $1 - 1/e$ or 0.63 times its asymptotic value. It is measured by the light pulse method: exposing the detector to a square-wave pulse of radiation.	

(continued)

Table 4.2 *(continued)*

Name, Symbol, and Preferred Units	Definition	Equation Definition	Functional Relationship
Time constant τ, s		(c) Responsive time constant $\tau_r = \dfrac{\mathcal{R}_{\max}^2}{4\int_0^\infty [\mathcal{R}(f)]^2\, df}$ (d) Detective time constant $\tau_d = \dfrac{(D^*_{\max})^2}{4\int_0^\infty [D^*(f)]^2\, df}$ (e) Empirical responsive time constant $\tau_{rs} = \dfrac{1}{2\pi} \times \left\{ \dfrac{[\mathcal{R}(f_1)]^2 - [\mathcal{R}(f_2)]^2}{[f_2\mathcal{R}(f_2)]^2 - [f_1\mathcal{R}(f_1)]^2} \right\}^{1/2}$ where f_1 and f_2 must be specified.	

Table 4.2 *(continued)*

Name, Symbol, and Preferred Units	Definition	Equation Definition	Functional Relationship
Time constant τ, s		(f) Empirical detective time constant $\tau_{ds} = \frac{1}{2\pi} \times \left\{ \frac{[D^*(f_1)]^2 - [D^*(f_2)]^2}{[f_2 D^*(f_2)]^2 - [f_1 D^*(f_1)]^2} \right\}^{1/2}$ where f_1 and f_2 must be specified.	
Spectral noise equivalent power NEP_λ, W	That value of monochromatic incident rms signal power of wavelength λ required to produce an rms signal-to-rms-noise ratio of unity. The chopping frequency, the electrical bandwidth used in the measurement, and the detector area should be specified.	$NEP_\lambda = \Phi_{s,\lambda} \Delta\lambda \left(\frac{v_n}{v_s} \right) = \frac{v_n}{\mathcal{R}_\lambda}$	Depends on λ, A, f, Δf, and in some cases Ω and T_B.
Blackbody noise equivalent power NEP_{BB}, W	That value of incident rms signal power (with a blackbody spectral character) required to produce an rms signal-to-rms-noise ratio of unity. The blackbody temperature must be specified along with the detector area, the electrical bandwidth used in the measurement, and the chopping frequency.	$NEP_{BB} = \Phi_{s,BB} \left(\frac{v_n}{v_s} \right) = \frac{v_n}{\mathcal{R}_{BB}}$	Depends on blackbody temperature A, f, Δf, and in some cases Ω and T_B.

(continued)

Table 4.2 *(continued)*

Name, Symbol, and Preferred Units	Definition	Equation Definition	Functional Relationship
Spectral detectivity $D(\lambda)$, W^{-1}	The reciprocal of spectral noise equivalent power. The chopping frequency, the electrical bandwidth used in the measurement, and the detector sensitive area should be specified.	$D(\lambda) = 1/\mathrm{NEP}_\lambda$	Depends on λ, A, f, Δf, and in some cases Ω and T_B.
Blackbody detectivity D_{BB}, W^{-1}	The reciprocal of the blackbody noise equivalent power. The blackbody temperature should be specified, along with the electrical bandwidth used in the measurement, the detector area, and the chopping frequency.	$D_{BB} = 1/\mathrm{NEP}_{BB}$	Depends on blackbody temperature, A, f, Δf, and in some cases Ω and T_B.
Spectral D-star, $D^*(\lambda,f_c)$, cm $Hz^{1/2}$ W^{-1} (Jones)	A normalization of spectral detectivity to take into account the area and electrical bandwidth dependence. The chopping frequency (f_c) used in the measurement is specified by inserting it in the parentheses as indicated in the last column. For detectors limited by the fluctuation in arrival rate of background photons, Ω and T_B must be specified.	$D^*(\lambda,f_c) = (A_d \Delta f)^{1/2} D(\lambda)$	For background-noise-limited detectors, $D^*(\lambda,f_c)$ depends on Ω and T_B.
Blackbody D-star $D^*(T_{BB},f_c)$, cm $Hz^{1/2}$ W^{-1} (Jones)	A normalization of blackbody detectivity to take into account the detector area and the electrical bandwidth. The chopping frequency (f_c) and the blackbody temperature (T_{BB}) are specified in the parentheses as indicated. For detectors that are background noise limited, Ω and T_B must also be specified.	$D^*(T_{BB},f_c) = (A_d \Delta f)^{1/2} D_{BB}$	For background-noise-limited detectors, $D^*(T_{BB},f_c)$ depends on Ω and T_B.

Table 4.2 *(continued)*

Name, Symbol, and Preferred Units	Definition	Equation Definition	Functional Relationship
Maximized D-star $D^*(\lambda_p, f_c)$, cm $Hz^{1/2}$ W^{-1} (Jones)	A quantity obtained when the wavelength is λ_p and the chopping frequency used yields a maximum rms signal-to-rms-noise ratio.	$D^*(\lambda_p, f_c) = \frac{(A_d \Delta f)^{1/2}}{\text{NEP}_{\lambda_p}}$	Same as for $D^*(\lambda, f)$.
Spectral D-double star $D^{**}(\lambda, f_c)$, cm $Hz^{1/2}$ $sr^{1/2}$ W^{-1}	A normalization of $D^*(\lambda, f_c)$ to account for the detector effective weighted field of view Ω_e. (Note: if $\Omega_e = \pi$, $D^{**} = D^*$.)	$D^{**}(\lambda, f_c) = (\Omega/\pi)^{1/2} D^*(\lambda, f_c)$	—
Peak wavelength λ_p or λ_{max}, μm	The wavelength at which spectral detectivity is a maximum.	—	Depends on cell temperature and detector material used.
Cutoff wavelength λ_c μm	The wavelength at which $D^*(\lambda, f_c)$ has degraded to one-half its peak value.	—	Depends on cell temperature and detector material used.
Responsive quantum efficiency RQE	The ratio of the number of countable output events, N_o, to the number of incident photons, N_p.	$\text{RQE} = N_o/N_p$	Depends on bias voltage, time constant, and cell geometry.
Detective quantum efficiency DQE	The square of the ratio of measured detectivity to the theoretical limit of detectivity. Both detectivities must be for the same set of conditions.	$\text{DQE} = \left(\frac{D(\lambda) \text{ measured}}{D(\lambda) \text{ theoretical limit}}\right)^2$	—
D^*f^*	The product of the maximum D-star and f^*, the highest frequency at which $D^*(f)$ has decreased to $2^{-1/2}$ of its maximum value.	D^*f^*, where $D^*(f^*) = 2^{-1/2} D^*_{max}$	—

Table 4.3 Detector Noises (from Ref. 2)

Type of Noise	Physical Mechanism	Detectors Concerned	Equation for v_n
Johnson (also called Nyquist or thermal)	At thermal equilibrium the random motion of charge carriers in a resistive element generates a random electrical voltage across the element. As the temperature of the resistor is increased, the mean kinetic energy of the carriers increases, yielding an increased electrical noise voltage.	All detectors	$v_n = (4kT_d R_d \Delta f)^{1/2}$
Temperature	The fluctuations in temperature of the sensitive element, due to either radiative exchange with the background or conductive exchange with the heat sink, produce a fluctuation in a signal voltage. For thermal detectors, the detector is said to be at its theoretical limit if the temperature noise is due to radiative exchange with the background.	All detectors but especially those made of thin films	For thermal detectors, $\overline{\Delta T^2} = \dfrac{4kT_d^2 \mathscr{G} \Delta f}{\mathscr{G}^2 + 4\pi^2 f^2 \mathscr{C}^2}$ The relation between $\overline{\Delta T^2}$ and v_n should be determined for each detector.
Modulation (or $1/f$)	The mechanism is not well understood. As its name implies, it is characterized by a $1/f^n$ noise power spectrum, where n varies from 0.8 to 2.	All detectors	$v_n \propto R_d I \left(\dfrac{\Delta f}{A_d d}\right)^{1/2} \left(\dfrac{1}{f}\right)^n$
Generation-recombination, G-R	Statistical fluctuations in the rate of generation and in the rate of recombination of charge carriers in the sensitive element result in an electrical noise. These fluctuations can be caused by charge carrier/photon interactions or by the random arrival rate of photons from the background. If the background photons are the prime contributors to the fluctuation in G-R rates, then the noise is often called photon, radiation, or background noise.	All photon detectors	$v_n = R_d I \left[\dfrac{2\tau \Delta f}{N(1 + 4\pi^2 f^2 \tau^2)}\right]^{1/2}$ For photovoltaic detectors, the value of v_n is smaller by a factor of $\sqrt{2}$ since only fluctuations in the generation rate of free charge carriers contribute to the noise. Fluctuations in the rate of recombination of free charge carriers do not affect the detector output voltage.

Table 4.3 *(continued)*

Type of Noise	Physical Mechanism	Detectors Concerned	Equation for v_n
Shot	Noise caused by the discreteness of electronic charge. The current I, flowing through the responsive element, is the result of current pulses produced by the individual electrons and/or holes.	Photovoltaic detectors and thin-film detectors	$v_n = R_d(2eI\Delta f)^{1/2}$ where e is the charge of an electron.

the responsive element is kept in a vacuum. The window affects the spectral distribution of photons incident on the responsive element.

4.1.5 Apertures

Apertures are used to restrict the field of view of the responsive element. This is often done in cooled detectors that are photon-noise limited to cut down on the extraneous background photons and thus reduce noise (see Sec. 4.4).

4.1.6 Dewar Flask

Dewar flasks are used to house the coolant needed to reduce the operating temperature of the responsive element and thus improve detectivity.

4.2 THEORETICAL DESCRIPTIONS OF THERMAL DETECTORS

As indicated earlier, thermal detectors rely on one of four basic processes to accomplish infrared radiation detection. The four processes are:

1. the bolometric effect
2. the thermovoltaic effect
3. the thermopneumatic effect
4. the pyroelectric effect.

The elementary theory of each process is given below.

4.2.1 Bolometers

The bolometric effect is a change in the electrical resistance of the responsive element due to temperature changes produced by absorbed, incident infrared radiation. Figures 4.1 and 4.2 show two electronic circuit configurations that use this effect.

When the bridge circuit is used (Fig. 4.1), the two detectors are placed close to each other with one shielded from any incident radiation in excess of the ambient levels. The bridge is balanced when no excess radiation is on the exposed detector. Incident infrared radiation will then cause a rise in the temperature of the exposed detector, thereby causing a change in its resistance. This electrically unbalances the bridge, causing a current to flow through R_2. In the ac circuit of Fig. 4.2, only changes in voltage across the bolometer pass through the coupling capacitor to the electronics.

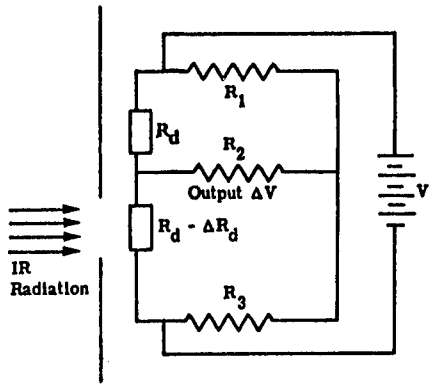


Fig. 4.1 Bolometer detector circuit with bridge configuration for dc operation.

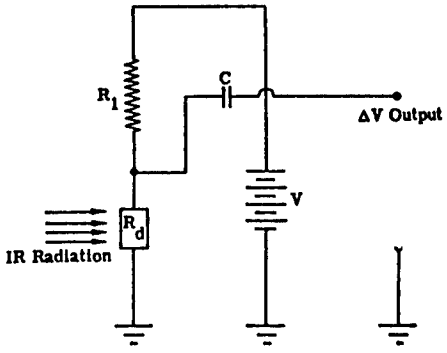


Fig. 4.2 Bolometer detector circuit.

The change in electrical resistance resulting from the increased temperature of the bolometer depends on the temperature coefficient of resistance α, which is

$$\alpha = \frac{1}{R_d}\frac{dR_d}{dT_d} , \tag{4.1}$$

where R_d is the resistance of the detector and T_d is the temperature of the detector.

The signal equations for the circuits in Figs. 4.1 and 4.2 are, respectively,

$$v_s = \Delta v = \frac{I(\Delta R_d)R_2}{2R_2 + R_1 + R_3} , \tag{4.2}$$

$$v_s = \Delta v = \frac{R_1 V \Delta R_d}{(R_d + R_1)^2} , \tag{4.3}$$

$$\Delta R_d = \frac{dR_d}{dT_d}\Delta T_d , \tag{4.4}$$

where

I = the steady-state current through the bolometers in the bridge circuit
$\Delta R_d = (dR_d/dT_d)\Delta T_d$
ΔT_d = the time variation of T_d
Δv = the resulting change in voltage
v_s = the ac signal voltage
V = the dc bias voltage.

Figures 4.1 and 4.2 identify R_1, R_2, and R_3, and R_d and T_d are as in Eq. (4.1).

Responsivity. The responsivity $\mathcal{R}$ is defined as

$$\mathcal{R} = \frac{\Delta v}{\Delta \Phi} , \tag{4.5}$$

where Δv is the open-circuit voltage appearing across the load resistor for an incremental increase in the infrared radiation power input, $\Delta\Phi$. The increase in bolometer temperature caused by $\Delta\Phi$ is ΔT_d, which is expressed in the following differential equation:

$$\mathscr{C}\frac{d\Delta T_d}{dt} + \mathscr{G}_0 \Delta T_d = W_h + \Delta\Phi , \tag{4.6}$$

where

$\mathscr{C}$ = the heat capacitance of the bolometer element, J K^{-1}
$\mathscr{G}_0 \Delta T_d$ = the conductive and radiative heat flow for the element
W_h = the thermal power generated in the bolometer due to I^2R_d heating.

In the steady-state condition,

$$\mathscr{G}_0 \Delta T_d = W_h = I^2 R_d . \tag{4.7}$$

From Eqs. (4.6) and (4.7), one can write the following equation when ΔT_d is small:

$$\mathscr{C}\frac{d\Delta T_d}{dt} + \mathscr{G}\Delta T_d = \frac{dW_h}{dT}\Delta T_d + \Delta\Phi , \tag{4.8}$$

where $\mathscr{G}$ is the thermal conductance defined for small temperature changes, in units of W K^{-1}. The rate of change of W_h with T depends on the electronic circuit arrangement. For the circuit shown in Fig. 4.2,

$$\frac{dW_h}{dT} = \alpha W_h\left(\frac{R_1 - R_d}{R_1 + R_d}\right) \tag{4.9}$$

$$= \alpha(\Delta T_d)\mathscr{G}_0\left(\frac{R_1 - R_d}{R_1 + R_d}\right) . \tag{4.10}$$

Equation (4.8) can now be rewritten

$$\mathscr{C}\frac{d\Delta T_d}{dT} + \mathscr{G}_e \Delta T_d = \Delta\Phi , \tag{4.11}$$

where $\mathscr{G}_e$ is the effective thermal conductance, given as

$$\mathscr{G}_e = \mathscr{G} - \alpha\mathscr{G}_0(\Delta T_d)\left(\frac{R_1 - R_d}{R_1 + R_d}\right) . \tag{4.12}$$

If $\mathcal{G}_e < 0$, then Eq. (4.11) has an exponentially increasing solution when $\Delta\Phi = 0$. The bolometer is unstable in this condition and will burn out. For stable operation, the requirement is

$$\mathcal{G} > \mathcal{G}_0\alpha(\Delta T_d) \ , \tag{4.13}$$

where R_1 is chosen such that $R_1 \gg R_d$ for maximizing the signal voltage Δv. The solution of Eq. (4.11) is given in Eq. (4.14) for a sinusoidally varying input radiation function ($\Delta\Phi = \Delta\Phi \cos\omega t$):

$$\Delta T_d = \frac{\varepsilon\Delta\Phi_0}{\mathcal{G}_e(1 + \omega^2\tau^2)^{1/2}} \ , \tag{4.14}$$

where

τ = $\mathcal{C}/\mathcal{G}_e$
ε = the emissivity of the bolometer
$\Delta\Phi_0$ = the periodic function with angular frequency ω and a peak amplitude $\Delta\Phi_{max}$.

The thermal response to an arbitrary, periodic radiation impact can be determined by expressing the arbitrary periodic function in terms of its Fourier series components and applying the superposition principle. From Eq. (4.1) it can be shown that

$$\Delta R_d = \Delta T_d R_d \alpha \ . \tag{4.15}$$

Therefore,

$$\Delta R_d = \frac{R_d\alpha\varepsilon\Delta\Phi_0}{\mathcal{G}_e(1 + \omega^2\tau^2)^{1/2}} \ . \tag{4.16}$$

The responsivity $\mathcal{R}$ of the bolometer in the circuit shown in Fig. 4.2 is obtained by combining Eqs. (4.16) and (4.3) to obtain the following expression:

$$\mathcal{R} = \left(\frac{R_1}{R_1 + R_d}\right) \frac{I\varepsilon R_d\alpha}{\mathcal{G}_e(1 + \omega^2\tau^2)^{1/2}} \ . \tag{4.17}$$

For the bridge circuit shown in Fig. 4.1, the responsivity becomes

$$\mathcal{R} = \frac{1}{2}\varepsilon I R_d \alpha \frac{1}{G_e} \ . \tag{4.18}$$

Several numerical examples given in Ref. 2 are presented below.

Case 1: An Ideal, Metal Bolometer with Predominantly Conductive Cooling. The temperature coefficient of resistance is given as $\alpha = (1/R_d)\,(dR_d/dT_d)$.

For a metal, the resistance over a wide temperature range is approximately proportional to the temperature so $\alpha \approx 1/T_d$. If one assumes that $R_1 \gg R_d$ and $\mathcal{G} \approx \mathcal{G}_0$, then Eq. (4.12) becomes

$$\mathcal{G}_e = \mathcal{G}\left[1 - \left(\frac{\Delta T_d}{T_d}\right)\right] . \tag{4.19}$$

If $T_0 = 300$ K, $T_d = 450$ K, $R_d = 50\ \Omega$, and $\mathcal{G} = \mathcal{G}_0 = 10^{-4}$ W K^{-1}, then $\mathcal{G}_e = 6.7 \times 10^{-5}$. Using the assumptions above, and assuming $\omega^2\tau^2 \ll 1$, one can simplify Eq. (4.17):

$$\mathcal{R} = \frac{IR_d\varepsilon}{T_d(6.7 \times 10^{-5})} . \tag{4.20}$$

If one sets $\varepsilon = 1$ and solves for I in Eq. (4.7), the responsivity $\mathcal{R}$ becomes 30 V W^{-1}.

Case 2: Metal Bolometer, Predominantly Conductive Cooling. Assume $\mathcal{G} = \mathcal{G}_0 = 10^{-4}$ W K^{-1}, α = constant, $\varepsilon = 1$, $\omega^2\tau^2 \ll 1$, $T_0 = 300$ K, $R_d = 50\ \Omega$, $R_1 \gg R_d$, and $T_d = 375$ K.

From Eq. (4.12),

$$\mathcal{G}_e = \mathcal{G}[1 - \alpha(\Delta T)] . \tag{4.21}$$

To prevent thermal instability and detector burnout, $(\Delta T_d) < 1/\alpha$. If $\Delta T_d = 1/2\alpha$, then $\mathcal{G}_e = (1/2)(\mathcal{G})$. The responsivity equation is

$$\mathcal{R} = \frac{I\varepsilon R_d\alpha}{\mathcal{G}_e} , \tag{4.22}$$

and I is computed using Eq. (4.7). Substitution into Eq. (4.22) gives $\mathcal{R} = 82.4$ V W^{-1}.

Case 3: Semiconducting Bolometer. Assume that $\alpha = -10T_0/T^2$, $\mathcal{G} = \mathcal{G}_0 = 10^{-4}$ W K^{-1}, $R_1 \gg R_d$, $\varepsilon = 1$, $\omega^2\tau^2 \ll 1$, $T_d = 315$ K, $T_0 = 300$ K, and $R_d = 10^6\ \Omega$:

$$\begin{aligned} \mathcal{G}_e &= \mathcal{G} - \alpha\mathcal{G}_0(\Delta T_d) \\ &= \mathcal{G}[1 - \alpha(\Delta T_d)] = 5.5 \times 10^{-5} . \end{aligned} \tag{4.23}$$

If one substitutes this value for $\mathcal{G}_e$, as well as the value for I derived from Eq. (4.7) into the responsivity equation, then

$$\begin{aligned} \mathcal{R} &= \frac{IR_d\alpha}{\mathcal{G}_e} \\ &= 21{,}000 \text{ V W}^{-1} . \end{aligned} \tag{4.24}$$

Noise. The noise voltage from commercially available thermistor bolometers is composed of $1/f$ noise (also called current, excess, or modulation noise) and Johnson noise. Current noise is expressed as

$$v_n \propto IR_d\left(\frac{\Delta f}{A_d}\right)^{1/2}\left(\frac{1}{f}\right)^{1/2} \tag{4.25}$$

and Johnson noise is expressed as

$$v_n = (4kT_dR_d\Delta f)^{1/2} \ . \tag{4.26}$$

For bias current values high enough to give maximum detector performance [optimum signal-to-noise ratio (SNR)], $1/f$ noise predominates throughout most of the useful part of the frequency spectrum to which the detector is responsive. If the bias current is reduced sufficiently, then the $1/f$ noise is reduced and the Johnson noise predominates. In this case, the spectrum of the detector noise is flat, depending only on the resistance and temperature of the responsive element.

Some bolometers have been specially designed and built to operate at low temperatures to increase detector sensitivity and decrease the time constant. Significant improvements in the detectivity, D^*, and time constant have been observed. These cooled bolometers have not done well commercially because of their increased complexity and cost, caused by the need for cryogenic apparatus. Photon detectors are more attractive by comparison.

4.2.2 Thermocouples and Thermopiles

A junction of two dissimilar materials will, when heated, produce a voltage across the two open leads. This is the thermovoltaic effect. Such a junction is called a *thermocouple.* When more than one thermocouple is combined in a single responsive element, it is termed a *thermopile.*

Figure 4.3 contains a schematic of a thermocouple made of two dissimilar materials, A and B, connected with an electrical conductor C. Junction J_1 is attached to the responsive element that is irradiated with infrared radiation. Upon absorbing the infrared radiation, the temperature of the responsive element increases from T_d to $(T_d + \Delta T_d)$, which causes heating at J_1. If one assumes that the temperature at J_1 is also $T_d + \Delta T_d$, then the open-circuit thermoelectric electromotive force (emf) established in the circuit is

$$V_0 = P_{ab}\Delta T_d \ , \tag{4.27}$$

where P_{ab} is a characteristic of the two materials and is known as the *thermoelectric* power. When radiation is incident on the responsive element, thus heating it, a current will flow in the circuit. This current will flow through the junction and tend to cool it by the Peltier effect. The cooling, ΔW_h, is given by

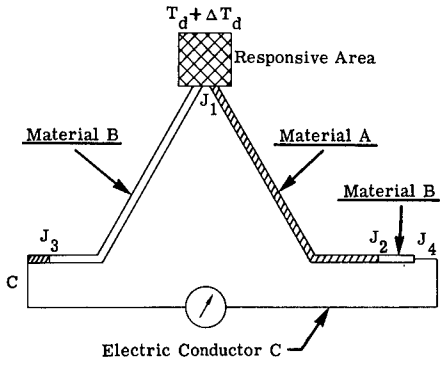


Fig. 4.3 Thermocouple schematic.

$$\Delta W_h = -\pi_{ab} I \ , \tag{4.28}$$

where π_{ab} is known as the Peltier coefficient. The quantity π_{ab} is related to P_{ab} by

$$\pi_{ab} = T_d P_{ab} \ . \tag{4.29}$$

As soon as current flows, the value of ΔT_d is changed by the Peltier cooling of the junction, J_1. If the cold junction, J_2, is kept at a constant temperature, then, using Eqs. (4.28) and (4.29), one can show that the hot junction will be cooled at a rate $IP_{ab}\Delta T_d$. If $\mathcal{Z}$ is the thermal impedance of the hot junction plus the responsive element, then

$$\Delta(\Delta T_d) = IP_{ab}\mathcal{Z}T_d \ . \tag{4.30}$$

This cooling induces an emf, V_p, in the circuit, where

$$V_p = P_{ab}\Delta(\Delta T_d) = -IP_{ab}^2\mathcal{Z}T_d \ . \tag{4.31}$$

The total emf, V_t, caused by the increased temperature at J_1 resulting from the incident infrared radiation and the Peltier effect is

$$V_t = V_0 + V_p = V_0 - IP_{ab}^2\mathcal{Z}T_d \ . \tag{4.32}$$

The current I is related to V_t by

$$I = \frac{V_t}{(R_d + R_m)} \ , \tag{4.33}$$

where R_d is the detector resistance and R_m is the meter resistance. Therefore,

$$V_0 = I(R_d + R_m + P_{ab}^2 \mathscr{L} T_d) \ . \tag{4.34}$$

This shows that, in effect, Peltier cooling increases the electrical resistance of the circuit by the dynamic resistance of the thermocouple,

$$R_{\text{dyn}} = P_{ab}^2 \mathscr{L} T_d \ . \tag{4.35}$$

When there is constant, external radiant power Φ feeding into the responsive element J_1, the equation for the balance of the heat flow becomes

$$\frac{\Delta T_d}{\mathscr{L}} = \Phi - P_{ab} I T_d \ . \tag{4.36}$$

The emf produced by Φ is $P_{ab}\Delta T_d$, and the current in the circuit is $P_{ab}\Delta T_d(R_d + R_m)^{-1}$. If one solves the latter expression for ΔT_d, Eq. (4.35) for T_d, and substitutes into Eq. (4.36), then

$$I = P_{ab} \mathscr{L} \Phi (R_d + R_m + R_{\text{dyn}})^{-1} \ . \tag{4.37}$$

Therefore,

$$\Delta T_d = \mathscr{L} \Phi (R_d + R_m)(R_d + R_m + R_{\text{dyn}})^{-1} \ . \tag{4.38}$$

The open-circuit emf, V_0, and the level of ΔT_d for the open-circuit case, ΔT_0, can be computed by omitting the Peltier heating term:

$$\Delta T_0 = \mathscr{L} \Phi \tag{4.39}$$

and

$$V_0 = P_{ab} \Delta T_0 = P_{ab} \mathscr{L} \Phi \ . \tag{4.40}$$

Therefore, Eq. (4.37) can be rewritten in terms of the open-circuit voltage V_0:

$$I = \frac{V_0}{R_d + R_m + R_{\text{dyn}}} \tag{4.41}$$

and ΔT_d can be written in terms of ΔT_0:

$$\Delta T_d = \Delta T_0 \frac{(R_d + R_m)}{(R_d + R_m + R_{\text{dyn}})} \ . \tag{4.42}$$

Responsivity $\mathscr{R}$ of a thermovoltaic detector for constant input power Φ is then

$$\mathscr{R} = \frac{V_0}{\Phi} = \varepsilon P_{ab} \mathscr{L} \ , \tag{4.43}$$

where ε is the fraction of incident infrared power that is absorbed. To obtain a high responsivity, materials with a high value of thermoelectric power and

high thermal resistance should be selected.

The responsivity of a thermovoltaic detector to an alternating input power is given as

$$\mathcal{R} = \varepsilon P_{ab} \mathcal{L} (1 + \omega^2 \tau^2)^{-1/2} , \tag{4.44}$$

where

τ = time constant, $\tau = \mathcal{L}\mathcal{C}$
$\mathcal{C}$ = the thermal capacitance of the responsive element
ω = the angular frequency of the alternating input power Φ.

A derivation of this equation is given in Ref. 2.

The time constant τ of evaporated thermopiles ranges from 4 to 50 ms, depending on the type and thickness of the radiation-absorbing material used on the thermopile surface. This absorbing material increases the thermal capacitance of the responsive element. If the responsive element is enclosed in a sealed housing, the thermal conductive paths will affect the time constant. A high, effective thermal conductance leads to a decreased time constant.

According to manufacturers' data, Johnson noise is the predominant noise mechanism in currently produced thermocouples and thermopiles.

4.2.3 Thermopneumatics

In this detection process, an infrared radiation absorbing element is placed in a chamber filled with gas. A window in one of the chamber walls allows incident infrared radiation to irradiate the absorbing element. When an increased flux of infrared radiation strikes the element, it heats and, by conduction, heats the gas in the chamber. The increase in temperature of the gas results in increased pressure in the chamber, which distorts a thin, flexible mirror mounted on one of the chamber walls. The degree of distortion is sensed by a separate optical system consisting of a light source and a detector. Golay[3] has described the theory of operation for this type of detector.

4.2.4 Pyroelectrics

When the responsive element absorbs an incrementally increased amount of infrared radiation, its temperature rises, changing its surface charge. With the appropriate external circuit, this change in surface charge leads to a signal voltage. The change in temperature ΔT_d of the responsive element is related to its thermal capacitance $\mathcal{C}$ and to its thermal conductance $\mathcal{G}$ by the following equation:

$$\mathcal{C} \frac{d\Delta T_d}{dt} + \mathcal{G} \Delta T_d = \Phi , \tag{4.45}$$

where Φ is the incident incremental infrared radiation. The solution of this equation for a periodic incident infrared radiation power Φ is

$$\Delta T_d = \varepsilon \Phi \mathcal{G}^{-1} \left[1 + \omega^2 \left(\frac{\mathcal{C}}{\mathcal{G}} \right)^2 \right]^{-1/2} , \tag{4.46}$$

where ε is emissivity and ω is the angular frequency of Φ. This analysis assumes that the radiation is absorbed uniformly throughout the sample. Putley[4] has derived the responsivity for a pyroelectric detector used in the circuit shown in Fig. 4.4:

$$v_s = I_p |Z| = I_p R_e (1 + \omega^2 \tau_e^2)^{1/2} , \tag{4.47}$$

where

I_p = $\omega \mathcal{P} A_d \Delta T_d$
$\mathcal{P}$ = the pyroelectric coefficient
Z = impedance
R_e = equivalent input resistance of the detector-preamplifier circuit
τ_e = $R_e C_e$
C_e = equivalent capacitance.

Therefore, voltage v_s can be expressed as

$$v_s = \omega \mathcal{P} A_d \Delta T_d(\omega) R_e (1 + \omega^2 \tau_e^2)^{-1/2} , \tag{4.48}$$

where A_d is the sensitive area of the responsive crystal. Substituting the expression for ΔT_d from Eq. (4.46) into Eq. (4.48), one gets

$$v_s = \omega \mathcal{P} A_d \varepsilon \Phi R_e \mathcal{G}^{-1} \left[1 + \omega^2 \left(\frac{\mathcal{C}}{\mathcal{G}} \right)^2 \right]^{-1/2} [1 + \omega^2 \tau_e^2]^{-1/2} . \tag{4.49}$$

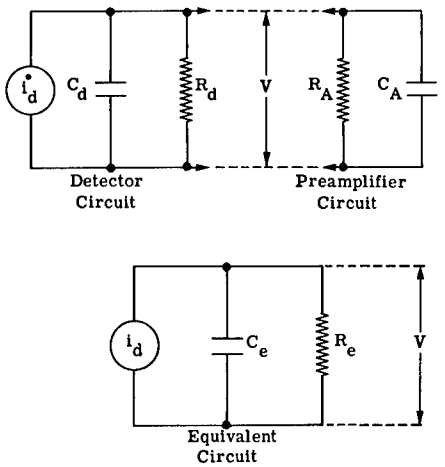


Fig. 4.4 Equivalent circuit for pyroelectric detector and amplifier input.

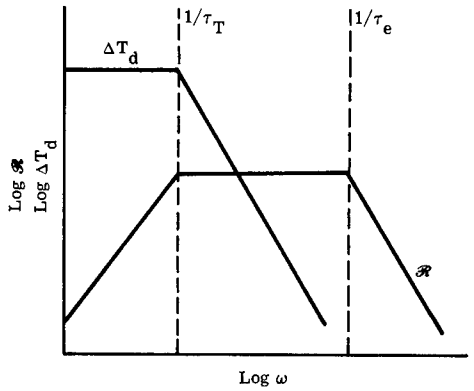


Fig. 4.5 Log-log plots of ΔT and $\mathcal{R}$ versus ω for a pyroelectric detector.

The expression for the responsivity $\mathcal{R}$ is then

$$\mathcal{R} = \frac{v_s}{\Phi} = \frac{\omega \mathcal{P} A_d \varepsilon R_e}{\mathcal{G}} (1 + \omega^2 \tau_T^2)^{-1/2} (1 + \omega^2 \tau_e^2)^{-1/2} , \tag{4.50}$$

where $\tau_T = \mathcal{C}/\mathcal{G}$ is the thermal time constant.

The relationship between the angular frequency ω of the incoming infrared radiation, the incremental temperature rise ΔT_d of the pyroelectrical crystal, and the responsivity $\mathcal{R}$ of the detector is shown in Fig. 4.5. The responsivity is 0 at $\omega = 0$ and increases as ω increases until the angular frequency reaches the value $\omega = 1/\tau_T$. In the range $1/\tau_T \leq \omega \leq 1/\tau_e$, the value of responsivity is a constant. For values of ω larger than τ_e, the responsivity is inversely proportional to ω (Ref. 4).

4.2.5 Theoretical Limit of Performance for Thermal Detectors

The radiation power Φ incident on the responsive element of a thermal detector and the power emitted by the responsive element consist of streams of photons. The rate of arrival and the rate of emission of these photons fluctuate randomly; they are not correlated spatially or temporally to any significant degree. These random arrival and emission rates lead to random fluctuations in the responsive element's temperature, which in turn produces a random output voltage. Smith et al.[2] have shown that fluctuations in arrival and emission rates of photons on a thermal detector lead to a mean square fluctuation in radiation power, $\overline{\Delta\Phi^2}$:

$$\overline{\Delta\Phi^2} = 4kT_d^2 \mathcal{G} \Delta f , \tag{4.51}$$

where

k = Boltzmann's constant
Δf = the electrical frequency bandwidth

T_d = the temperature of the responsive element
$\mathcal{G}$ = the thermal conductance between the responsive element and its surroundings.

For the case of a detector responsive element with an area A_d and constant emissivity ε connected to its surroundings by radiation alone,

$$\mathcal{G} = 4\sigma\varepsilon A_d T_d^3 \ , \tag{4.52}$$

where σ is the Stefan-Boltzmann constant.

Substituting for $\mathcal{G}$ into Eq. (4.51), one gets

$$\overline{\Delta\Phi^2} = 16 A_d k\sigma\varepsilon T_d^5 \Delta f \ . \tag{4.53}$$

For the case of a thermal detector at 300 K with a 1-mm^2 area and $\Delta f = 1$ Hz, the value of $[(\Delta\Phi^2)]^{1/2}$ is 5.5×10^{-12} W. This means that a thermal detector (with a 1-mm^2 area and 300 K temperature), limited only by the fluctuation in power flowing to and from the responsive element (neglecting all other noise sources), will have a minimum detectable power of

$$\text{NEP} = [(\overline{\Delta\Phi^2})]^{1/2} = 5.5 \times 10^{-12} \text{ W} \ . \tag{4.54}$$

The theoretical limit for D^* in this case is 1.8×10^{10} cm Hz$^{1/2}$ W^{-1}.

In the case of bolometers (excluding $1/f$ noise), the minimum detectable power can be expressed as the sum of the mean square noise resulting from temperature fluctuations and Johnson noise:

$$(\text{NEP})^2 = \overline{\Delta\Phi^2} + \mathcal{R}^{-2} v_n^2 \ , \tag{4.55}$$

where $\overline{\Delta\Phi^2}$ is the same as that shown in Eq. (4.53), $\mathcal{R}$ is the responsivity, and v_n is the rms Johnson noise voltage. Using Eq. (4.52), one obtains

$$(\text{NEP})^2 = 4kT_d^2 \mathcal{G} \Delta f + \mathcal{R}^{-2} 4kT_d R_d \Delta f \ . \tag{4.56}$$

This equation can be rearranged and Eqs. (4.7), (4.12), and (4.17) can be combined with Eq. (4.56) to produce the following expression (It is assumed that $\omega^2\tau^2 \ll 1$.):

$$(\text{NEP})^2 = 4kT_d^2 \Delta f \left\{ \mathcal{G} + \frac{[\mathcal{G} - \alpha\mathcal{G}_0(\Delta T_d)]^2}{T_d \mathcal{G}_0 (\Delta T_d) \varepsilon^2 \alpha^2} \right\} \ . \tag{4.57}$$

If $\varepsilon \approx 1$ and $\alpha(\Delta T_d) \ll 1$, then mean square fluctuation $(\text{NEP})^2$ reduces to

$$(\text{NEP})^2 = 4kT_d \Delta f \left[\frac{\mathcal{G}^2}{\mathcal{G}_0 (\Delta T_d) \alpha^2} \right] \ . \tag{4.58}$$

This expression shows that for a bolometer, $(\text{NEP})^2$ is not dependent on detector resistance R_d. It also points out that a bolometer can be optimized by choosing

materials with a higher α and a low thermal conductance $\mathcal{G}$. Since the thermal time constant is expressed as $\tau_T = \mathcal{C}/\mathcal{G}$, reducing the value of $\mathcal{G}$ leads to a longer time constant unless the thermal capacitance $\mathcal{C}$ is reduced to the same extent.

For the case of the thermopile, the minimum detectable power NEP can be obtained using Eqs. (4.56) and (4.44):

$$(\text{NEP})^2 = 4kT_d^2\Delta f\left[\mathcal{G} + \frac{R_d(1 + \omega^2\tau_T^2)\mathcal{G}^2}{\varepsilon^2 P_{ab}^2 T_d}\right] \tag{4.59}$$

and, since $\tau_T = \mathcal{C}/\mathcal{G}$,

$$(\text{NEP})^2 = 4kT_d^2\Delta f\left[\mathcal{G} + \frac{R_d(\mathcal{G}^2 + \omega^2\mathcal{C}^2)}{\varepsilon^2 P_{ab}^2 T_d}\right] . \tag{4.60}$$

In this equation the thermal conductance $\mathcal{G}$ and the thermal capacitance $\mathcal{C}$ are made up of contributions from the following: the responsive element, $\mathcal{G}_R$ and $\mathcal{C}_R$; the gas used to fill the chamber containing the responsive element, $\mathcal{G}_g$ and $\mathcal{C}_g$; and the electrical leads from the responsive element to the connector pins, $\mathcal{G}_c$ and $\mathcal{C}_c$. Thus,

$$\mathcal{G} = \mathcal{G}_R + \mathcal{G}_g + \mathcal{G}_c , \tag{4.61}$$

$$\mathcal{C} = \mathcal{C}_R + \mathcal{C}_g + \mathcal{C}_c . \tag{4.62}$$

If the responsive element of a thermopile detector is located in a housing that has been evacuated, then the contributions resulting from the ambient gas are zero (i.e., $\mathcal{G}_g = 0$ and $\mathcal{C}_g = 0$). This reduces the NEP somewhat; but since $\tau_T = \mathcal{C}/\mathcal{G}$, a decrease in $\mathcal{G}$ also increases the thermal time constant. The most fundamental method of increasing the performance of a thermopile is to use materials with higher thermoelectric powers, P_{ab}.

The ultimate limit of a pyroelectric detector is given by Eq. (4.53). The other source of noise contributed by the responsive element is Johnson noise. Therefore, Eq. (4.55) applies for the pyroelectric detector as well as in cases where responsivity $\mathcal{R}$ is given by Eq. (4.50). Assuming that $\omega^2\tau_e^2 \ll 1$, $\omega^2\tau_e^2 > 1$, and that $\varepsilon \approx 1$, we obtain the following expression for the mean square noise:

$$(\text{NEP})^2 = 4kT_d^2\mathcal{G}\Delta f + 4kT_dR_e\Delta f\mathcal{R}^{-2} . \tag{4.63}$$

Therefore,

$$(\text{NEP})^2 = 4kT_d\Delta f\left(T_d\mathcal{G} + \frac{\mathcal{G}^2\tau_T^2}{\mathcal{P}^2A_d^2R_e}\right) , \tag{4.64}$$

where $\mathcal{P}$ is the pyroelectric coefficient and the other symbols are as previously defined. Since $\tau_T = \mathcal{C}/\mathcal{G}$, Eq. (4.64) can be rewritten as

$$(\text{NEP})^2 = 4kT_d\Delta f\left(T_d\mathcal{G} + \frac{\mathcal{C}^2}{\mathcal{P}^2A_d^2R_e}\right) . \tag{4.65}$$

This means that the mean square noise can be reduced by decreasing the area A_d of the detector and by choosing a material with a high value of pyroelectric coefficient $\mathcal{P}$ and a low thermal capacitance $\mathcal{C}$. [The terms $\mathcal{C}$ and $\mathcal{G}$ are both area-dependent terms as analyzed by Putley.[4] Therefore, the area dependency cancels in the second term in brackets on the right side of Eq. (4.65) and remains only in the first term. This analysis ignores noise contributions from other parts of the infrared system in which these detectors are used, such as in postdetector electronics.]

In 1946, Havens[5] developed an empirical relationship between the signal power generated by a thermal detector and the change in incident optical power.

$$\frac{V_s^2}{R_d} = N\, d\Phi\, \frac{\Delta T}{T} , \tag{4.66}$$

where

V_s = signal voltage due to the change in optical power
R_d = resistance of the detector element
$d\Phi$ = optical power change incident on the detector
ΔT = temperature change due to the change in incident optical power $d\Phi$
T = temperature of detector, K
N = empirically determined constant that is a function of thermal detection type and ranged from 1 to 100.

He further postulated that the theoretical minimum detectable optical energy would be approximately the same for all thermal detectors operating at room temperature:

$$d\Phi_\tau = 3 \times 10^{-12}\ \text{J} , \tag{4.67}$$

where τ is the duration of the pulse in seconds and the detector size is 1 mm^2.

Equating $d\Phi$ in this case with NEP we have:

$$\text{NEP} = \frac{3 \times 10^{-9}}{\tau}\ \text{W} \tag{4.68}$$

for τ measured in milliseconds. Hudson[6] has converted this limit to its equivalent D^* value and obtained the value

$$D^*_{\text{Havens limit}} = 5.38 \times 10^8\ (\tau)^{1/2} , \tag{4.69}$$

where τ is in milliseconds.

Contemporary room-temperature detectors exhibit performance characteristics that are close to (or slightly exceed) this empirically devised, and somewhat arbitrary, limit.

4.3 THEORETICAL DESCRIPTIONS OF PHOTON DETECTORS

In photon detectors, incident infrared photons are absorbed producing free charge carriers that change an electrical characteristic of the responsive element. This process is carried out without any significant temperature change in the responsive element. Brief theoretical descriptions of the four most commonly used processes are given here.

4.3.1 Photoconductive Effect

In this type of photon detector, incident infrared photons are absorbed, producing free charge carriers that change the electrical conductivity of the responsive element. This change in conductivity is detected in the associated electronic circuit (Fig. 4.6). If the conductivity of the responsive element (Fig. 4.7) increases because of absorbed infrared photons, then resistance R_d of the responsive element will decrease since

$$R_d = \frac{l}{\sigma_e A_c} , \tag{4.70}$$

where l is the length, A_c is the cross-sectional area wd, and σ_e is the electrical conductivity of the responsive element. This change in resistance produces a signal voltage, which is fed to a preamplifier. The signal voltage can be expressed as

$$v_s = I\Delta R_d + R_d \Delta I , \tag{4.71}$$

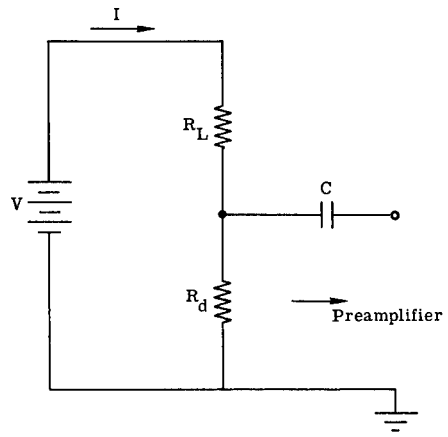


Fig. 4.6 Photoconductor detector circuit.

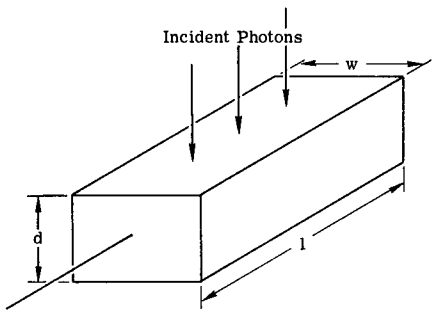


Fig. 4.7 Photoconductor geometry.

where I is the current through the circuit. If the circuit is operated in a constant current condition (i.e., $\Delta I = 0$ and $R_L \gg R_d$), then this equation reduces to

$$v_s = I\Delta R_d \frac{R_L}{(R_L + R_d)}$$

$$\approx I\Delta R_d \tag{4.72}$$

and further:

$$v_s = \frac{V\Delta R_d}{(R_L + R_d)} \tag{4.73}$$

$$= \frac{V_B R_d}{(R_L + R_d)} \frac{\Delta N}{N} , \tag{4.74}$$

where

R_d = $l/\sigma_e wd$
σ_e = the electrical conductivity; $\sigma_e = ne\mu$
l = length of detector
w = width of detector
d = depth of detector
n = the density of free charge carriers per unit volume
N = the total number of free charges in the responsive element when there are no incident, infrared signal photons; $N = nlwd$
ΔN = the increase in total number of free charges caused by the incident infrared signal photons
V_B = bias voltage.

According to Petritz,[7] the small detector signal properties are governed by the following two equations:

$$\frac{d}{dt}\Delta N = A_d \eta E_{q,s} - \frac{\Delta N}{\tau_c} , \tag{4.75}$$

$$\Delta N = N(t) - N , \tag{4.76}$$

where

η = the efficiency in converting incident infrared photons into free charge carriers
A_d = the detector area, wl
E_q = incident photon flux density, photons $\text{cm}^{-2}\,\text{s}^{-1}$
τ_c = the average free charge carrier lifetime.

The solution for a sinusoidal input rate $E_{q,s}$ at a modulation frequency f_c of $\omega_c/2\pi$, where $E_q = E_{q,o} + E_{q,s}\cos\omega_c t$, is

$$|\Delta N(f)| = \frac{A_d \eta E_{q,s} \tau_c}{(1 + \omega^2 \tau_c^2)^{1/2}} . \tag{4.77}$$

Therefore,

$$\frac{|\Delta N(f)|}{N} = \frac{\Delta N}{N} = \frac{A_d \eta E_{q,s} \tau_c}{N(1 + \omega^2 \tau_c^2)^{1/2}} \ . \tag{4.78}$$

The signal voltage expression for a photoconductor is obtained by substituting Eq. (4.78) into (4.74):

$$v_s = \frac{V_B R_d}{(R_L + R_d)} \frac{A_d \eta E_{q,s} \mathrm{T}_c}{N(1 + \omega^2 \tau_c^2)^{1/2}} \ . \tag{4.79}$$

Equation (4.79) is an expression for the magnitude of the detector output voltage for an input flux that has a sinusoidal waveform. The superposition theorem for linear systems permits use of Eq. (4.79) for other more complex waveforms, and for modulation at any frequency. One can also write this in terms of the spectra of the quantities involved:

$$\mathcal{F}\{v_s\} = \frac{VR_d}{R_L + R_d} \frac{A_d \eta \tau_c}{N(1 + i\omega\tau_c)} \mathcal{F}\{\Phi\} \ , \tag{4.80}$$

where $\mathcal{F}\{v_s\}$ is the Fourier transform of the voltage change, or the voltage spectrum, and $\mathcal{F}\{\Phi\}$ is the modulation spectrum of the signal flux. The complex voltage spectrum can be separated into its modulus and phase:

$$\mathcal{F}\{v_s\} = |\mathcal{F}\{v_s\}| + i \arg(\mathcal{F}\{v_s\}) \ , \tag{4.81}$$

$$\mathcal{F}\{|v_s|\} = \frac{VR_d}{R_L + R_d} \frac{A_d \eta \tau_c}{N(1 + \omega^2 \tau_c^2)^{1/2}} \ , \tag{4.82}$$

$$\arg(\mathcal{F}\{v_s\}) = \arctan(-\omega\tau_c) + \arg(\mathcal{F}\{\Phi\}) \ . \tag{4.83}$$

The modulus of the transfer function relating the output signal voltage spectrum to the input flux spectrum is given by

$$\left| \frac{\mathcal{F}\{v_s\}}{\mathcal{F}\{\Phi\}} \right| = \frac{VR_d}{R_L + R_d} \frac{A_d \eta \tau_c}{N(1 + \omega^2 \tau_c^2)^{1/2}} \ . \tag{4.84}$$

The phase shift is given by $\arctan(-\omega\tau_c)$.

The noise encountered in photoconductors and other photon detectors is described in Sec. 4.3.8. In that section, expressions for the SNR will be examined using Eq. (4.79) as the signal voltage for photoconductors.

4.3.2 Photovoltaic Effect

In the photovoltaic process, a *p-n* junction is formed in or on a semiconductor. Infrared photons, which are absorbed at or near the junction, are separated by the junction producing an external electrical current. The magnitude of this current is related to the number of incident infrared photons.

A simplified, energy-band picture of the photovoltaic process is shown in Fig. 4.8. Figure 4.8(a) shows the energy-band arrangement for an unirradiated p-n junction. Incident infrared photons with energy $h\nu$ greater than the energy-band gap will create electron-hole pairs in the semiconductor at or near the junction region. As shown in Fig. 4.8(b), if the photons are absorbed in the p region, the free electron will move down along the conduction band (c-band) to the n region. This means that the Fermi levels in the n and p regions will be displaced because of the presence of the free electron and hole, respectively. This shift in the Fermi level produces a voltage ΔV that can be observed with an external electronic circuit. When the electrical characteristics of the responsive element are observed with an external electronic circuit, the current-voltage relationship is that shown in Fig. 4.9, where curve (a) is the unilluminated case and (b) is the illuminated case. In the unilluminated case, no short-circuit current I_{sc} will be observed if the diode is externally shorted. Also, no open-circuit voltage V_0 will be observed. In practice, however, some background photons with sufficient energy to cause band-to-band transitions will be incident on the responsive element, producing free electron-hole pairs. Thus the unilluminated I-V curve really lies between curves (a) and (b). In curve (b), the open-circuit voltage V_0 and the short-circuit current I_{sc} are shown. The current through the junction is given by

$$I = I_{sa}\left[\exp\left(\frac{eV}{\beta kT_d}\right) - 1\right], \tag{4.85}$$

where β is an efficiency factor for photodiodes ($\beta = 1$ for an ideal diode and $\beta = 2$ to 3 for the real case) and I_{sa} is the saturation current, which is given by

$$I_{sa} = e\left(\frac{D_h p}{L_h} + \frac{D_e n}{L_e}\right), \tag{4.86}$$

where D_h and D_e are the hole and electron diffusion constants, respectively, in $\text{cm}^2\ \text{s}^{-1}$; and L_h and L_e are the hole and electron diffusion lengths, respectively, in centimeters. These are defined by the equations

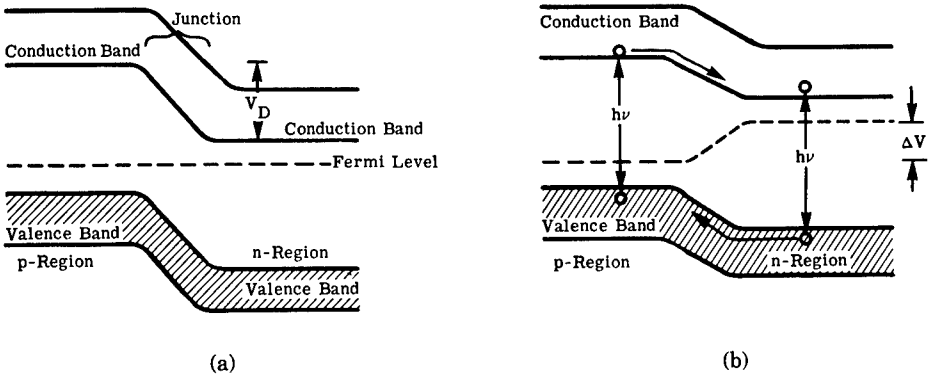


Fig. 4.8 Energy-band model of a p-n junction: (a) unilluminated and (b) illuminated.

$$L_h = (D_h\tau_h)^{1/2} = \left[\left(\frac{kT}{e}\right)\mu_h\tau_h\right]^{1/2} ,$$
$$D_h = \frac{kT}{e}\mu_h , \tag{4.87}$$

$$L_e = (D_e\tau_e)^{1/2} = \left[\left(\frac{kT}{e}\right)\mu_e\tau_e\right]^{1/2} ,$$
$$D_e = \frac{kT}{e}\mu_e , \tag{4.88}$$

where μ_h and μ_e are the hole and electron mobilities, respectively, and τ_h and τ_e are the hole and electron lifetimes, respectively.

However, Eq. (4.85) is not complete. As pointed out earlier in this section, when the photodiode is operated in the short-circuit condition (i.e., $V_B = 0$), the current is not zero as curve (a) in Fig. 4.9 implies. The existence of background photons produces some free charges that lead to a short-circuit current I_{sc}. Also, in practice, there is a leakage current in photodiodes that can be represented by the term $G_{sh}V$, where G_{sh} is the effective shunt conductance. Equation (4.85) can now be completed[7]:

$$I = I_{sa}\left[\exp\left(\frac{eV}{\beta kT_d}\right) - 1\right] - I_{sc} + G_{sh}V . \tag{4.89}$$

Since I_{sc} is due to background radiation, it can be expressed as

$$I_{sc} = e\eta E_{q,B}A_d \tag{4.90}$$

where

e = the electronic charge
η = the quantum efficiency
A_d = the sensitive area of the responsive element
$E_{q,B}$ = the photon flux density from the background,

which is given by

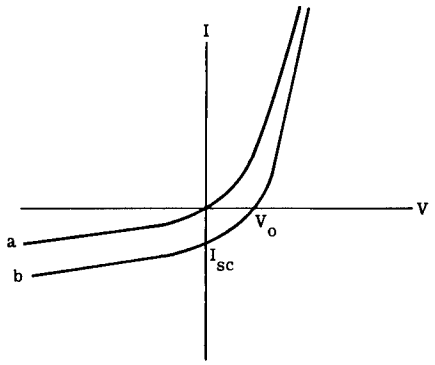
$$E_{q,B} = \int_0^{\lambda_c} E_{q,\lambda,B}(\lambda)\, d\lambda , \tag{4.91}$$

where λ_c is the long-wavelength limit of the detector's spectral response and $E_{q,\lambda,B}(\lambda)$ is the spectral photon flux density from the background.

Pruett and Petritz[7] have shown that for the small signal case, the signal current I_s can be derived from Eq. (4.89). The expression for I_s becomes

$$I_s = -e\eta A_d E_{q,s} , \tag{4.92}$$

where $E_{q,s}$ is the signal photon flux density.



a = unilluminated condition
b = illuminated condition

Fig. 4.9 Current voltage characteristics of a photodiode.

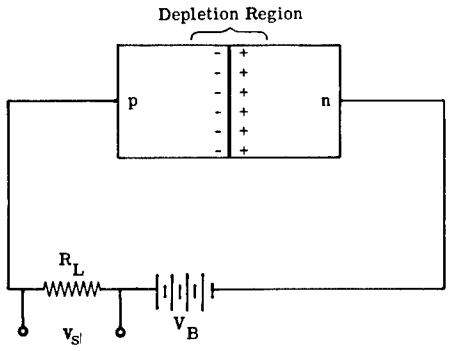


Fig. 4.10 Back bias in photovoltaic detector.

The best operating point (producing a maximum SNR) is at or near $V_B = 0$, which necessitates a back-bias configuration, shown in Fig. 4.10. The back bias also increases the width of the depletion region, thereby reducing the time constant, because the wider depletion region ensures a higher probability that signal photons will be absorbed in the vicinity of the depletion region. This cuts down the time it takes free charge carriers to move from the site of absorption to the depletion region.

If the back-bias voltage V_B is raised to a large enough value, the free electrons and holes moving in the field will be accelerated and, thus, acquire sufficient energy to produce additional free charge carriers on collision with the lattice. As V_B is increased further, the multiplication of free charge carriers increases. The limit to this is a breakdown condition where the electric field produced by the applied bias is causing free charge carriers to be generated. The multiplication factor M, defined as the average number of electron-hole pairs produced from a single initiating electron-hole pair, is related to the bias voltage V_B and the breakdown voltage V_{bd} as shown:

$$M^{-1} \propto \left[1 - \left(\frac{V_B}{V_{bd}}\right)^m\right] . \tag{4.93}$$

The values of the empirical constant m have been observed to be between 1.4 and 4. Avalanche photodiodes made from silicon with M values up to 10^6 have been reported by Haitz et al.[8] The signal current in an avalanche diode can be expressed as

$$I_s = -e\eta A_d E_{q,s} M . \tag{4.94}$$

However, the mean square noise current, as calculated by McIntyre,[9] is given by

$$i_n^2 = 2eIM^3\Delta f , \tag{4.95}$$

where I is the current from Eq. (4.89). The signal-to-noise current ratio is then proportional to $1/\sqrt{M}$. This means that although the signal current rises with M, the noise rises faster. Because of this, avalanche photodiodes are very useful in situations where amplifier noise predominates. In these cases, an increase in noise caused by M would not significantly increase the system noise but would substantially increase the signal level.

4.3.3 Photoelectromagnetic Effect

The photoelectromagnetic (PEM) effect is not often used in modern infrared photon detectors. In this effect, incident infrared photons are absorbed at or near the surface of a semiconductor, producing free electron-hole pairs. These pairs diffuse from the surface and down into the crystal. The presence of a magnetic field B directed along the y axis causes the electrons to separate from the holes as they diffuse away from the surface (see Fig. 4.11). This separation of charge produces an electrical voltage across the terminals. If the external circuit is shorted, a current will continue to flow as long as infrared photons are arriving at the surface. If the exterior circuit is left open, as shown in the figure, a voltage will appear across the open terminals and remain there as long as the surface is irradiated. Moss et al.[10] have derived the open-circuit voltage V_0 and the short-circuit current I_{sc} per unit length for PEM detectors. These expressions are

$$I_{sc} = \frac{eE_{q,s}\theta L}{(1 + \gamma)} , \tag{4.96}$$

$$V_0 = \frac{eE_{q,s}\theta Ll}{(1 + \gamma)\sigma d} , \tag{4.97}$$

where

E_{qs} = photon flux density for signal radiation, photons s^{-1} cm^{-2}
L = ambipolar diffusion length, either L_e or L_h, whichever is the majority carrier

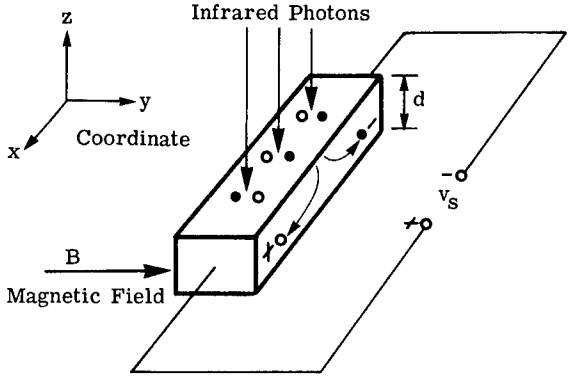


Fig. 4.11 Schematic of the PEM effect.

γ = $\tau s/L$
l = the electrode separation
σ = electrical conductivity of the crystal
d = thickness of the responsive element
e = electronic charge
τ = time constant
s = surface recombination velocity
θ = $\theta_e + \theta_h$
θ_e = $B\mu_e$, Hall angle for electrons
θ_h = $B\mu_h$, Hall angle for holes.

These expressions are valid under the following conditions:

1. Magnetic field B is small enough that $\mu^2B^2 \ll 1$.
2. The specimen is thick enough that $d/2L_d \approx 1$, where L_d is the effective diffusion length:

$$L_d = \left[\frac{bD_h\tau(n + p)}{bn + p + (bp + n)\theta_h\theta_e}\right]^{1/2}, \tag{4.98}$$

   where $b = \mu_e/\mu_h$, the ratio of electron to hole mobility; n is the electron volume density; and p is the hole volume density.
3. The crystal has high enough conductivity to assume charge neutrality throughout the responsive element.
4. The surface recombination velocity is small enough to allow a large fraction of free charge carriers to diffuse from the generation site near the surface down into the crystal bulk.[10]

4.3.4 Photoemissive Effect

In the photoemissive effect, an incident photon is absorbed by the sensitive surface. It gives up all its energy, hc/λ, to a free electron at or near the surface of the sensitive material. Thus, the kinetic energy of the electron is increased by an amount equal to the photon energy. Before the electron can escape from the surface, it may give up part of its kinetic energy to atoms through collisions. The amount of energy lost in this manner varies considerably. If the electron still has enough kinetic energy by the time it arrives at the surface, it can escape from the surface. The maximum kinetic energy of the escaped electron can be expressed as follows:

$$E_k = \frac{hc}{\lambda} - e\phi\ , \tag{4.99}$$

where

E_k = the kinetic energy of the freed electron
hc/λ = the energy of the absorbed photon
ϕ = the surface work function of the material
e = the charge of the electron.

If such a sensitive surface is placed in an evacuated chamber along with an anode and attached to an exterior circuit, as shown in Fig. 4.12, the electrons

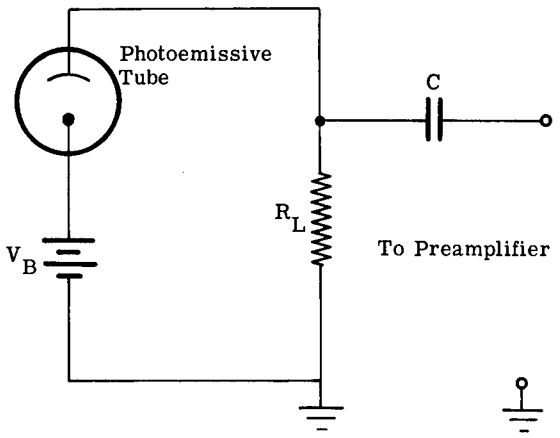


Fig. 4.12 Photoemissive bias circuit.

freed from the cathode surface by absorbed photons will be attracted to the anode. A current will then flow in the circuit through R_L as long as photons of sufficient energy arrive at the sensitive surface.

The time constant is determined by the spread in transit times of the electrons between the cathode and the anode. This can be as small as 10^{-9} s.

In addition to the current induced by infrared photons, a small current flows even when the cell is in the dark. This dark current sets the limit to the minimum detectable radiation. This current is primarily a result of thermionic emission, and its level depends on cathode temperature, sensitive area, and the work function.

The lowest work functions that have been achieved to date are approximately 1 eV. This means that the photoemissive detectors are sensitive in the very short wavelength region of the infrared spectrum. Spectral responses as long as 1.5 μm have been achieved.

A considerable gain in responsivity can be produced by adding a small amount of inert gas to the detector chamber. Electrons emitted from the cathode may then be accelerated by the field toward the anode. Before reaching the anode, an electron may collide with a gas atom and ionize it. With this technique, the number of electrons that reach the anode could be 100 times the number of electrons photoemitted from the cathode. The presence of positive ions in the cell lengthens the transit time for the electrons, which increases the dispersion and thus increases the time constant.

Electrons are emitted from a material surface when it is bombarded with high-velocity electrons. This process of secondary emission has been used to develop a photocell having high internal amplification. Such cells are called *photomultipliers*. Photoemitted electrons are focused onto another electrode where each electron produces a number of secondary electrons. These, in turn, are focused onto a third electrode, and the process is repeated several times. The photomultiplier has advantages over both the simpler photocells. The current can be multiplied by about 10^6, compared with a gain of about 10^2 for the gas-filled photocell. Since no positive ions are involved in the photomulti-

plier photocells, the time constants are shorter than for the gas photocell, but longer than for the simple vacuum photocell.

The responsivity $\mathcal{R}$ of the photoemissive detector can be derived as follows. The effective spectral, signal, photon flux density, $\Phi_{q,\lambda,s}(\lambda)$, on the sensitive cathode is given by

$$\Phi_{q,\lambda,s}(\lambda) = \{[\Phi_{s,\lambda}(\lambda)\eta]\}\left(\frac{\lambda}{hc}\right) , \tag{4.100}$$

where $\Phi_{s,\lambda}(\lambda)$ is the spectral signal power and $\eta(\lambda)$ is the quantum efficiency. If one assumes that all the electrons will be collected at the anode, the photocurrent i_s is expressed as

$$i_s(\lambda) = e\eta(\lambda)\Phi_{s,\lambda}(\lambda)\frac{\lambda}{hc} . \tag{4.101}$$

The signal voltage v_s will be the change in voltage across the load resistor R_L (see Fig. 4.12):

$$v_s(\lambda) = i_s(\lambda)R_L = R_L e\eta(\lambda)\Phi_{s,\lambda}(\lambda)\frac{\lambda}{hc} . \tag{4.102}$$

Since the responsivity $\mathcal{R}$ is defined as the signal voltage per signal watt input, $\mathcal{R}$ becomes

$$\mathcal{R} = \frac{v_s(\lambda)}{\Phi_{s,\lambda}(\lambda)} = R_L e\eta(\lambda)\frac{\lambda}{hc} . \tag{4.103}$$

The total current is given as

$$i_s = \frac{e}{hc}\int_0^\infty \eta(\lambda)\Phi_{s,\lambda}(\lambda)\lambda \, d\lambda . \tag{4.104}$$

The responsivity is then given by

$$\mathcal{R} = R_L \frac{\dfrac{e}{hc}\displaystyle\int_0^\infty \eta(\lambda)\Phi_{s,\lambda}(\lambda)\lambda \, d\lambda}{\displaystyle\int_0^\infty \Phi_{s,\lambda}(\lambda) \, d\lambda} . \tag{4.105}$$

For photomultipliers, the responsivity equation should be multiplied by the gain. The limitation to sensitivity will be set by fluctuations in the dark current manifested by changes in voltage across the load R_L.

4.3.5 Quantum Well Detectors

The development of advanced artificially structured material fabrication techniques, such as molecular beam epitaxy (MBE) and photon-assisted chemical

beam epitaxy (CBE), allows molecular layer-by-layer customization of semiconductors for detector applications. Two applications will be introduced here: tailoring the detector response by fabricating a desired bound state energy level and achieving true deterministic ("noiseless") solid-state photomultiplication.[11–13] These applications allow us to consider the possibility of custom fabrication, for example, of detector arrays with peak response at any desired wavelength, and with an individual solid-state photomultiplier connected to each detector. Such an array is analogous to a large number of photomultiplier tubes, but with higher quantum efficiency, the ability to operate at longer wavelengths, and smaller size and lower power requirements.

Both of the applications mentioned are made possible by the relationship between quantum well width and allowed energy levels. In Fig. 4.13(b), notice

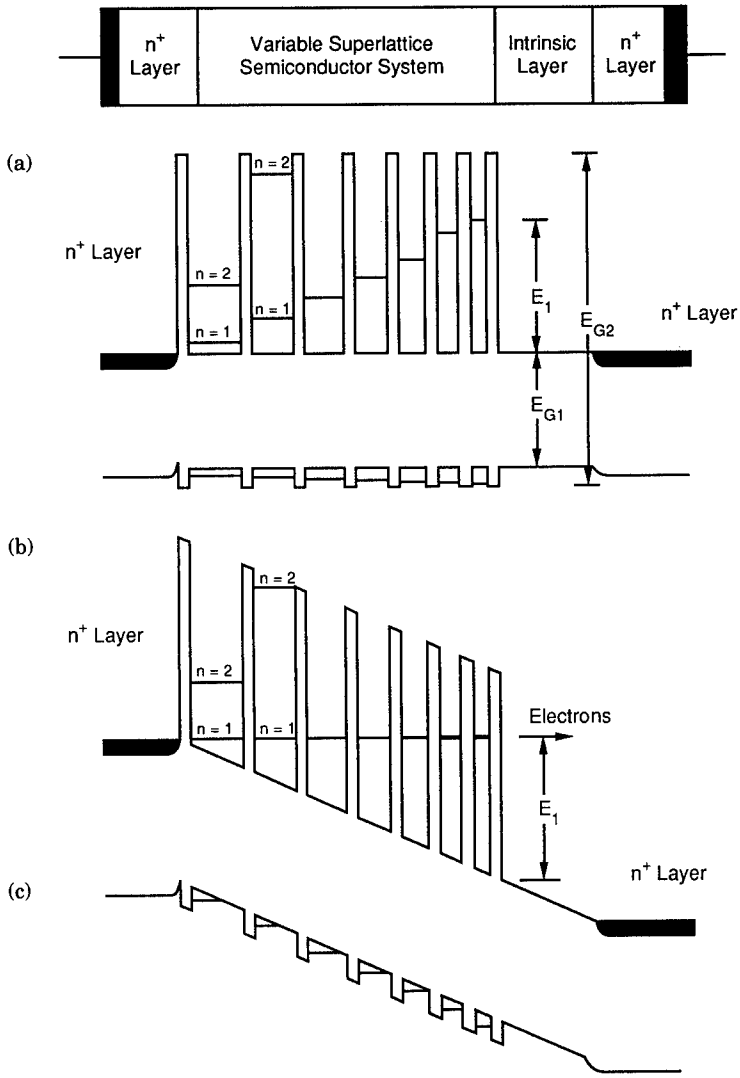


Fig. 4.13 Variably spaced superlattice energy filter (VSSEF): (a) map of device layers, (b) effect of quantum well width on energy levels, and (c) energy diagram with aligning bias applied.[11]

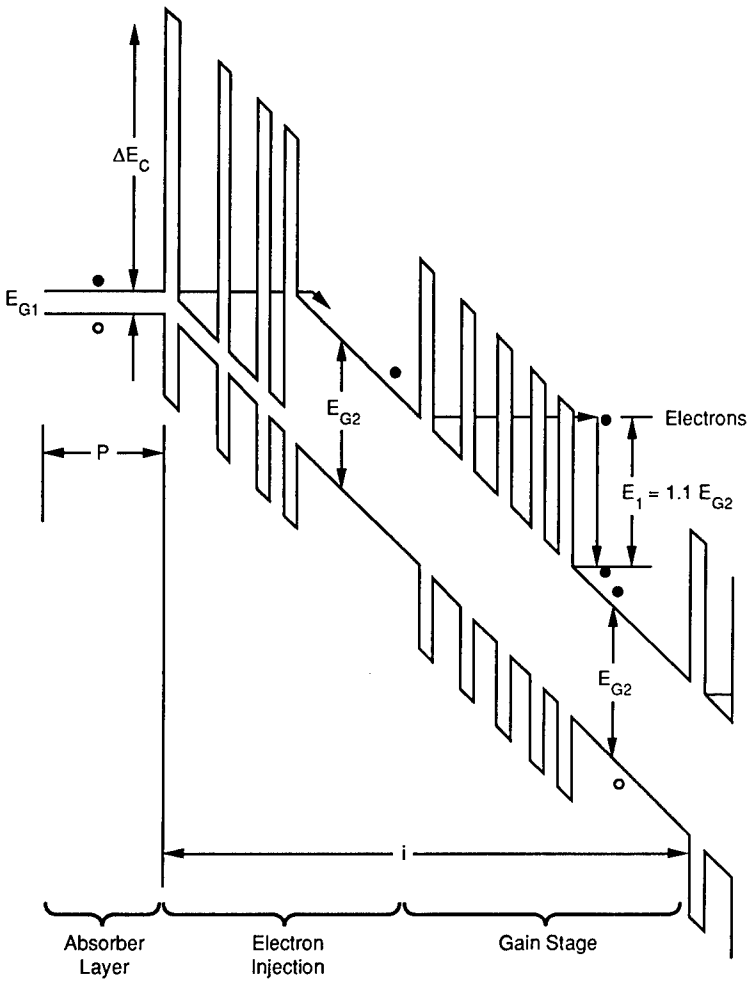


Fig. 4.14 Gain stage in a VSSEF solid-state photomultiplier.[12]

how the $n = 1$ energy level is successively higher in each of the sequences of more closely spaced material layers. Quantum wells, for this purpose, are of the order of 10 molecular layers wide. Since the width of the wells can be controlled to within one molecular layer, it is possible to make a material with essentially any desired energy gap for use as a detector at any desired wavelength. For example, even though GaAs has a 1.35-eV bandgap and, hence, cannot ordinarily be used as a detector at wavelengths longer than 0.92 μm, quantum well detectors fabricated of GaAs and AlGaAs layers have performed[14] well at 10 μm. Figure 4.13(c) indicates how this particular layer design can have all the $n = 1$ energy levels aligned by applying a bias voltage. Now the material is an energy filter passing an electron tunneling current with a specific electron energy E_1 present in the stage output current. In Fig. 4.14, a portion of a variably spaced superlattice energy filter (VSSEF) photomultiplier is shown. The key to deterministic photomultiplication is that a monoenergetic electron current emerges from each VSSEF stage, and that energy can be tailored to

excite an integer number of electrons in the successive stage. This offers several advantages over avalanche photodiodes (which sometimes are misleadingly termed "solid-state photomultipliers"). Avalanche photodiodes exhibit very noisy gain in the sense that the output signal per photon varies widely. This forces operation in a photon-counting mode for high gains. While satisfactory for single detectors at low light levels, photon-counting operation is not suitable for detector arrays, with their long integration times per frame, or for higher photon arrival rates. At present, however, photon counting can be done in the infrared with available avalanche detectors, while quantum well devices must be custom made by MBE.

4.3.6 Regenerative Detectors

A recently demonstrated detection mechanism,[15] which has exceeded[16] a dynamic range of 70 dB, is predicted to have a dynamic range of up to 100 dB with a very narrow-band spectral response and a noise equivalent power (NEP) as low as 10^{-16} W. This detector measures the turn-on time of a laser that is below oscillation threshold before an external signal is applied. Figure 4.15 shows that a strong signal causes the laser intensity to build up sooner than a weak signal. Figure 4.16 shows the agreement between model predictions and experimental results, while Fig. 4.17 illustrates the narrow measured detector bandwidth. By fixing the end time of each pulse with a Q switch, the laser pulse length becomes a measure of incident signal intensity. The laser pulse duration should be minimized in the absence of signal, and the pulse repetition rate should be maximized consistent with the desired dynamic range in order to maximize the detection duty cycle. When this method is compared to heterodyne detection, it has the advantages of accepting all polarizations and of being very much less critical in alignment. A disadvantage is that it has a detection duty cycle of less than one, but perhaps greater than one-half. These trade-offs could favor the super-regenerative laser receiver over a heterodyne detector for specialized applications.

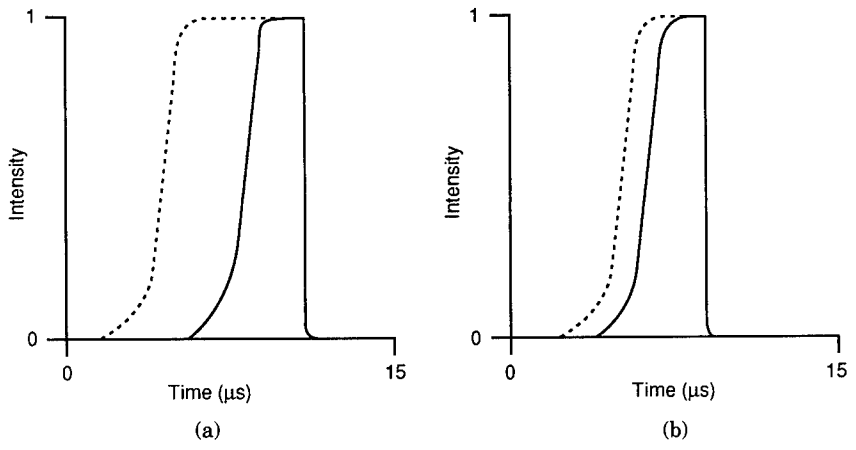


Fig. 4.15 Dependence of laser turn-on time on external signal strength: (a) strong signal and (b) weak signal.

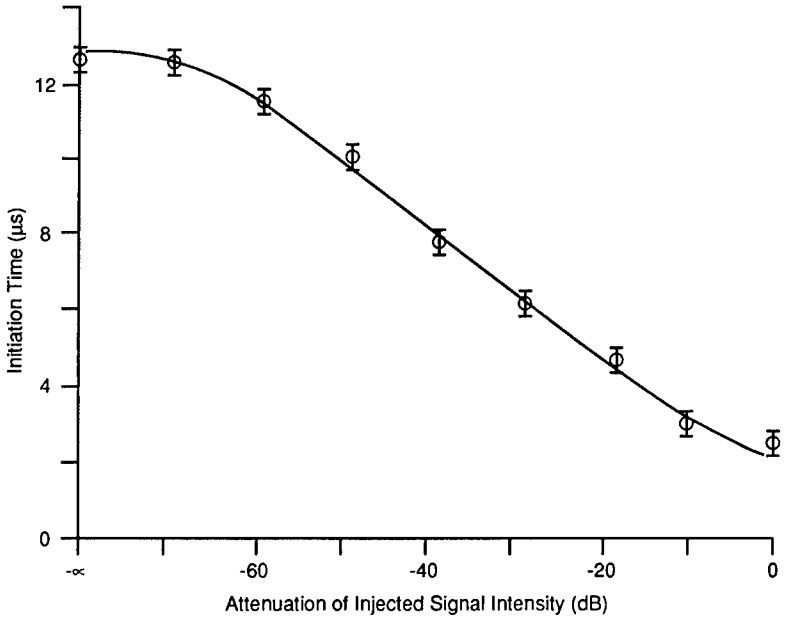


Fig. 4.16 Agreement between theory and experiment.[17]

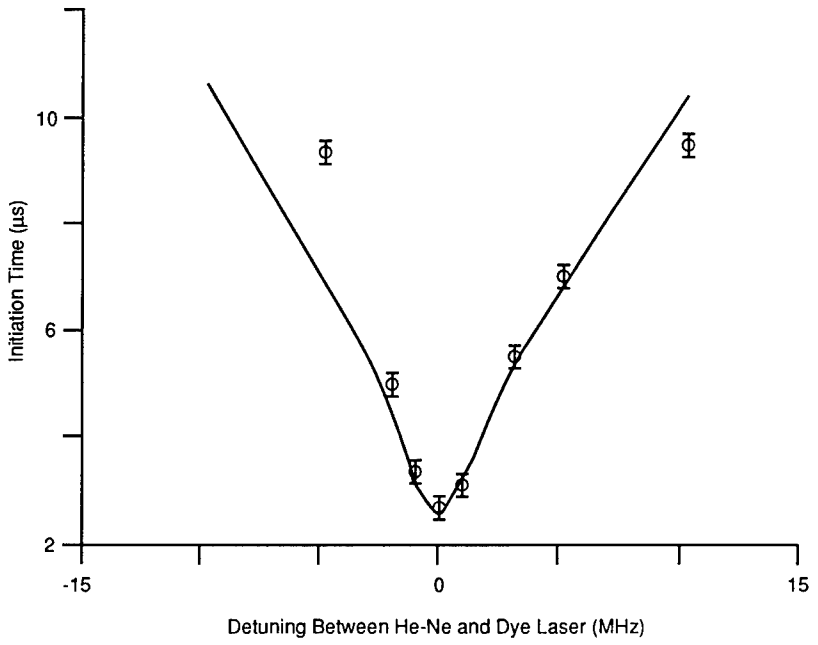


Fig. 4.17 Narrow bandwidth of detector response.[17]

The unique characteristics of this detector are its narrow spectral response (a matched filter to the laser linewidth) and its large dynamic range. Most detectors have less than 70 dB of linear dynamic range in incident optical power, and most detector arrays have less than 40 dB of dynamic range due to charge transfer device limitations. Arrays of optical regenerative receivers could be fabricated as arrays of laser diodes or as bundles of titanium sapphire fibers for applications such as extreme dynamic range optical computing and active imaging. Potential single-detector applications include free-space optical communications.

4.3.7 Coherent Heterodyne Detection

Applications for heterodyne detection primarily exploit two features of the technique: fine wavelength resolution at the electrical baseband and the ability to achieve shot-noise-limited performance with high quantum efficiency, but rather noisy, room temperature detectors. Once at electrical baseband, electronic spectrum analyzers with resolutions of perhaps a small fraction of 1 Hz can be used, compared to optical interference filters, for example, with 0.1-Å (10^9-Hz) passbands. A room temperature detector with an NEP of 10^{-10} W $Hz^{-1/2}$ may be used in a heterodyne receiver system that has an NEP below 10^{-16} W $Hz^{-1/2}$.

The heterodyne receiver of Fig. 4.18 will be analyzed as two optical sources, the signal and the local oscillator, incident on a square-law detector with output current proportional to optical intensity, or the square of the incident amplitude:

$$\phi_d = |A_{\text{in}}|^2 = (A_{\text{sig}} \cos\omega_{\text{sig}} t + A_{\text{LO}} \cos\omega_{\text{LO}} t)^2 \ . \tag{4.106}$$

Equation (4.106) may be written

$$\begin{aligned}\phi_d = {} & A_{\text{sig}}^2\left(\frac{\cos 2\omega_{\text{sig}} + 1}{2}\right) + A_{\text{LO}}^2\left(\frac{\cos 2\omega_{\text{LO}} t + 1}{2}\right) \\ & + A_{\text{sig}}A_{\text{LO}} \cos(\omega_{\text{sig}} + \omega_{\text{LO}})t \\ & + A_{\text{sig}}A_{\text{LO}} \cos(\omega_{\text{sig}} - \omega_{\text{LO}})t \ .\end{aligned} \tag{4.107}$$

This expression contains three terms with very high optical frequencies, which are well above the electrical frequency response of the detector. The first three

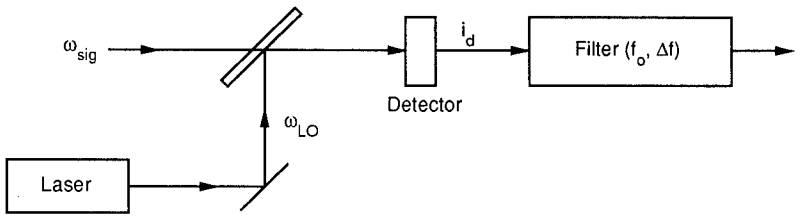


Fig. 4.18 Heterodyne optical receiver.

terms will contribute to the detector dc output only, with the fourth (difference frequency) term contributing the only ac detector output. Therefore, electrical filtering of the detector output will select a signal frequency corresponding to the difference in signal and local oscillator frequencies:

$$i_d = i_{dc} + i_{ac} = \frac{\eta q \lambda}{hc}\left[\frac{A_{sig}^2 + A_{LO}^2}{2} + A_{sig}A_{LO}\right] + \frac{\eta q \lambda}{hc}[A_{sig}A_{LO}\cos(\omega_{sig} - \omega_{LO})t] \ . \tag{4.108}$$

For a shot-noise-limited detector, the SNR is

$$\text{SNR} = \frac{i_{ac}}{(2qi^{dc}\Delta f)^{1/2}} \ . \tag{4.109}$$

Combining Eqs. (4.108) and (4.109), we have

$$\text{SNR} = \left(\frac{\eta\lambda}{hc\Delta f}\right)^{1/2} \frac{A_{sig}A_{LO}}{A_{sig} + A_{LO}} \cos(\omega_{sig} - \omega_{LO})t \ . \tag{4.110}$$

Heterodyne receivers are usually operated with $A_{LO} \gg A_{sig}$, which simplifies our expression to

$$\text{SNR} = \left(\frac{\eta\lambda}{hc\Delta f}\right)^{1/2} A_{sig}\cos(\omega_{sig} - \omega_{LO})t \ . \tag{4.111}$$

Equation (4.111) is a shot-noise-limited expression in the desired signal, yet the detector need only be shot noise limited by the much stronger local oscillator signal for Eq. (4.111) to hold.

4.3.8 Theoretical Limit of Performance of Photon Detectors

In general, the limit of performance of infrared photon detectors is set by the fluctuation in arrival rate of background photons. This fluctuation, called *photon noise,* appears as a random voltage or current at the output terminals of the detector. The total noise from the detector is the rms sum of the inherent noise generated within the detector plus the photon noise.

This section deals with the various noise mechanisms encountered in the use of infrared detectors, including photon noise, lattice G-R noise, $1/f$ noise, Johnson noise, and shot noise.

Photon Noise. The mean square fluctuations in the number of photons arriving at a surface in a small wavelength interval $\Delta\lambda$ are given by

$$\overline{[\Delta\phi_{q,\lambda}(\lambda)\Delta\lambda]^2} = \Phi_{q,\lambda}(\lambda)\Delta\lambda[1 + \gamma(e^x - 1)^{-1}] \ , \tag{4.112}$$

where $\Phi_{q,\lambda}(\lambda)$ is the photon flux per unit wavelength, γ is a coherence factor, and $x = c_2/\lambda T$, $c_2 = hc/k$. The quantity $\Phi_{q,\lambda}(\lambda)\Delta\lambda$ can be expressed as the sum

of a signal flux and a background flux. The mean square fluctuation in total photon flux on the detector is

$$\overline{[\Phi_{q,\lambda}(\lambda)\Delta\lambda]^2} = \overline{[\Phi_{q,\lambda,s}(d)\Delta\lambda]^2} + \overline{[\Phi_{q,\lambda,B}(\lambda)\Delta\lambda]^2} \ , \tag{4.113}$$

where $\Phi_{q,\lambda,s}(\lambda)\Delta\lambda$ is the photon flux from the signal and $\Phi_{q,\lambda,B}(\lambda)\Delta\lambda$ is the photon flux from the background. The total photon flux incident on the detector is given by

$$\Phi_q = \int_0^\infty [\Phi_{q,\lambda,s}(\lambda) + \Phi_{q,\lambda,B}(\lambda)]\, d\lambda \ . \tag{4.114}$$

A monochromatic signal flux generates free charge carriers at a rate G_{gen} given by

$$(G_{\text{gen}})_s = \eta(\lambda)\frac{\lambda}{hc}\Phi_{s,\lambda}(\lambda)\Delta\lambda \ . \tag{4.115}$$

The variance (noise squared) is thus

$$\overline{(\Delta G_{\text{gen}})_s^2} = 2\eta(\lambda)E_{q,\lambda,s}(\lambda)\Delta\lambda A_d\Delta f \ . \tag{4.116}$$

The noise from the background is given by

$$\overline{(\Delta G_{\text{gen}})_B^2} = 2\eta(\lambda)E_{q,\lambda,B}(\lambda)\Delta\lambda A_d\Delta f \ . \tag{4.117}$$

The total mean square noise from the two sources is given by the sum of these two generation rates:

$$\overline{(\Delta G_{\text{gen}})_t^2} = 2\eta(\lambda)A_d\Delta f[E_{q,\lambda,s}(\lambda) + E_{q,\lambda,B}(\lambda)]\Delta\lambda \ . \tag{4.118}$$

The NEP is the flux (power) divided by the SNR:

$$\text{NEP} = \Phi_s/\text{SNR} \tag{4.119}$$

$$= \Phi_s[\overline{(\Delta G_{\text{gen}})_t^2}]^{1/2} \times (G_{\text{gen}})_s^{-1}$$

$$= \{2\eta(\lambda)A_d\Delta f[E_{q,\lambda,s}(\lambda) + E_{q,\lambda,B}(\lambda)]\Delta\lambda\}^{1/2}\left[\eta(\lambda)\frac{\lambda}{hc}\right]^{-1} \tag{4.120}$$

The specific detectivity D^* is given by

$$D^*(\lambda) = \frac{(A_d\Delta f)^{1/2}}{\text{NEP}} \ . \tag{4.121}$$

Therefore,

$$D^*(\lambda) = \eta(\lambda)\frac{\lambda}{hc}\{2\eta(\lambda)[E_{q,\lambda,s}(\lambda) + E_{q,\lambda,B}(\lambda)]\Delta\lambda\}^{-1/2} . \tag{4.122}$$

If the assumption of monochromatic signal and noise is relaxed,

$$D^* = \frac{\int_0^\infty \eta(\lambda)\frac{\lambda}{hc}\Phi_{s,\lambda}(\lambda)\, d\lambda}{\left[\int_0^\infty \Phi_{s,\lambda}(\lambda)\, d\lambda\right]\left\{\int_0^\infty 2\eta(\lambda)[E_{q,\lambda,s}(\lambda) + E_{q,\lambda,B}(\lambda)]\, d\lambda\right\}^{1/2}} . \tag{4.123}$$

Wolfe[18] has defined D_n^*, the response of detectors to photons rather than power, as

$$D_n^* = \eta(\lambda)\left\{\int_0^\infty 2\eta(\lambda)[E_{q,\lambda,s}(\lambda) + E_{q,\lambda,B}(\lambda)]\, d\lambda\right\}^{-1/2} . \tag{4.124}$$

Several cases can be conveniently evaluated: noise due to signal alone, noise due to background alone, coherent detection, and incoherent detection.

If the background can be neglected (which is almost never the case) and the signal is monochromatic, then

$$D^*(\lambda) = \eta(\lambda)\frac{\lambda}{hc}[2\eta(\lambda)E_{q,\lambda,s}\Delta\lambda]^{-1/2} = \left[\frac{\eta(\lambda)}{2}\right]^{1/2}\frac{\lambda}{hc}E_{q,\lambda,s}^{-1/2}\Delta\lambda , \tag{4.125}$$

$$D_n^* = \left[\frac{\eta(\lambda)}{2}\right]^{1/2}E_{q,\lambda,s}^{-1/2}\Delta\lambda = [2\eta(\lambda)E_{q,\lambda,s}\Delta\lambda]^{-1/2} . \tag{4.126}$$

The noise is contained in the (assumed) narrow band of the signal and can be evaluated as

$$E_{q,\lambda,s} \propto \frac{e^x - 1 + \gamma}{(e^x - 1)^2} , \tag{4.127}$$

where $x = c_2/\lambda T$, and $c_2 = hc/k$. If the noise in the signal can be neglected (almost always the case) and the signal is monochromatic then

$$D^*(\lambda_s) = \frac{\lambda_s}{hc}\eta(\lambda_s)[2\eta(\lambda_s)E_{q,\lambda,B}\Delta\lambda]^{-1/2} . \tag{4.128}$$

While the signal carriers are excited only in a narrow spectral band, the background contribution is generally over a wide spectral band, so that

$$D^*(\lambda_s) = \frac{\lambda_s}{hc}\eta(\lambda_s)\left[2\int_0^\infty \eta(\lambda)E_{q,\lambda,B}(\lambda)\, d\lambda\right]^{-1/2} . \tag{4.129}$$

For many detectors $\eta(\lambda)$ is essentially constant, in which case

$$D^*(\lambda_s) = \frac{\lambda_s\sqrt{\eta}}{hc}\left[2\int_0^{\lambda_c} E_{q,\lambda,B}(\lambda)\, d\lambda\right]^{-1/2}, \tag{4.130}$$

where λ_c is the cutoff wavelength of the detector:

$$D^*_n = \sqrt{\eta}\left[2\int_0^{\lambda_c} E_{q,\lambda,B}(\lambda)\, d\lambda\right]^{-1/2}. \tag{4.131}$$

The effective background photon flux is also often referred[9] to as Q_B, where:

$$Q_B \equiv \int\int_0^{\lambda_c} E_{q,\lambda,B}(\lambda)\, d\lambda \tag{4.132}$$

and

$$D^*(\lambda_c) = \frac{\lambda_c}{\sqrt{2}hc}\left(\frac{\eta}{Q_B}\right)^{1/2}. \tag{4.133}$$

If λ_c is expressed in micrometers, c in cm s^{-1}, h in W s^2, and $\eta = 1$, then:

$$D^*(\lambda_c) = 3.56 \times 10^{18}\, \lambda_c\left(\frac{1}{Q_B}\right)^{1/2}. \tag{4.134}$$

For a photoconductor, the recombination noise introduces a factor of $\sqrt{2}$ into the denominator of Eq. (4.134), which, therefore, reduces the value of the BLIP D^*. Therefore,

$$D^*_{pc}(\lambda_c) = 2.52 \times 10^{18}\, \lambda_c\left(\frac{1}{Q_B}\right)^{1/2}. \tag{4.135}$$

The background flux can be reduced by the use of a cold spectral filter and by a cold aperture stop that limits the field of view. The $D^*(\lambda)$ value for the filter spectral band is given by

$$D^*(\lambda_s) = \frac{\lambda_s\sqrt{\eta}}{hc}\left[2\int_{\lambda_1}^{\lambda_2} E_{q,\lambda,B}(\lambda)\, d\lambda\right]^{-1/2}, \tag{4.136}$$

$$D^*_n = \sqrt{\eta}\left[2\int_{\lambda_1}^{\lambda_2} E_{q,\lambda,B}(\lambda)\, d\lambda\right]^{-1/2}. \tag{4.137}$$

The reduction of the background flux by a cooled aperture is a geometrical consideration involving the ratio of the square root of the projected solid angle

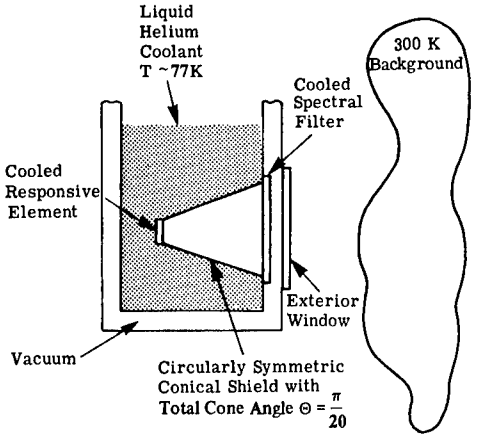


Fig. 4.19 Schematic of spectrally filtered cooled detector.

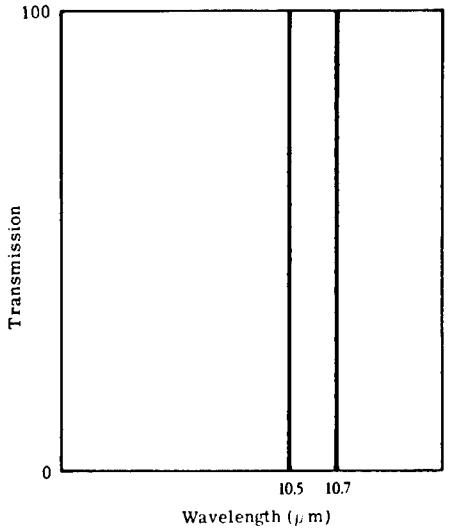


Fig. 4.20 Transmission characteristics of an ideal 10.5- to 10.7-μm spectral filter.

of view to that of a hemisphere. For a conical field of view, $D^*(\lambda)$ can be multiplied by $(\sin\Theta/2)^{-1}$, where Θ is the total cone angle.

Figure 4.19 is an example of the above problem, done for $\gamma = 1$ and $\gamma = 0$, respectively. If $\lambda_c = 10.7$ μm, the background is a 300 K blackbody, and the spectral filter characteristics are those shown in Fig. 4.20, then the theoretical limit in D^* for this case is

$$D^*(\lambda_c) = \frac{\lambda_c}{2hc\sqrt{\pi c}}\left\{\int_0^{10.5\ \mu m} \frac{\exp(c_2/\lambda T_1)}{\lambda^4[\exp(c_2/\lambda T_1) - 1]^2}\,d\lambda + \int_{10.5\ \mu m}^{10.7\ \mu m} \frac{\exp(c_2/\lambda T_2)}{\lambda^4[\exp(c_2/\lambda T_2) - 1]^2}\,d\lambda\right\}^{-1/2} (\sin\pi/40)^{-1} \quad (4.138)$$

The factor $(\sin\pi/40)$ is due to the cooled cone, which has the effect of increasing $D^*(\lambda_c)$ by $(\sin\Theta/2)^{-1}$, with Θ equal to $\pi/20$ in the example. One can assume that, since the absorption of the filter is 100% outside the 10.5- to 10.7-μm region, the filter has an emittance of unity in that region. Also, $T_1 = 77$ K and $T_2 = 300$ K. This also assumes that $\eta(\lambda) = 1$ for $0 < \lambda < \lambda_c$. It is reasonable to assume further that the first integral within the brackets has an effective value of zero because of the small value of T_1. Then D^* can be written

$$D^*(\lambda_c) = \frac{\lambda_c}{2hc\sqrt{\pi c}}\left\{\int_{10.5}^{10.7} \frac{\exp(c_2/\lambda T_2)}{\lambda^4[\exp(c_2/\lambda T_2) - 1]^2}\,d\lambda\right\}^{-1/2} (\sin\Theta/2)^{-1} \quad (4.139)$$

For $\gamma = 0$, one has

$$D^*(\lambda_c) = \frac{\lambda_c}{2hc\sqrt{\pi c}}\left\{\int_{10.5}^{10.7} \frac{1}{\lambda^4[\exp(c_2/\lambda T) - 1]}\, d\lambda\right\}^{-1/2} (\sin\Theta/2)^{-1} \ . \qquad (4.140)$$

Using the mks system of units, one can evaluate the integral where

h = Planck's constant = 6.627×10^{-34} W s^2 or J Hz^{-1}
c = velocity of light = 3×10^8 m s^{-1}
k = Boltzmann's constant = 1.38×10^{-23} W s K^{-1} or J K^{-1}
λ = optical wavelength in meters.

Therefore, the theoretical limit in D^* is given by

$$D^*(\lambda_c) = 1.3 \times 10^{11} (\sin\Theta/2)^{-1} \ , \qquad \text{cm Hz}^{1/2}\,\text{W}^{-1} \qquad (4.141)$$

$$= 1.7 \times 10^{12} \ , \qquad \text{cm Hz}^{1/2}\,\text{W}^{-1} \ . \qquad (4.142)$$

This is about a factor of 50 times better than the $D^*(\lambda_c)$ value obtained without a cooled spectral filter and a cold aperture stop.

Figure 4.21 gives the $D^*(\lambda_c)$ value versus long wavelength threshold λ_c at different background temperatures, assuming the detector looks into a hemisphere. Figure 4.22 gives the effect on the theoretical limit of D^* caused by the reduction in the solid angle through which the detector views the background.

Van Vliet[19] has shown that, at equilibrium, a photoconductor has a total noise power that can be no less than twice the value of the photon noise power

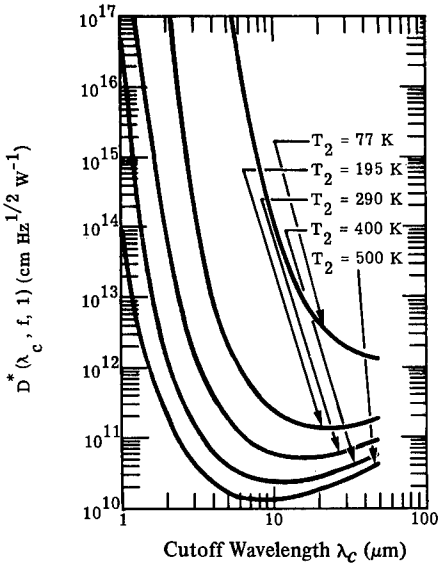


Fig. 4.21 Photon noise-limited D^* at peak wavelength λ for various background temperatures and a hemispherical field of view.

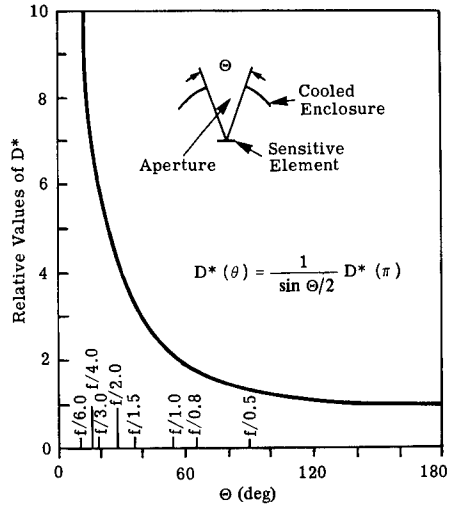


Fig. 4.22 Improvement factor in D^* versus cone angle Θ.

alone. This means that for photoconductors, the values of $D^*(\lambda_c)$ given in Fig. 4.21 should be divided by the square root of 2.

Lattice G-R Noise. This noise is a random generation and recombination of free charge carriers resulting from interactions with the vibrating atoms of the crystal lattice. The noise-voltage spectrum caused by this process has been derived by Van Vliet[19] and Kruse et al.[20]:

$$v_n = R_d I \left[\frac{4\tau \Delta f}{\overline{N}(1 + 4\pi^2 f^2 \tau^2)} \right]^{1/2} , \tag{4.143}$$

where $\overline{N}$ is the average total number of free charge carriers in the responsive element and τ is the carrier lifetime.[18]

Current, 1/f, or Modulation Noise. The physical mechanisms causing this noise are not completely understood. It has been observed in non-ohmic contacts, crystal surfaces, and in some cases the bulk of the crystal itself. The noise can be minimized with proper fabrication procedures. A general empirical expression for the noise voltage is

$$v_n \propto \left(\frac{I^2}{f A_d} \Delta f \right)^{1/2} R_d , \tag{4.144}$$

where I is the dc current through the responsive element, A_d is the detector area, and f is the electrical frequency.

Johnson Noise. This noise is caused by the random motion of free charge carriers in the responsive element, which is the result of collisions between the free charge carriers and the lattice atoms and exists regardless of the presence of a bias current. Since the degree of motion of the lattice sites will depend on the lattice temperature, the degree of Johnson noise depends on lattice temperature. The expression derived by Johnson for this noise voltage is

$$v_n = (4kTR_d \Delta f)^{1/2} , \tag{4.145}$$

where k is Boltzmann's constant, T is the temperature of the responsive element, and Δf is the electrical frequency bandwidth.[21]

Shot Noise. This noise is due to the discreteness of free electrons and holes as they pass across p-n junctions. They appear as minute current pulses, which show up as a random noise current or voltage in the exterior circuit. The expression for the shot-noise spectrum given by van der Zeil[22] is

$$v_n = R_d (2eI \Delta f)^{1/2} , \tag{4.146}$$

where I is the dc current across the p-n junction and e is the charge of an electron. The total noise in a photoconductive detector is the sum of the noises (excluding shot noise) discussed in the above sections:

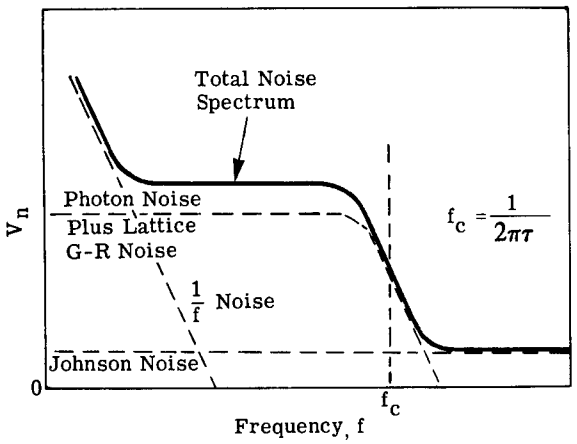


Fig. 4.23 Frequency spectrum for the total noise spectrum in a photoconductor.

$$(v_n)_{\text{total}} = [(v_n)_{ph}^2 + (v_n)_J^2 + (v_n)_{1/f}^2 + (v_n)_{gr}^2]^{1/2} \ . \tag{4.147}$$

Since a photoconductive detector does not have a *p-n* junction, shot noise is not present. The frequency spectrum of the sum of all these noises is given in Fig. 4.23.

4.4 DETECTOR CHARACTERIZATION

This section contains an enumeration and description of the important parameters associated with infrared detectors. The figures of merit that are used for comparing detectors are also discussed. This is followed by a description of the currently accepted procedures for testing detectors.

4.4.1 Detector Parameters

The following parameters are important when one measures the properties of infrared detectors or interprets data measured by others:

1. the infrared radiation incident on the detector
2. the electrical output of the detector (which consists of the signal and noise voltages mixed)
3. the geometrical properties of the detector
4. the detector as a circuit element
5. the detector temperature
6. the bias.

Incident Infrared Radiation. The radiation incident on a detector is characterized by the distribution of the radiant power with respect to wavelength, modulation frequency, position on the responsive element, and the direction of arrival. In the laboratory, the source of infrared radiation is usually a blackbody operated at 500 K. The accuracy of the measurements will be determined by the accuracy of the temperature setting and the uniformity of temperature in the source cavity.

The axis of the conical blackbody cavity should be slightly off-axis so that the apex is not visible, normal to the sensitive surface of the detector. The radiant power incident on the detector is measured in watts. The sensitive surface of the detector is also irradiated by optical power from the surroundings; this is called *ambient radiation.* The signal radiation is usually modulated periodically in time. This modulation is usually achieved by rotating a multibladed wheel at constant angular frequency ω between the radiation source and the detector. This device, called a *chopper*, produces a fundamental frequency f. The instantaneous power $\Phi(t)$ incident on the detector can be expressed as the sum of Fourier components:

$$\Phi(t) = \Phi_0 + \Phi_1 \cos(2\pi ft + \delta_1) + \Phi_2 \cos(2 \cdot 2\pi ft + \delta_2) + \ldots . \tag{4.148}$$

Each Fourier component has amplitude Φ_k and phase δ_k. Each component represents a sinusoidally modulated signal, with frequency f, $2f$, $3f$, etc. The fundamental component is the sinusoidal component of frequency f. The rms value is

$$\Phi_s = 2^{-1/2}\Phi_1 . \tag{4.149}$$

The radiation can have a 500 K blackbody spectral distribution, or it can be optically filtered in various ways to produce a specific character (see Sec. 4.4.3).

In measuring detector characteristics, one uses the rms value of the sinusoidal fundamental component to compute the detector figures of merit.

The Electrical Output of the Detector. There are two additive components of the electrical output from a detector. The first is the signal voltage v_s (or signal current i_s), and the second is the noise voltage v_n (or the noise current i_n). Over a sufficiently long time period, the two voltages (or currents) can be distinguished because the signal voltage v_s is coherent with the signal radiation power Φ_s, while the noise is not. If the signal power is a periodic function of time, the signal voltage is as well:

$$\begin{aligned} v_s(t) &= V_0 + V_1 \cos(2\pi ft + \psi_1) \\ &\quad + V_2 \cos(4\pi ft + \psi_2) + \ldots . \end{aligned} \tag{4.150}$$

In general, the magnitude of the signal is a function of the bias voltage V_B applied to the detector, the modulation frequency f, the wavelength λ, the incident irradiance E, and the detector area A_d. In functional notation, this can be written

$$v_s = v_s(V_B, f, \lambda, E, A_d) . \tag{4.151}$$

Over a wide range of input powers, the signal is a linear function of the incident radiation power. Since only those detectors that have this property will be considered, v_s/EA_d is a constant for fixed V_B, f, and λ. In many detectors, the signal's dependence on the modulation frequency can be separated from its dependence on the wavelength of the incident signal radiation power. Detectors

are said to have the factorability property in those wavelength regions in which separation is possible. Only detectors having this factorability property are considered here. Under these conditions the detector signal is given by

$$v_s = v_s(V_B,E,A_d)v(\lambda)v(f) \ , \tag{4.152}$$

where $v(\lambda)$ and $v(f)$ are normalized functions with values ranging from 0 to 1.

The functional dependence of the signal on the applied bias may also be separable. There are useful detectors that do not have this property, however; therefore caution must be exercised in making this assumption.

The dependence of the detector signal and noise on the applied bias is measured at a specified modulation frequency with specified incident radiation. These data are usually reported as a graph labeled "determination of optimum bias."

The dependence of a detector signal on the frequency at which the incident radiation is modulated is a measure of its temporal response. The measurements are usually made with constant radiation signal power and bias value. The results can be reported in a plot of relative signal voltage versus modulation frequency. Because such a plot represents the frequency response of the detector signal, a time constant, called the *responsive time constant*, can be determined. The responsive time constant is a function of applied bias. The peak detective frequency is the modulation frequency that maximizes the SNR.

The signal dependence of a detector on the wavelength of the radiation signal power is a measure of its ability to respond to the radiation of different wavelengths. These spectral responsivity measurements are made at a constant bias value and modulation frequency. Such data are reported on a plot of relative signal, normalized to constant radiation signal power at each wavelength, versus wavelength. Such a plot is usually labeled "relative spectral response."

The second additive component of the electrical voltage from a detector is the noise voltage. The detector noise is the electrical voltage or current output from the detector that is not coherent with the signal radiation power. The rms voltage (or current) of the electrical noise is defined as the square root of the time average of the square of the difference between the instantaneous voltage (or current) and the time average voltage (or current):

$$v_n = \{[\mathbf{v}_n(t) - \overline{\mathbf{v}_n(t)}]^2\}^{1/2} \ , \tag{4.153}$$

$$i_n = \{[\mathcal{I}_n(t) - \overline{\mathcal{I}_n(t)}]^2\}^{1/2} \ . \tag{4.154}$$

In practice, the average values of $\mathbf{v}(t)$ and $\mathcal{I}(t)$ are usually zero because ac-coupled amplifiers are usually used. Then v_n and i_n become

$$v_n = \{\overline{[\mathbf{v}_n(t)]^2}\}^{1/2} \ ,$$

$$i_n = \{\overline{[\mathcal{I}_n(t)]^2}\}^{1/2} \ . \tag{4.155}$$

The detector noise, which is random, is a function of the bias voltage, modulation frequency, and area of the detector. With certain assumptions, it can

be shown that the detector SNR varies directly as the square root of the detector area when the modulation frequency and bias are constant. The noise voltage is measured with the detector shielded from the radiation signal source. The detector noise, as reported, is referred to the output terminals of the detector and normalized to a 1-Hz effective noise bandpass.

The detector noise may be reported as a family of curves representing the different noise values obtained at several bias values plotted against frequency and labeled noise spectrum.

Geometrical Properties of the Detector. There is usually no ambiguity in selecting the responsive planes of an infrared detector. If the responsive element itself is in the form of a thin, flat layer (such as evaporated lead salt photoconductors), the responsive plane is the plane of the sensitive lead salt film. In the case of curved photoemissive detectors, the adopted responsive plane is the plane that contains the straight edges of the photocathode. Positions on the adopted responsive plane are defined by a rectangular Cartesian coordinate system. The responsivity $\mathcal{R}(x,y)$ of a small spot on the plane is measured with a small spot of radiation. The effective area of the adopted responsive plane is defined as

$$A_e = \int_{A_d} \int \frac{\mathcal{R}(x,y)\, dx\, dy}{\mathcal{R}_{\max}} \ . \tag{4.156}$$

The detector solid angle Ω is the solid angle from which the detector receives radiation from the outside. For flat detector responsive elements, the effective solid angle is the solid angle weighted at each angle by the cosine of the angle of incidence. This is called the *effective weighted solid angle* and is designated as Ω_e. For detectors with circular symmetry (i.e., the detector responsivity is independent of the azimuthal angle ϕ_a), the total cone angle Θ may be used to define Ω_e by

$$\Omega_e = \pi \sin^2(\Theta/2) \ . \tag{4.157}$$

The projected solid angle of a detector that has a hemispherical field of view is π sr.

The Detector as a Circuit Element. The properties of a detector as a circuit element will usually depend on the electrical frequency and on the amount of ambient radiant power Φ and on the dc current I_B through the detector. The impedance of the detector can be written as a complex impedance:

$$\tilde{Z}_d = R_d + iX_d \ , \tag{4.158}$$

where R_d is the resistance of the detector, given by the ratio of the dc voltage to the dc current, i.e.,

$$R_d = \frac{V}{I} \ , \tag{4.159}$$

where i is $\sqrt{-1}$ and X_d is the detector reactance.

When a detector is tested, it is connected to an amplifier and in some cases to bias sources. The load impedance $\bar{Z}_L$ is the impedance of the external circuit as seen from the terminals of the detector. Usually the load impedance is almost purely resistive.

The Detector Temperature. Detectors that operate without refrigerators have responsive elements with temperatures equal to or higher than the ambient temperature. Refrigerated detectors have designated temperatures equal to the temperature of the coolant. The detector signal and noise voltages and the resistance are greatly influenced by the operating temperature of the detector.

The Bias. Many detectors require an external bias of some kind. In the case of a photoconductive detector, the bias is an applied electrical voltage. In general, the signal and noise characteristics of a detector are measured over the entire useful range of bias values. Since both the detector signal and noise are functions of the bias and modulation frequency, the optimum bias value reported is the value that maximizes the SNR at a stated modulation frequency.

4.4.2 Detector Figures of Merit

The most basic performance quantity is the voltage (or analogous current) responsivity

$$\mathcal{R}_v(\lambda,f) = \frac{V_s}{\phi_e(\lambda)} , \tag{4.160}$$

where

λ = optical wavelength
f = modulation frequency
V_s = signal voltage due to ϕ_e
$\phi_e(\lambda)$ = spectral radiant incident power.

As an alternative to the above monochromatic quantity, the blackbody responsivity may be specified:

$$\mathcal{R}_v(T,f) = \frac{V_s}{\int_0^\infty \phi_e(\lambda)\, d\lambda} . \tag{4.161}$$

Another quantity of interest is the noise equivalent power (NEP). The NEP is the incident power on the detector generating a signal output equal to the noise output:

$$\text{NEP} = \frac{V_n}{\mathcal{R}_v} = \frac{i_n}{\mathcal{R}_i} . \tag{4.162}$$

The detectivity D is the reciprocal of NEP:

$$D = \frac{1}{\text{NEP}} . \tag{4.163}$$

Both NEP and detectivity are functions of electrical bandwidth and detector area, so a normalized detectivity is defined as

$$D^* = D(A_d\Delta f)^{1/2} = \frac{(A_d\Delta f)^{1/2}}{\text{NEP}} \ . \tag{4.164}$$

Either a spectral or blackbody D^* can be defined in terms of the corresponding type of NEP. Useful equivalent expressions to (4.164) include:

$$D^* = \frac{(A_d\Delta f)^{1/2}}{V_n}\mathcal{R}_v = \frac{(A_d\Delta f)^{1/2}}{i_n}\mathcal{R}_i = \frac{(A_d\Delta f)^{1/2}}{\phi_e}(\text{SNR}) \ . \tag{4.165}$$

Blackbody $D^*(T,f)$ may be found from spectral detectivity:

$$D^*(T,f) = \frac{\int_0^\infty D^*(\lambda,f)\phi_e(T,\lambda)\,d\lambda}{\int_0^\infty \phi_e(T,\lambda)\,d\lambda} = \frac{\int_0^\infty D^*(\lambda,f)E_e(T,\lambda)\,d\lambda}{\int_0^\infty E_e(T,\lambda)\,d\lambda} \ , \tag{4.166}$$

where $\phi_e(T,\lambda)$ is the incident blackbody radiant flux (W), and $E_e(T,\lambda)$ is the blackbody irradiance (W cm^{-2}).

The expression for shot noise can be used to derive the background-limited infrared photodetector (BLIP) detectivity

$$D^*_{\text{BLIP}}(\lambda,f) = \frac{\lambda}{hc}\left[\frac{\eta}{2\sin^2\theta_{1/2}\int_0^{\lambda_{\max}} M_P(\lambda,T_B)\,d\lambda}\right]^{1/2} \ , \tag{4.167}$$

where η is the detector quantum efficiency (photoelectrons per photon), $\theta_{1/2}$ is the half-angle field of view, and $M_P(\lambda,T_B)$ is the blackbody spectral radiant flux exitance (s^{-1} cm^{-2} μm^{-1}):

$$M_{P,\lambda}(\lambda,T) = \frac{2\pi c}{\lambda^4[\exp(hc/\lambda KT) - 1]} = \frac{1.885 \times 10^{23}}{\lambda^4[\exp(14{,}388/\lambda T) - 1]} \ . \tag{4.168}$$

Equation (4.167) holds for photovoltaic detectors, which are shot-noise limited. Photoconductors that are generation-recombination noise limited will have a lower D^*_{BLIP} by a factor of $\sqrt{2}$. All of the photon-noise-limited expressions given here are only for Poisson statistics, when the Bose-Einstein factor is near 1.

The relation of blackbody D^*_{BLIP} as a function of the peak spectral D^*_{BLIP} is

$$D^*_{\text{BLIP}}(T_S,f) = D^*(\lambda_P,f)\left[\frac{(hc/\lambda_P)\int_0^{\lambda_P} M_P(T_s,\lambda)\,d\lambda}{\sigma_e T_S^4}\right] \ . \tag{4.169}$$

All of the D^*_{BLIP} expressions have assumed a Lambertian source subtending a half-angle of $\pi/2$ radians. A special figure of merit is designed *only* for the BLIP case:

$$D^{**}(\lambda,f) = \sin\theta_{1/2} D^*_{BLIP}(\lambda,f) \ . \tag{4.170}$$

This allows direct comparison of BLIP detectors without reference to the field of view.

Detectivity-Frequency Product (D^*f^*). Another figure of merit for performance of an infrared detector suggested by Williams[23] and Borrello[24] is the product D^*f^* where D^* is the maximum specific detectivity as defined previously and f^* is the upper frequency at which D^* has decreased to 0.707 of its maximum value. This figure of merit allows the performance potential of detectors to be compared for high-frequency, low-background applications. Borrello has derived the maximum value of D^*f^*:

$$(D^*f^*)_{max} = \frac{1}{2\pi}\left(\frac{\sigma_c}{4E_g h}\right)^{1/2} , \tag{4.171}$$

where σ_c is the photon capture cross section, E_g is the energy of a photon, and h is Planck's constant.

Curves of maximum D^* versus frequency for some typical detectors are shown in Fig. 4.24. The experimental results[24] give a D^*f^* value of about 10^{17}, whereas calculations based on capture cross sections predict values of 10^{18} and 10^{19}.

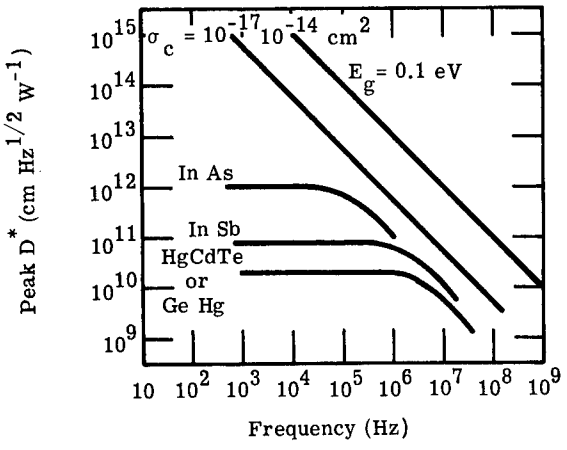


Fig. 4.24 Maximum D^* versus frequency: calculations for two σ_c values compared with measured detectors.

4.4.3 Detector Performance Tests[a]

Test Equipment. The equipment used to measure important detector parameters consists of electronic equipment, a blackbody source, a variable frequency source, a monochromatic source, and a reference detector. For the most part, these instruments can be obtained from various commercial suppliers and can be assembled to meet specific measurement requirements.

The electronics generally used to measure detector signal and noise characteristics consist of a preamplifier, amplifier, and a spectrum analyzer. Auxiliary equipment consists of an oscillator and calibrated attenuator used for electrical calibration, and a variable voltage supply suitable for providing detector bias. Figure 4.25 is a block diagram of a typical arrangement of equipment suitable for signal and noise measurements in the 1-kHz to 2-MHz frequency range. If a variety of detectors is to be measured, it is convenient to have several similar arrangements of instruments to cover the necessarily wide frequency range. A number of preamplifiers will also be necessary to cover a wide range of detector impedances.[25] When possible, the preamplifier should be placed with the detector in a shielded enclosure to minimize input capacity. Apparatus suitable for measuring the signal and noise of thermal-type detectors over a frequency range of 1 to 100 Hz are shown[26] in Fig. 4.26. Multiplexed detector arrays require clock signals, sample-and-hold circuits, and A/D converters to characterize the detector elements.

The primary requirement for measuring detector responsivity is a stable, modulated source of infrared radiation with known spectral characteristics. A blackbody simulator is a convenient means of meeting these requirements. If fixed-frequency modulators are used, it is convenient to have at least three available, operating at frequencies of approximately 10, 100, and 1000 Hz. The source must produce an accurately known irradiance over the responsive plane of the detector, and the irradiance must be uniform over the sensitive area of the detector. The spectral irradiance used is the rms value of the fundamental component of the modulation frequency. In determining this value, one must take into account the radiation from the modulator. To compare measured data easily, operators of many detector laboratories operate their blackbody sources at a temperature of 500 K. Irradiance values of the order of microwatts per centimeter squared are appropriate for many types of detectors.

A variable frequency source is required to measure the dependence of detector signal on modulation frequency. Basically, any stable source having a suitable spectral output and equipped with a variable speed modulator may be used. The irradiance E produced by the source must be uniform over the sensitive surface of the detector. At audio and subaudio frequencies, satisfactory mechanical modulators are easily fabricated.[27] However, at more than 50 kHz mechanical modulators become awkward, blade diameters become large,

[a]Sections 4.4.3 and 4.4.4 essentially reproduce material prepared by W. L. Eisenman, J. D. Merrian, and R. F. Potter of the Electro Materials Sciences Center, Naval Electronics Laboratory Center, Corona, California (now the Naval Ocean Systems Center, San Diego) and originally published in the 1965 *Handbook of Military Infrared Technology*.

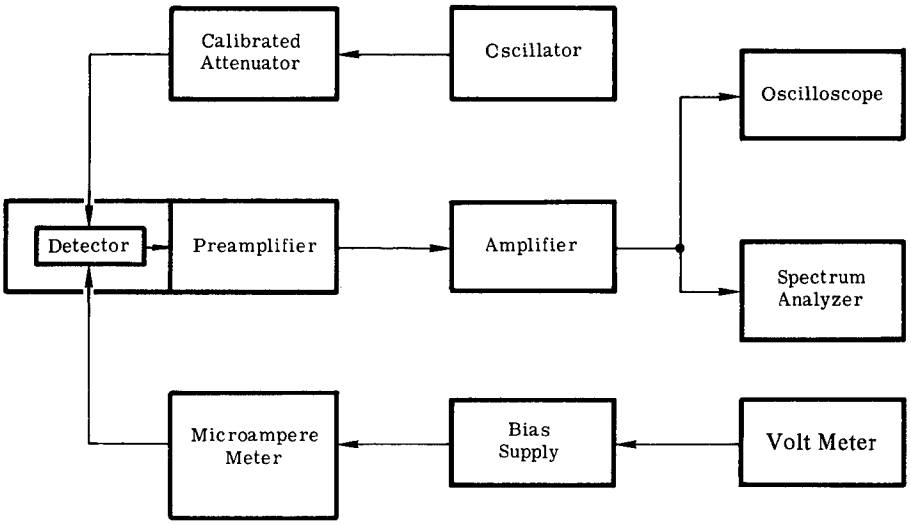


Fig. 4.25 Block diagram for detector signal and noise measurements with a range of 1.0 kHz to 1.0 MHz.

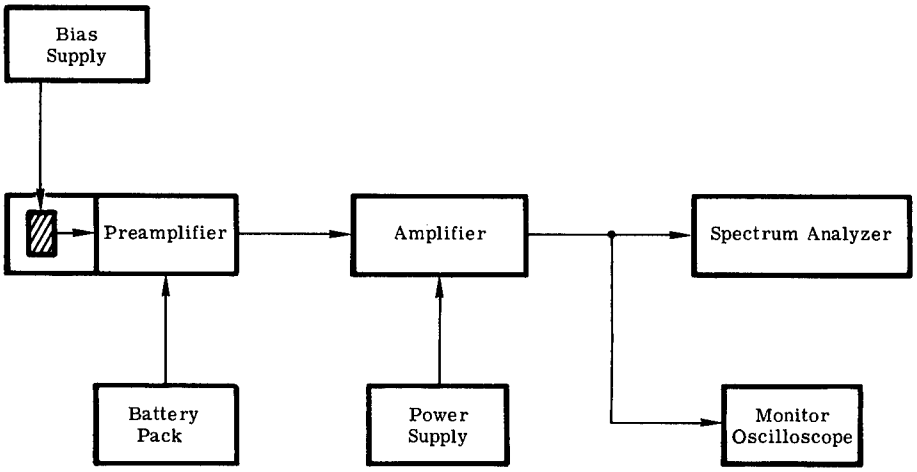


Fig. 4.26 Block diagram for detector signal and noise measurements with a range of 1.0 to 100.0 Hz.

aperture sizes are reduced to small dimensions, and the blades must be driven at quite dangerous speeds.

Light-emitting diodes (LEDs) can also serve as variable frequency sources for frequencies into the megahertz region. The apparatus is quite simple. One instrument arrangement is shown in Fig. 4.27. The system utilizes the beat frequency oscillator (BFO) output voltage available on several models of audio-frequency wave analyzers. This BFO voltage has a frequency that is always

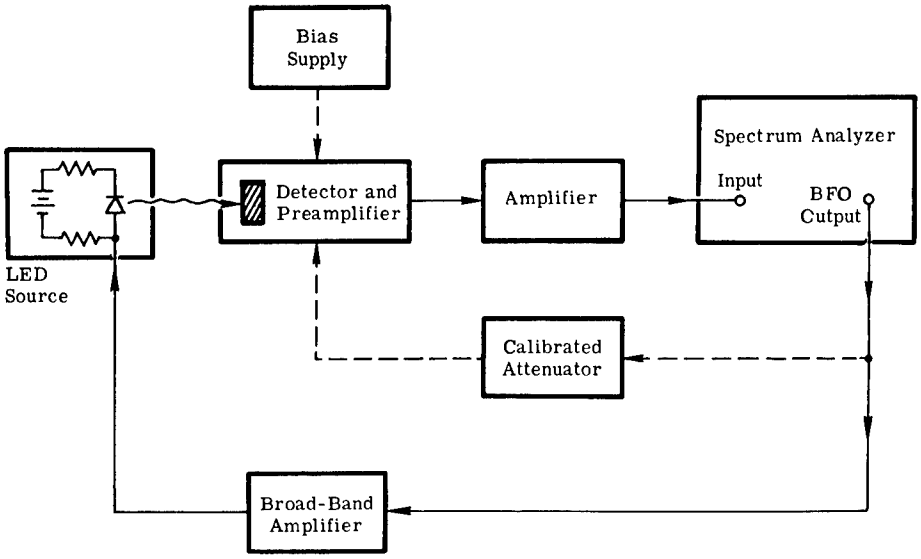


Fig. 4.27 Block diagram for a variable frequency source using LEDs with a range up to 1.5 MHz.

equal to the input frequency of the analyzer. Thus, when the input frequency of the analyzer is varied over the frequency range of the equipment, the frequency of the BFO voltage automatically follows. The BFO voltage from the analyzer is amplified and applied to an LED. The modulated radiation from the diode is focused on the detector being measured. The electrical signal from the detector is amplified and then applied to the input of the analyzer.

This arrangement has two distinct advantages in addition to its simplicity. The measurement of detector frequency response is quite rapid, since only one control on the wave analyzer needs be adjusted, and the narrow bandpass of the wave analyzer produces a large detector SNR. Thus, a relatively small amount of input radiation is necessary.

The major disadvantage of this method is the narrow spectral distribution of the infrared radiation emitted from the LED. Gallium arsenide and indium arsenide diodes emitting at 0.8 and 3.2 μm, respectively, are commercially available. Diodes of the ternary alloys are available that provide emission at longer wavelengths.

The monochromatic, infrared radiation source consists of a stable, broad spectral-band source of infrared radiation, a modulator, a monochromator or other spectral filter, and an optical system that directs the monochromatic radiation to the detectors. Tungsten filament lamps, Nernst glowers, and glow bars are commonly used sources depending on the wavelength region of interest. Since the radiant output of both the glower and glow bar will be affected by air currents, these sources should be provided with a suitable housing or chimney. The modulator may be a fixed-frequency device and is normally placed at the entrance slit of the monochromator, which should be capable of providing a wavelength band of radiation not wider than about 1/25th of the

center wavelength. Because scattered radiation may be a problem, particularly at the longer wavelengths, a double monochromator is preferable for this application. However, satisfactory performance may be obtained from a single monochromator by fitting the instrument with suitable rejection filters.[27] The detector being measured and the reference detector are alternately illuminated by an optical system placed at the exit slit of the monochromator. A typical arrangement is shown in Fig. 4.28.

The spectral response of a detector is obtained by comparing the signal from the detector to the signal from a reference detector as a function of the wavelength of the incident radiation. Detectors utilizing a cavity as the radiation receiver have been used as reference standards,[28–30] but they are difficult to obtain and hard to use because of their low sensitivity and slow speed of response. A radiation thermocouple is a convenient detector for use as a reference standard, provided its spectral sensitivity has been determined (by comparison to a cavity-type detector). The relative spectral sensitivity of a typical radiation thermocouple compared to such a cavity is shown in Fig. 4.29. The decline in sensitivity is relatively smooth and there is no difficulty in

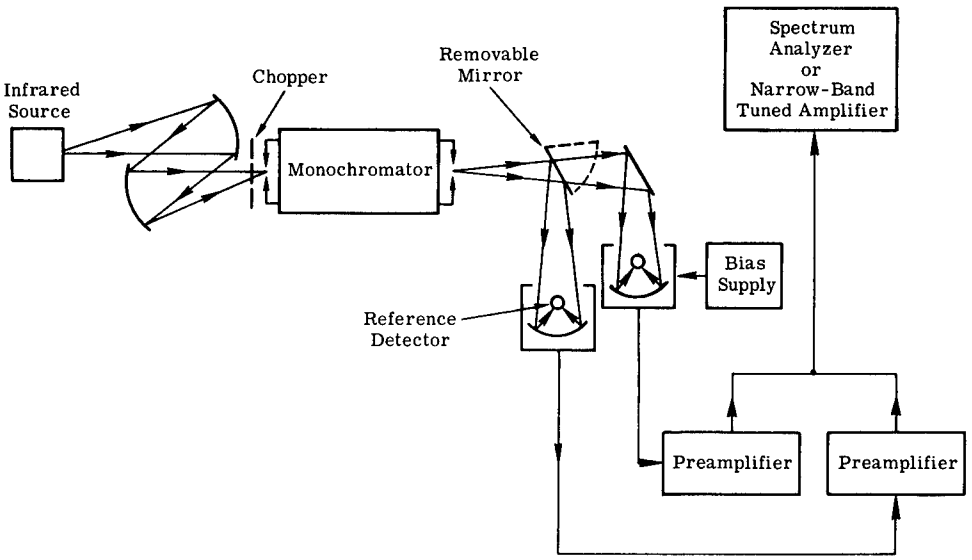


Fig. 4.28 Block diagram of instrument arrangement for measuring spectral response.

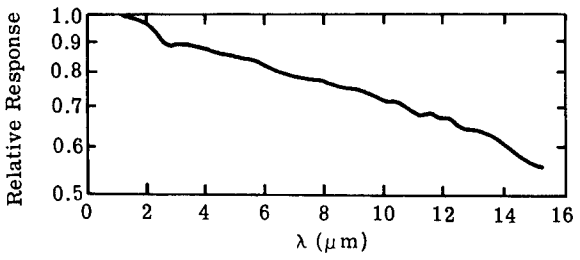


Fig. 4.29 Relative spectral response of a typical radiation thermocouple as compared to a special black conical cavity reference detector.

correcting the data. Since the sensitivity of a thermal detector may not be uniform across the responsive plane,[31] the optical system used at the exit slit of the monochromator should be adjusted to flood completely the thermocouple receiver. The monochromator should also have a reasonably uniform exit pupil.

Test Procedures. Test procedures may be divided into two groups. In the first group are the measurements necessary to determine the detector responsivity, and in the second are measurements that yield the root power spectrum of the noise. The arrangement of the electrical equipment is the same for both groups. Figure 4.25 shows such an arrangement for use on common photoconductive detectors that require a single bias voltage V_B.

The determination of detector responsivity involves three separate measurements. In each of these measurements the incident signal radiation must be normal to the detector's responsive plane, and the amount of signal radiation must be confined to a range in which the output signal from the detector is proportional to the incident radiant power. Confirmation of this linearity may be necessary in some cases.

The measurement of the responsivity involves the use of the factorability property [see Eq. (4.151)]:

$$v_s(V_B,\Phi_s,A_d,\lambda,f) = v_s(V_B,\Phi_s,A_d)v(\lambda)v(f) \ . \tag{4.172}$$

The blackbody and the variable frequency sources must be equipped with a filter that limits the radiation to a wavelength band within which the factorability property holds.[30]

The first step in measuring responsivity is to establish the range of bias values to be used with the detector being measured. Experience with similar detectors will usually indicate the approximate range of bias values. The range will normally cover at least one decade of bias voltage or current. The highest value of bias, normally known as the manufacturer's maximum bias, is explicitly stated by the manufacturer. Considerable care should be exercised if measurements are to be made at biases greater than this value. The detector noise should be carefully monitored and the bias should be increased in small steps. Operating some types of detectors in a region of high bias is very risky and you must rely on experience.

The blackbody source equipped with a modulator of frequency f is then used to irradiate the detector. The center frequency of the spectrum analyzer is set at f and the signal generator is set to zero. The reading v_s of the output meter is noted. Then the irradiance is adjusted to a value that gives the same reading v_s on the meter.

The open-circuit detector signal voltage v_s is the voltage across the calibrating resistor. These measurements are then repeated until the complete range of bias values has been covered.

The radiant power Φ_s incident on the detector is obtained from the known irradiance E upon multiplication by the detector area A_d:

$$\Phi_s = A_d E \ . \tag{4.173}$$

The corresponding detector responsivity is given by

$$\mathcal{R}_{BB}(V_B,f) = \frac{v_s(V_B,f)}{\Phi_s} \quad . \tag{4.174}$$

Next, with the spectrum analyzer set at frequency f and the bias applied, the detector is irradiated by the monochromatic source. The center wavelength of the monochromator is varied over the wavelength range of interest, and the relative signal voltage v_s of the detector is recorded as a function of wavelength. The detector under measurement is then replaced by the reference detector, and the relative signal voltage v_{ref} of the reference detector is recorded as a function of wavelength. (Some detectors may exhibit changes in spectral response as a function of bias. If these changes are significant, then several spectral response curves must be obtained for different bias values.)

The relative response $v(\lambda)$ as a function of the wavelength is then calculated by

$$v(\lambda) = \frac{v_s \mathcal{R}_{ref}(\lambda)}{v_{ref}} \quad , \tag{4.175}$$

where $\mathcal{R}_{ref}(\lambda)$ is the relative spectral responsivity of the reference detector.

The detector is irradiated with the variable frequency source. As the modulation frequency of the source is varied, the center frequency of the spectrum analyzer is continuously adjusted to the modulation frequency. The detector signal voltage v_s read on the meter is recorded as a function of frequency. The source is then removed, and the signal generator with a fixed attenuator setting is varied over the same range of frequencies. As the frequency of the signal generator is varied, the center frequency of the spectrum analyzer is continuously adjusted. The voltage v_c read on the meter is recorded as a function of frequency. The relative response $v(f)$, as a function of modulation frequency, is then computed by

$$v(f) = \frac{v_s(f)v_c(f_0)}{v_c(f)v_s(f_0)} \quad , \tag{4.176}$$

where f_0 is such that $v(f_0) = 1$.

The frequency response of some detectors may vary as a function of applied bias. If the change in frequency response is significant, then several frequency response curves must be measured over the entire range of bias values.

The measurement of the noise characteristics of a radiation detector requires good judgment and experience to ensure that the noise recorded is only the noise generated in the detector, load resistor, and the amplifier. Constant attention is required to prevent external sources of noise from influencing the results. It may be convenient to place a wideband oscilloscope in the electronic system ahead of the bandpass filter. The appearance of the noise trace on the oscilloscope is helpful in determining the presence of any extraneous noise.

In particular, the bias supply must not contribute appreciable noise. The bias source can be checked for internal noise by substituting a wire-wound resistor in place of the detector in the input circuit. The resistance of the wire-wound resistor should be approximately equal to the detector resistance. Bias

is then applied to the circuit and the noise noted on the output meter. The noise generated in the wire-wound resistor should be independent of the current flowing through the resistor.

Bias is then applied to the detector. All radiation sources except ambient background are removed. With the signal generator producing zero signal, the rms noise voltage indicated by the output meter is recorded as a function of frequency over the entire frequency range of interest. The voltage read is then denoted v_0.

The detector and load resistor are replaced by a wire-wound resistor having approximately the same resistance as the parallel combination of the detector and load resistor. The temperature of this wire-wound resistor is maintained such that the Johnson noise generated in the resistor is small compared to the noise generated in the amplifier. The rms noise voltage indicated by the output meter is again recorded as a function of frequency. This voltage is denoted v_a.

The signal generator is then adjusted to produce a calibration signal v_c across the calibrating resistor. This calibration signal is made approximately 100 times larger than the detector noise, the spectrum analyzer is tuned to the frequency of the calibration signal, and the voltage indicated on the output meter recorded. This procedure is repeated over the entire frequency range of interest. The system gain $g(f)$ is thus determined as a function of frequency.

The root power spectrum v_n^*, referred both to the terminals of the detector and to an infinite load impedance and corrected for amplifier noise, is calculated in units of rms volts per root Hertz from the following formula:

$$v_n^*(f,V_B) = \frac{\left[\dfrac{v_0^2(f,V_B) - v_a^2(f)}{g^2(f)} - v_T^2\left(\dfrac{R_d}{R_L}\right)^2\right]^{1/2}}{(\Delta f)^{1/2}} , \tag{4.177}$$

where R_d is the resistance of the detector. The thermal noise voltage v_T generated by the load resistor R_L in the noise bandwidth Δf is given by

$$v_T^2 = 4kTR_L\Delta f . \tag{4.178}$$

Note that Δf, which is the effective noise bandwidth of the measurement equipment, is defined by

$$\Delta f = \int_0^\infty \frac{g^2(f)}{g_m^2}\, df , \tag{4.179}$$

where $g(f)$ is the gain of the system as a function of frequency f, and g_m is the maximum value of the gain. The frequency f_m that corresponds to g_m is defined as the center frequency of the passband.

4.4.4 Performance Calculations

A limited set of conditions must be selected for the tests, but if these are properly chosen the detector response can be predicted under a variety of other operating conditions. Responsivity in functional form is written

$$\mathcal{R} = \mathcal{R}(V_B,f,\lambda) \ . \tag{4.180}$$

The parameters V_B, f, and λ enter into $\mathcal{R}$ only through the signal, because E and A are independent of them. Therefore, we have

$$v_s = v_s(V_B,f,\lambda) \ . \tag{4.181}$$

Because the parameters V_B, f, and λ are factorable (separable) in any usable detector, one may write

$$v_s = v_s(V_B)v(f)v(\lambda) \ . \tag{4.182}$$

The parameters $v(V_B)$, $v(f)$, and $v(\lambda)$ can be measured separately and reported graphically in the charts as follows:

- $v_s(V_B)$ in a chart for the determination of optimum bias
- $v(f)$ in a frequency response chart
- $v(\lambda)$ in a spectral response chart.

The subscript r will be used to designate the reported value and the subscript 1 to designate a desired value. The responsivity at bias V_{B1}, modulation frequency f_1, and radiation wavelength λ_1, are given by the relation

$$\mathcal{R}(V_{B1},f_1,\lambda_1) = \frac{\mathcal{R}(V_{Br},f_r,\lambda_r)v_s(V_{B1})v(f_1)v(\lambda_1)}{v_s(V_{Br})v(f_r)v(\lambda_r)} \ . \tag{4.183}$$

Since the responsivity is measured with a blackbody, usually at 500 K, the following relation is used:

$$\mathcal{R}(V_{Br},f_r,\lambda_{\max}) = \frac{\mathcal{R}_{BB}(V_{Br},f_r,T)}{\gamma_p} \ , \tag{4.184}$$

where $\lambda_{\max}$ is the infrared wavelength at which $v(\lambda)$ is a maximum. The symbol γ_p is given by the expression

$$\gamma_p = \frac{\int_0^\infty v(\lambda)\Phi_{s,\lambda}(T)\,d\lambda}{\int_0^\infty \Phi_{s,\lambda}(T)\,d\lambda} \ , \tag{4.185}$$

where $\Phi_{s,\lambda}(T)$ is the spectral radiant power from a blackbody at temperature T and wavelength λ. Also, since $v(\lambda)$ is normalized to the peak wavelength, one can evaluate at $\lambda_r = \lambda_{\max}$ and $v(\lambda_{\max}) = 1$. Then

$$\mathcal{R}(V_{B1},f_1,\lambda_1) = \frac{\mathcal{R}_{BB}(V_{Br},f_r,T)}{\gamma_p}\frac{v_s(V_{B1})}{v_s(V_{Br})}\frac{v(f_1)}{v(f_r)}\frac{v(\lambda_1)}{1} \ , \tag{4.186}$$

where $\mathcal{R}_{BB}(V_{Br},f_r,T)$ and γ_p [sometimes reported as the ratio $\mathcal{R}(\lambda_{\max})/\mathcal{R}_{BB}$] are reported parameters.

Unfortunately, noise is not factorable into independent functions of the parameters V_B and f. (The internal noise does not depend on λ.) Thus noise values desired at other than the measured values must be interpolated or estimated by the theoretical formulas given in Table 4.3.

The set of values of V_B, f, and λ that give the maximum or *peak* D^*, i.e., $D^*_{\max}$, is designated (V_{Bp}, f_p, and λ_p). At bias value V_{Bp}, the SNR is maximum; at chopping frequency f_p, the SNR is maximum; and at λ_p, the signal voltage is maximum. (The noise voltage is independent of λ.) Given $D^*_{\max}$, one can find D^* at other values of the set V_B, f, and λ values with the following equation:

$$D^*(V_{B1},f_1,\lambda_1) = D^*_{\max}(V_{Bp},f_p,\lambda_p) \times \frac{v_n(V_{Bp},f_p)v_s(V_{B1})}{v_n(V_{B1},f_1)v_s(V_{Bp})} \frac{v(f_1)}{v(f_p)} \frac{v(\lambda_1)}{v(\lambda_p)} . \tag{4.187}$$

A sample calculation of $\mathcal{R}$, D^*, and NEP will be performed next as an example for a detector operated at conditions different from those given in standard reported data. For this example, a photoconductive lead selenium (PbSe) detector operated at liquid nitrogen temperature will be used. Its characteristics are listed in Table 4.4 and shown in Figs. 4.30 and 4.31.

Suppose one wishes to find the responsivity and SNR at the following operating point:

$$\begin{aligned} f_1 &= 10^3 \text{ Hz} , \\ I_{B1} &= 75 \ \mu\text{A} , \\ \lambda_1 &= 6.0 \ \mu\text{m} . \end{aligned} \tag{4.188}$$

From the frequency response plot [Fig. 4.31(a)], the relative response $v(f)$ at 90 and 10^3 Hz is

$$\begin{aligned} v(90 \text{ Hz}) &= 1.0 , \\ v(10^3 \text{ Hz}) &= 0.78 . \end{aligned} \tag{4.189}$$

From the determination of an optimum bias plot [Fig. 4.31(b)], the values of $v(I_B)$ at bias currents of 50 and 75 μA are

$$\begin{aligned} v(50 \ \mu\text{A}) &= 9.0 \times 10^{-3} \text{ V} , \\ v(75 \ \mu\text{A}) &= 1.3 \times 10^{-2} \text{ V} . \end{aligned} \tag{4.190}$$

From the spectral response plot (Fig. 4.30) $v(\lambda)$, one finds the values at 6.0 μm:

$$v(6.0 \ \mu\text{m}) = 1.5 \times 10^{-1} . \tag{4.191}$$

From the test results in Table 4.4 one finds

$$\mathcal{R}_{BB}(50 \ \mu\text{A}, 90 \text{ Hz}, 500 \text{ K}) = 3.2 \times 10^4 \text{ V W}^{-1}$$

Table 4.4 Detector Characteristics Performance Parameters, and Test Conditions for a PbSe Detector (from Ref. 32)

Test Results	
$\mathcal{R}$ (500 K, 90 Hz)	3.2×10^4 V W^{-1}
NEE (500 K, 90 Hz) (Δf = 1 Hz)	2.9×10^{-9} W cm^{-2}
NEP (500 K, 90 Hz) (Δf = 1 Hz)	1.1×10^{-10} W
D^* (500 K, 90 Hz)	1.8×10^9 cm Hz$^{1/2}$ W^{-1}
Peak wavelength, λ_p or λ_{max}	4.5 μm
Peak modulation frequency	7.0×10^2 Hz
$D^*_{m,m}$ †	1.2×10^{10} cm Hz$^{1/2}$ W^{-1}
Effective time constant	1.2×10^2 μs
$\mathcal{R}(\lambda_{max})/\mathcal{R}_{BB}$	4.8
Conditions of Measurement	
Blackbody temperature	500 K
Blackbody flux density	7.7 μW cm^{-2}, rms
Chopping frequency	90 Hz
Noise bandwidth	5 Hz
Cell temperature	78 K
Cell current	50 μA
Load resistance	1.0×10^6 Ω
Transformer	—
Relative humidity	31%
Responsive plane from window	2.5 cm
Ambient temperature	24°C
Ambient radiation on detector	297 K (only)
Cell Description	
Type	PbSe (chemical)
Area	3.9×10^{-2} cm^2
Dark resistance	1.0×10^6 Ω
Field of view	180 deg
Window material	Sapphire

†D^* evaluated at maximum wavelength and maximum chopping frequency.

and

$$\frac{1}{\gamma^p} = \frac{\mathcal{R}(\lambda_{max})}{\mathcal{R}^{BB}} = 4.8 \ . \tag{4.192}$$

Using Eq. (4.186) and substituting in the above values, one has

$$\begin{aligned} \mathcal{R}(75\ \mu A,\ 10^3\ Hz,\ 6.0\ \mu m) &= 3.2 \times 10^4 \frac{1.3 \times 10^{-2}}{9.0 \times 10^{-3}} \frac{0.78}{1.0} \\ &\quad \times (1.5 \times 10^{-1})4.8 \\ &= 2.6 \times 10^4 \text{ V W}^{-1} \ . \end{aligned} \tag{4.193}$$

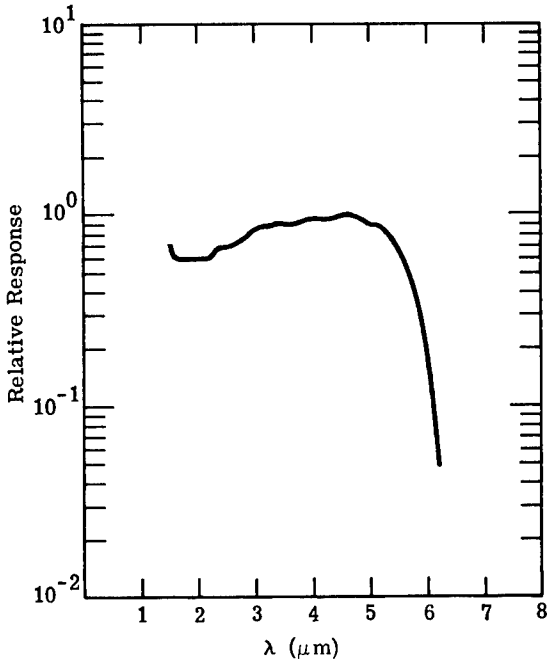


Fig. 4.30 Spectral response.[32]

Therefore, if this detector were of unit area and the radiation source had a flux density of 1.0 μW cm^{-2}, the expected signal level would be 2.6 × 10^4 μV from the detector at I_B = 75 μA, f = 10^3 Hz, and λ = 6.0 μm.

To find D^* at f_1, V_{B1}, λ_1, one must find the noise level at f_1, V_{B1}. From the noise spectrum plot of Fig. 4.31(d), interpolate at f = 10^3 to find the noise level for I_B = 75 μA. For a 1-Hz system bandwidth, this gives

$$v_n(75\ \mu\text{A}, 10^3\ \text{Hz}) = 3.1 \times 10^{-6}\ \text{V}\ ,$$
$$v_n(150\ \mu\text{A}, 7 \times 10^2\ \text{Hz}) = 6.2 \times 10^{-6}\ \text{V}\ . \tag{4.194}$$

The D^* at V_{B1}, f_1, λ_1 can be found with Eq. (4.187). Since the peak frequency is reported as 7.0 × 10^2 Hz and the peak bias is shown in the detectivity versus frequency plot at 150 μA,

$$D^*(V_{B1},f_1,\lambda_1) = D^*_{\max}(V_{B1},f_p,\lambda_p)\frac{v_n(150\ \mu\text{A}, 7.0 \times 10^2\ \text{Hz})}{v_n(75\ \mu\text{A}, 10^3\ \text{Hz})}$$
$$\frac{v_s(75\ \mu\text{A})}{v_s(150\ \mu\text{A})}\frac{v(10^3\ \text{Hz})}{v(7.0 \times 10^2\ \text{Hz})}\frac{v(\lambda_1)}{1}$$

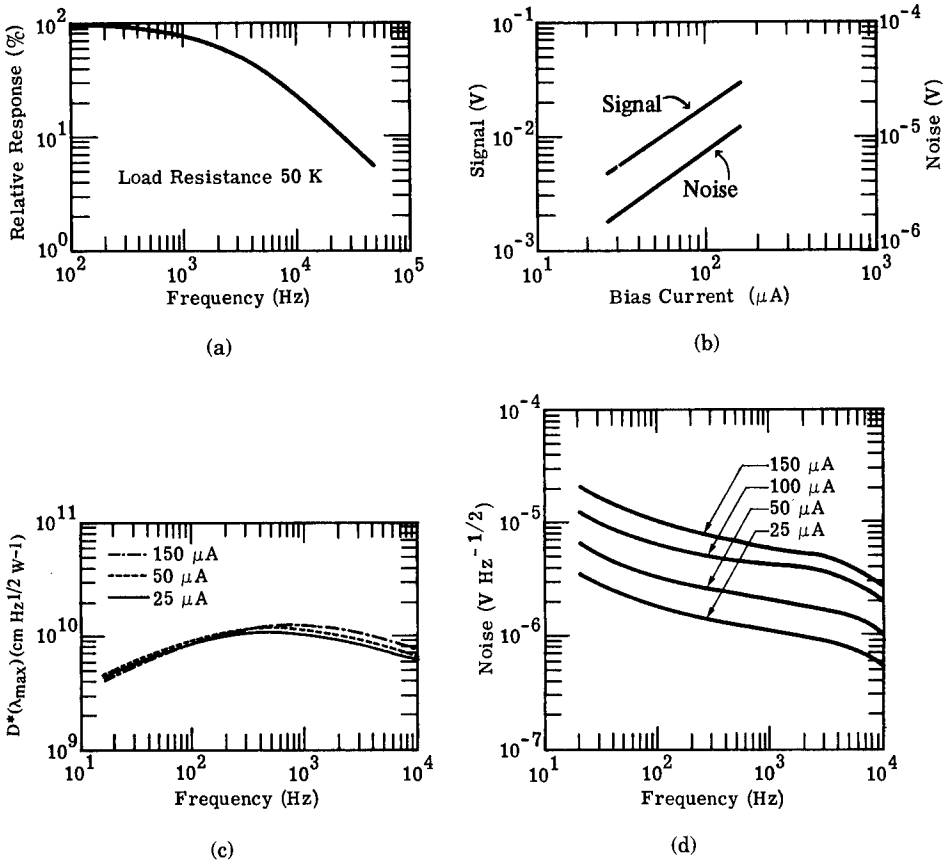


Fig. 4.31 Photoconductive lead selenium (PbSe) detector operated at liquid nitrogen temperature[33]: (a) frequency response, (b) determination of optimum bias, (c) detectivity versus frequency, and (d) noise spectrum.

$$= (1.2 \times 10^{10})\frac{6.2 \times 10^{-6}}{3.1 \times 10^{-6}}\frac{1.3 \times 10^{-2}}{2.9 \times 10^{-2}}\frac{0.78}{0.83}$$

$$\times (1.5 \times 10^{-1})$$

$$= 1.5 \times 10^{9}\ \text{Hz}^{1/2}\ \text{cm}\ \text{W}^{-1}\ . \tag{4.195}$$

If the detector has a 1-cm^2 area, a postdetector electronic system with unit bandwidth, and an incident radiation flux density of 1.0×10^{-6} W cm^{-2}, the expected SNR at I_{B1}, f_1, λ_1, would be 1.5×10^3.

The noise equivalent power at I_{B1}, f_1, λ_1, is

$$\text{NEP} = \frac{A_d^{1/2}}{D^*(I_{B1}, f_1, \lambda_1)} = \frac{(3.9 \times 10^{-2})^{1/2}}{1.5 \times 10^{9}} = 1.3 \times 10^{-10}\ \text{W}\ . \tag{4.196}$$

Some precautions and reservations should be kept in mind when extrapolating the performance data to low-background operations.

Figure 4.22 shows the theoretical improvement in D^* that can be achieved by reducing the field of view of the detector and thus reducing the background radiation flux. A reduction in background radiation can also occur through the use of cooled filters, or in space applications where the background temperatures can be very low. In general, the improvement in D^* is accompanied by a decrease in some other performance parameter.

Since the resistance of the detector is an inverse function of the background radiation, operation at low backgrounds results in an increased resistance and can cause the time constant of the system to be RC-limited. Thus, one price that the system designer should be prepared to pay is a decrease in frequency response.

Reduction of the background noise will improve the D^* only if the background noise is the dominant noise mechanism. Thus, successful operation at low backgrounds can involve the reduction of thermal generation-recombination noise by operating at lower temperatures. This can increase the complexity and power consumption of an infrared system.

A number of complex and, in general, nonlinear phenomena can arise from low-background operations, including high-field carrier sweep-out effects[34–37] and nonlinearities due to bias and memory effects (the previous irradiation history of the detector).[38,39] These nonlinear phenomena make precise calibration measurements a very complex and difficult undertaking.

4.5 SUMMARY OF COMMERCIAL DETECTOR PERFORMANCE

4.5.1 Performance Overview

A comparison of photon and thermal detector performance is given in Fig. 4.32. At shorter wavelengths, intrinsic semiconductor photon detectors provide essentially background photon flux-limited performance at 300 K for their quantum efficiency. At longer wavelengths, extrinsic silicon and germanium also approach background-limited performance. Thermal detectors offer flat (*bolometric*) wavelength dependance at a generally lower level of performance. Thermal detectors are also relatively slow devices with time constants in the millisecond range, while the fastest photon detectors have time constants measured in picoseconds.

Photoemissive detectors exhibit quantum efficiencies as high as 0.3 or more at their relatively short wavelengths of operation (Fig. 4.33). The interest in photoemissive materials is primarily for use as photocathodes in devices with photoelectric gain, such as photomultiplier tubes and image intensifiers.

Table 4.5 lists visible detector focal plane arrays announced by commercial manufacturers by 1990. Market forces and the relatively mature silicon technology result in larger arrays for visible and near-infrared wavelengths. Arrays of significant size and high quantum efficiency are starting to revolutionize the infrared imaging field, however, as discussed in Sec. 4.5.2.

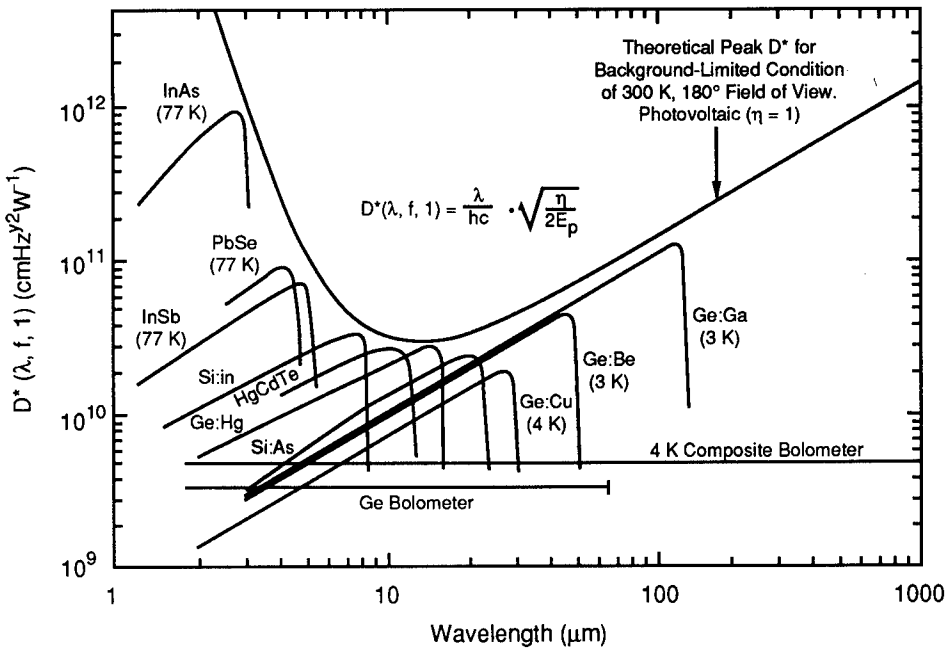


Fig. 4.32 Typical D^* for selected detectors.

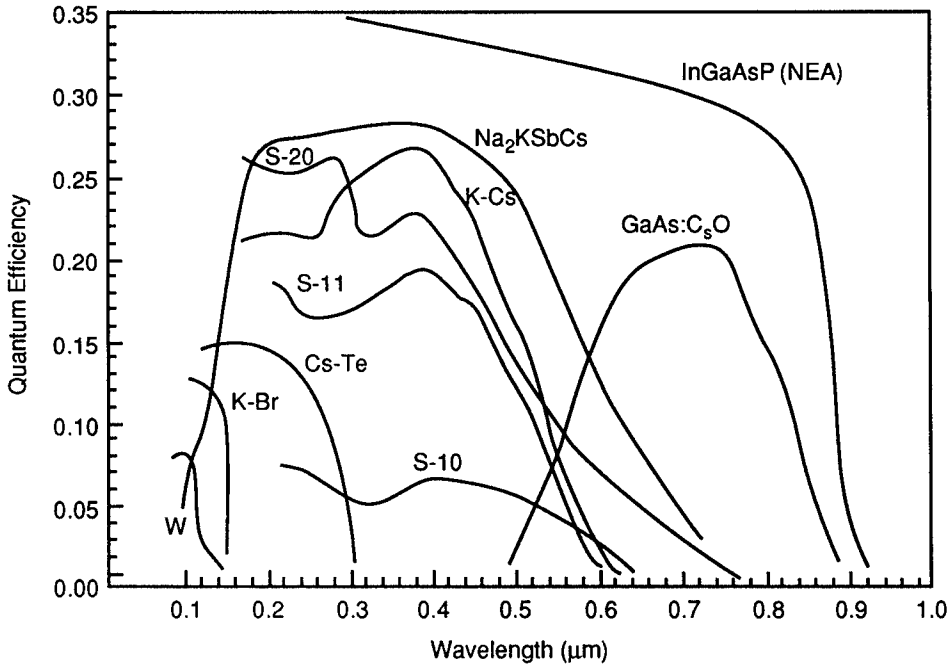


Fig. 4.33 Spectral response of photocathodes.

Table 4.5 Selected Commercial Detector Arrays

Detector Material	Array Size	Approximate Spectral Range (μm)	Manufacturer	Comments
Silicon	4096 × 4096	0.3–0.9	Ford Aerospace (Loral)	8192 × 8192 being designed
Silicon	8008 × 3	0.3–0.7	Kodak	Color
Silicon	2048 × 2048	0.3–0.9	Kodak	100% fill factor
Silicon	2048 × 2048	0.3–0.9	Reticon	Buried channel
Silicon	2048 × 2048	0.3–0.9	Tektronix	Line address

4.5.2 Imaging Sensor Performance

Infrared sensors are used to detect, image, and measure patterns of thermal heat radiation, which all objects emit. These sensors have been extensively developed[16] since the 1940s. Early devices consisted of single detector elements. This section considers only photon detectors, which sense the photon energy directly, as opposed to bolometers, which detect a rise in temperature caused by the absorption of IR radiation.

PbS was the first practical IR detector. It was developed prior to World War II in Germany and deployed in a variety of applications during the war. Detector research in the United States previous to 1940 and into the early 1940s was concentrated on Tl_2S (thallous sulfide). This effort was later dropped in favor of PbS, which had a longer spectral response and higher performance.

Beginning in the late 1940s and continuing into the 1950s, a wide variety of new materials was developed for IR sensing.[40] PbSe, PbTe, and InSb extended the spectral range beyond that of PbS in order to utilize the 3- to 5-μm wavelength (MWIR) atmospheric window. At the same time, extrinsic photoconductive response from copper, zinc, and gold impurity levels in germanium made devices possible in the 8- to 14-μm long-wavelength (LWIR) spectral window and beyond to the 14- to 30-μm very long wavelength (VLWIR) region. The end of this decade saw the first introduction of semiconductor alloys in III-V, IV-VI, and II-VI material systems. These alloys allowed the bandgap of the semiconductor, and hence its spectral response, to be custom tailored for specific applications. HgCdTe, which was introduced in the late 1950s in England,[41] has today become the most widely used of the tunable bandgap materials.

As photolithography became available in the early 1960s, it was applied to the production of IR sensor arrays. Linear array technology was first applied to PbS, PbSe, and InSb detectors. Photovoltaic (PV) detector development began with the availability of single-crystal InSb material. The discovery in the early 1960s of *extrinsic* Hg-doped germanium[42] led to the first forward-looking IR (FLIR) systems operating in the LWIR spectral window using linear arrays. Although Hg-doped germanium with a 0.09-eV activation energy was a good match to the LWIR spectral window, because it was an *extrinsic* detector it required a two-stage cooler to operate at 25 K. In the late 1960s and early 1970s, "first-generation" linear arrays of *intrinsic* HgCdTe photoconductive (PC) detectors were developed.[43] These allowed LWIR FLIR systems to operate at 80 K with a single-stage cryoengine, making them much more compact,

lighter, and requiring significantly less power consumption. PC HgCdTe linear arrays have been in high rate production for tactical applications for more than a decade.

The 1970s witnessed a mushrooming of IR applications combined with the start of high-volume production of first-generation sensor systems using linear arrays. Photoconductive HgCdTe linear array technology has been the workhorse for tactical imaging systems in the LWIR spectral region for more than a decade. Tens of thousands of these arrays have been produced. At the same time, other significant detector technology developments were taking place. Extrinsic silicon devices evolved largely to replace extrinsic germanium for many applications in the VLWIR spectral region.[44,45,b] However, germanium is still used for far-infrared (FIR) wavelengths beyond 32 μm and in selected production established in the late 1960s that used Hg as a dopant. Strained germanium detectors have recently extended the long-wavelength response of the shallowest dopant levels by about a factor of 2 from ~120 to 200 μm.[46,47] Impurity band conduction (IBC) devices[48] have eliminated the quirky behavior caused by dielectric relaxation effects in extrinsic devices under very low background conditions, while also extending the long-wavelength response of these devices.[49] Silicon technology also spawned novel platinum silicide (PtSi) detector devices,[50] which have demonstrated significant potential for a variety of MWIR high-resolution applications.

Note that for a period of several years from the late 1960s until the mid-1970s, HgCdTe alloy detectors were in serious competition with IV-VI (PbSnSSeTe) alloy devices for developing PV device prototypes.[51] Development of detectors using IV-VI compounds was discontinued in the United States because the chalcogenides suffered two significant drawbacks. The first was a high dielectric constant, which resulted in high diode capacitance and therefore limited frequency response. For scanning imaging systems under development at that time, this was a serious limitation. For staring imaging systems under development today using 2-D arrays, this would not be as significant an issue. The second drawback to PbSnTe and similar compounds was their very high thermal coefficients of expansion. This limited their applicability in hybrid configurations with silicon multiplexers. Today, with the ability to grow these materials on alternative substrates such as silicon, this too would not be a fundamental limitation. Although little or no work is currently under way in the United States on these alloys for detector applications, several European investigators have continued to pursue this technology and have made significant progress.[52]

First-generation IR imaging systems used linear arrays coupled with room temperature preamplifiers. Consequently, each detector element in the array had an individual conductive signal lead, which had to be fed through the cryogenic vacuum dewar wall. This approach limited first-generation linear

[b]Extrinsic silicon detectors can be made with significantly larger IR absorption coefficients than germanium in the LWIR and VLWIR spectral regions, due to the lower dielectric constant of silicon and the higher solubility of impurities having suitable energy levels. Device processing technology, refined in the production of silicon integrated circuits, also contributed to a preference for extrinsic detectors made from silicon. The opportunity to integrate the detector with the silicon multiplexer on the same chip was another potentially attractive feature of silicon detectors.

arrays to fewer than 200 elements. A novel British invention, the SPRITE detector,[53] extended conventional PC HgCdTe technology by incorporating signal time delay and integration (TDI) within a single elongated detector element to provide an effective 1.5-generation technology. Although only used in small arrays of ~10 elements, these devices have been produced in quantities of thousands.

The invention[54] of CCDs in the late 1960s made it possible to envision second-generation detector arrays coupled with on-focal-plane electronic analog signal readouts, which could multiplex the signal from a very large array of detectors. Early assessments of this concept showed that photovoltaic detectors such as InSb, PtSi, and HgCdTe or high-impedance photoconductors such as PbSe, PbS, and extrinsic silicon detectors were promising candidates because they have impedances suitable for interfacing with the FET input of readout multiplexers. PC HgCdTe was not suitable due to its low impedance. Therefore, in the late 1970s and through the 1980s, HgCdTe technology efforts focused almost exclusively on PV device development because of the need for low power and high impedance in large arrays in order to interface to readout input circuits. This effort is finally paying off with the birth of second-generation IR sensors, which provide large 2-D arrays in both linear formats with TDI for scanning imagers, and in square and rectangular formats for staring systems. Monolithic extrinsic silicon detectors were demonstrated[55–57] first in the 1970s. The monolithic extrinsic silicon approach was subsequently set aside because the process of integrated circuit fabrication degraded the detector-quality material properties. Monolithic PtSi detectors, however, in which the detector can be formed after the readout is processed, are now widely available. The first high-performance second-generation hybrid array demonstration was made with InSb in 1978 in a 32 × 32 array format.[58] Second-generation devices have now been demonstrated with many detector materials and device types, including PbS, PbSe, InSb, extrinsic Si bulk and IBC devices, PtSi, and PV HgCdTe.

It has taken nearly two decades since the invention of the CCD to mature the integration of IR detectors coupled with electronic readouts on the focal plane. IR image sensing technology is now in the midst of a very significant transition from first- to second-generation devices. Second-generation devices and their advantages have been convincingly demonstrated in small quantities.[59–65] The remaining issue of producibility in sufficient quantity and at low cost is now being addressed. In the longer term we anticipate that IR image sensors will evolve to incorporate new functionality analogous to the functions of the biological retina as they are interfaced with artificial neural network signal processors.

Range of Detector Performance and Sensor Choice. IR detector applications today span a broad range of requirements. A few guidelines are reviewed here about detector choice and performance specifications.

Wavelength and Temperature. Modern terminology generally defines[c] short-wavelength IR (SWIR) as 1 to 3 μm; MWIR as 3 to 5 μm; LWIR as 8 to 14 μm,

[c]Near-infrared (0.7- to 1-μm) detection, which falls within the spectral range of silicon detectors, is not addressed here. Also, the far-infrared (30- to 1000-μm) region is not addressed here, although technology described here could apply to extrinsic Ge devices with responses to 100 to 200 μm.

and VLWIR as 14 to 30 μm. Table 4.6 summarizes the spectral response coverage of many common IR detectors in four temperature regimes. These four temperature regimes are 300 K (commonly referred to as "room temperature" or uncooled), 190 K (representative of a four-stage thermoelectric cooler, freon-13, or dry ice), 80 K (liquid nitrogen, a Joule-Thompson cryostat with N_2 or Ar gas or a single-stage mechanical cooler), and 1.5 to 60 K (two- or three-stage mechanical cooler, liquid Ne, H, or He). It is evident that the alloy compound HgCdTe is the most versatile of common IR detectors with respect to spectral response coverage. This versatility comes from the ability to grow custom alloy mixtures with virtually any bandgap between 1 and 25 μm. However, when the spectral range requirement of a specific application does match one of the compound or extrinsic detector materials, they are often attractive choices.

Detectivity D and Background Flux.* Photon detectors are fundamentally limited by statistical fluctuations in the background and signal photon flux. To reach this fundamental limit, thermal fluctuation noise in the detector element must be reduced by cooling the detector. As noted above, the detection of longer wavelength, lower energy photons requires more cooling in order to avoid performance degradation due to thermal noise.[66]

If the background photon flux is reduced by narrowing the field of view, observing through a narrow-band cold filter, or by virtue of an inherent low-background flux environment such as outer space, the fundamental photon background noise is reduced and higher performance can be achieved. Additional cooling may also be required to reduce correspondingly the thermal noise under these circumstances.

Detectivity D^* is a common figure of merit for infrared detectors, which reflects the measured SNR per watt of signal flux under specific conditions, and is normalized by the statistical fluctuation sampling conditions (square root of detector area and observation bandwidth):

$$D^* = (\text{SNR})\text{cm}\sqrt{\text{Hz}}/\text{W} \ . \tag{4.197}$$

The units of D^* are commonly called *Jones*.[67,68] The conditions that must be specified include the magnitude and spectral distribution of the flux source,

Table 4.6 Detector Spectral Cutoff Range for Various Detector Materials*

	Temperature (K)			
Detector	300	190	80	1.5 to 60
PbS	3.0	3.3	3.6	—
PbSe	4.4	5.4	6.5	—
InSb	7.0	6.1	5.5	5.0
PtSi	—	—	4.8†	—
PV HgCdTe	1–3	1–5	3–12	10–16
PC HgCdTe	1–11	3–11	5–25	12–25
Extrinsic Si	—	—	—	8–32
Extrinsic Ge	—	—	—	7–200

*Spectral cutoff is defined as the long-wavelength 50% response limit except as noted by †.

such as the temperature of the blackbody (or power from a monochromatic source such as a laser), detector field of view, background temperature, chopping frequency at which the noise was measured, and wavelength at which the measurement applies. Detectivity D^* may be quoted as *blackbody*, which is the integral of the signal and background spectral characteristics convolved with the spectral response of the detector itself. Also, D^* can be quoted as *peak*, in which case the value applies only at the wavelength of maximum spectral responsivity. In other cases, D^* may be specified as average over a given spectral band, or at any other particular wavelength, such as that defined by a filter.

Photovoltaic detector performance is frequently reported in terms of the R_0A product, where R_0 is the dynamic junction impedance $(dI/dV)^{-1}$ at zero bias and A is the detector area (*caution:* this area may be the optical area or the junction area). The R_0A product relates directly to D^* in the absence of other noise sources such as $1/f$ noise or background photon noise according to the following relationship:

$$D^*_\lambda = \eta q (R_0A)^{0.5}/2E_\lambda(kT)^{0.5} \ , \tag{4.198}$$

where η is the quantum efficiency, q is the electronic charge (1.6×10^{-19}), R_0A is the resistance-area product (Ωcm^2), E_λ is the photon energy ($1.6 \times 10^{-19} \times 1.24/\lambda$), and kT is the product of Boltzmann's constant with temperature ($1.38 \times 10^{-23}\,T$). Figure 4.34 shows this theoretical relationship between D^* and R_0 At $T = 80$ K and $\eta = 0.8$ [under these conditions $D^* = 9.71 \times 10^9 (R_0A)^{0.5}\lambda$], where λ is in micrometers. The detector R_0A product, in general, depends on the photon flux present during the measurement.[69] In comparing diode performance, it is therefore important to note the background flux conditions used in the measurement.

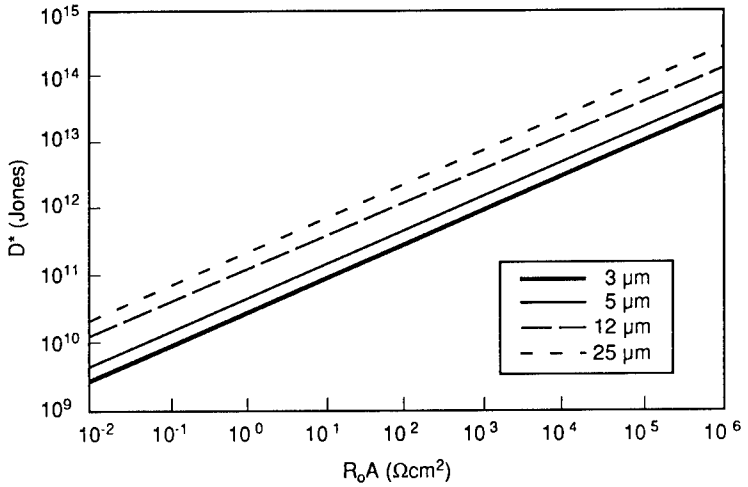


Fig. 4.34 The D^*_λ limit as a function of photodiode R_0A product for four spectral wavelengths at 80 K. Quantum efficiency of 80% is assumed. In many applications background flux will limit D^* rather than the R_0A product.

Detector Array Uniformity and Sensor Dynamic Range. Uniformity specifications apply to detector arrays, either linear or two dimensional. Uniformity is normally quoted as the standard deviation (one sigma) of detector responsivity. Other detector parameters are sometimes important with respect to uniformity, but are not discussed here. Figure 4.35 shows representative uniformity data for a variety of infrared detector materials.[70] In this figure, uniformity is also compared with a given set of limitations (diagonal lines) imposed by statistical geometrical variation in pixel dimensions. For example, if the optical area is 10^{-4} cm^2 and the critical dimensional control is 1 μm, then the uniformity is limited to at best ~2%. Such variations can occur due to mask, lithography, and/or processing variables. For some detector technologies, uniformity may already be limited by pixel dimensional control,[71] but in other cases, detector material[72] or processing variables appear to limit uniformity at the present time. Materials and processing developments are anticipated to improve the uniformity of detector arrays substantially in the future, particularly in the case of HgCdTe.

Uniformity has important consequences on the imaging system.[73,74] Correction of nonuniformity requires signal processing overhead as well as power and consumes some of the sensor's dynamic range (in a digital system some number of bits). Residual nonuniformity, which cannot be corrected ultimately, limits the SNR of the imaging system. Figure 4.36 compares two hypothetical detectors; one having high quantum efficiency (100%) and residual nonuniformity of 1%, while the second has low quantum efficiency (1%) and residual nonuniformity of 0.1%, as noted on the right-hand vertical axis. As can be seen, at low values of flux, high quantum efficiency is dominant in controlling

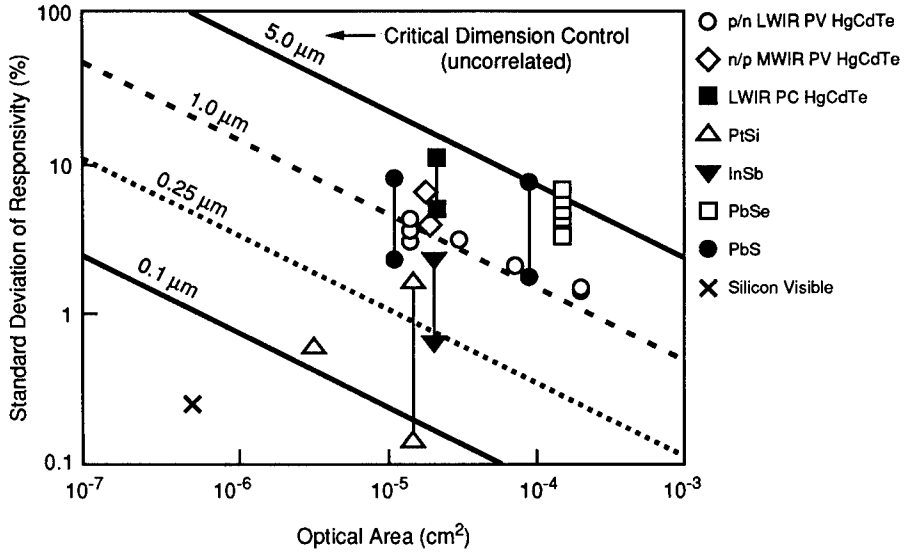


Fig. 4.35 Responsivity uniformity (one sigma standard deviation/average) for a variety of IR sensor arrays. Comparison is shown with uniformity limits imposed by pixel size variability (control of critical dimensions).

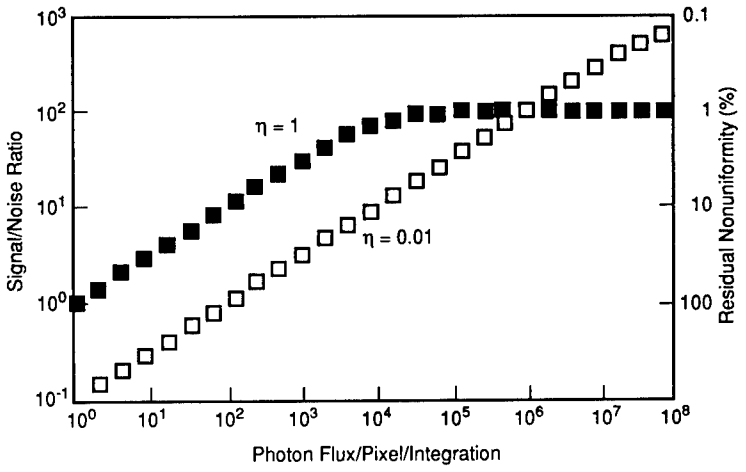


Fig. 4.36 Comparison of two hypothetical detectors: one with 100% quantum efficiency and 1% residual nonuniformity and a second with 1% quantum efficiency and 0.1% residual nonuniformity. Quantum efficiency limits SNR at low-flux levels while uniformity is the limit at high-flux levels.

the performance, while at high flux levels, uniformity plays a significant role. Note that the charge that must be stored per pixel per integration period is equal to the horizontal scale times the quantum efficiency. Readout charge storage capacity limits the upper dynamic range. Charge storage capacity itself is a complex function of the detector configuration, pixel size and spacing, and operating conditions. Some examples of the dynamic range of various IR image sensors are given in later sections.

Detector Formats and Architectures. Infrared detectors are available as single elements in circular, rectangular, cruciform, and other geometries for reticle systems, linear arrays, and 2-D arrays. For imaging applications, the imaging system and array format are related as shown in Table 4.7. Note that 2-D arrays may be used in either scanning or staring systems. When 2-D arrays are used in scanning imagers, performance enhancement is achieved by time delaying the output and integrating the signal sampled by each pixel.

Linear and 2-D arrays may be fabricated with a variety of device and signal output architectures. First-generation linear arrays are usually frontside illuminated and the detector signal output is connected by wire bonding to each element in the array. The signal from each element is then brought out of the

Table 4.7 Detector Formats Applicable to Various Imaging Systems

	Detector Format		
Imaging System	Single Element	Linear Array	2-D Array
Scanning, two axes	√	TDI	
Scanning, one axis		√	TDI
Staring			√

vacuum package and connected to an individual preamplifier prior to interfacing with the imaging system display. Gain adjustments are usually made in the preamplifier circuitry, although PC HgCdTe array packages in some cases have internal laser-trimmed load resistors to make the detector output responsivity uniform to within 1%.

Figure 4.37 illustrates the several alternative focal plane architectures for second-generation devices. Second-generation arrays, both linear and 2-D, are frequently backside illuminated through a transparent substrate. Figure 4.37(a) illustrates a detector array that is electrically connected directly to an array of preamplifiers and/or switches called a *readout*. The electrical connection is made with indium "bumps," which provide a soft metal interconnect for each pixel. This arrangement, commonly referred to as a *direct hybrid* arrangement, facilitates the interconnection of large numbers of pixels with individual preamplifiers coupled to row and column multiplexers. A PtSi direct hybrid having 480 × 640 pixels is shown in Fig. 4.38. The reliability of this detector/readout assembly under repeated cycling to cryogenic operating temperature is determined by the differential thermal coefficient of expansion between the two components and by the chip size. In the case of HgCdTe detectors on Cd(ZnSe)Te substrates, the chip size is limited to ~ 8 mm^2 for a direct hybrid

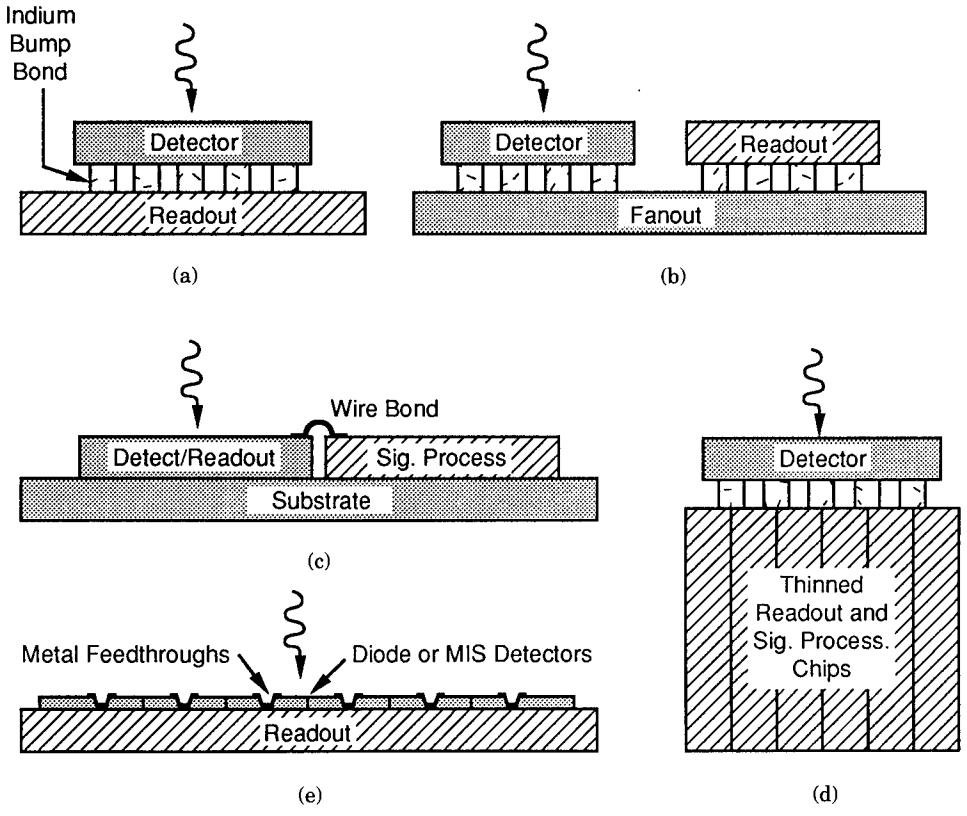


Fig. 4.37 Second-generation IR detector readout architectures: (a) direct hybrid, (b) indirect hybrid, (c) monolithic, (d) Z technology, and (e) loophole and VIMIS.

[Fig. 4.37(a)] if the readout is made from silicon. To extend direct hybrids to very large array sizes, silicon detector substrates are being developed that will have the same expansion coefficient as the silicon readout.[75]

Indirect hybrid configurations [Fig. 4.37(b)] may be used with large linear arrays (with or without TDI) to interface the detector with a substrate having a close-matching thermal coefficient of expansion; or to allow serial hybridization so that the detector may be tested prior to committing the readout and/or to accommodate readout unit cells having dimensions larger than the detector unit cell, for example, to increase the charge storage capacity and thereby extend the dynamic range. Readouts and detectors are electrically interconnected by a patterned metal bus on a fanout substrate. Figure 4.39 shows an LWIR 960 × 4 HgCdTe indirect hybrid having a single detector array and four readout arrays on such a fanout.

Monolithic detector arrays [Fig. 4.37(c)] such as IRCCDs and IRCIDs have integrated detector and readout functions, generally with command and control signal processing electronics adjacent to the detector array, rather than underneath.[76] In this case, the signal processing circuits may be connected to the detector by wire bonds. In the monolithic configuration, the signal processing circuits do not need to be on the same substrate as the detector/readout (as shown in the figure) or at the same temperature as the detector. Monolithic PtSi detector arrays can be made with signal processing incorporated on the periphery of the detector/readout chip by virtue of using silicon-based detector technology. Figure 4.40 shows a monolithic PtSi CCD array having 2048 × 16 detectors.

As illustrated in Fig. 4.37(d), Z technology[77] provides extended signal processing real estate for each pixel in the readout chip by extending the structure in the orthogonal direction. In the approach illustrated, stacked, thinned readout chips are glued together and the detector array is connected to the edge of this signal processing stack with indium. Although currently the least mature of the technologies, Z technology has potential for providing increased on-focal-plane signal processing functions.[78] Figure 4.41 shows a Z technology module with a 64 × 64 detector array.

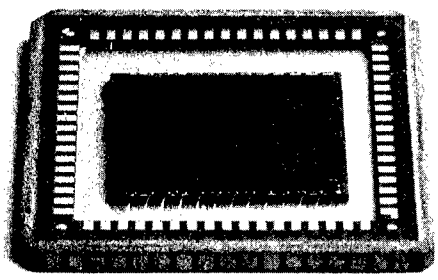


Fig. 4.38 PtSi direct hybrid.

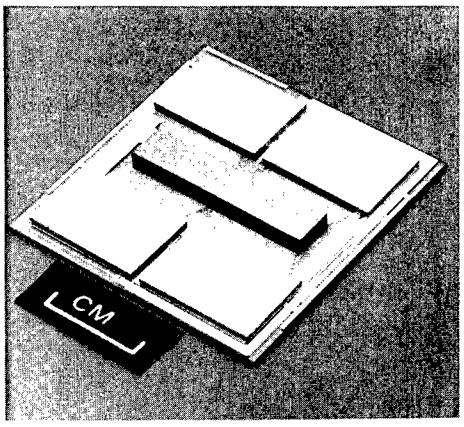


Fig. 4.39 LWIR 960 × 4 HgCdTe.

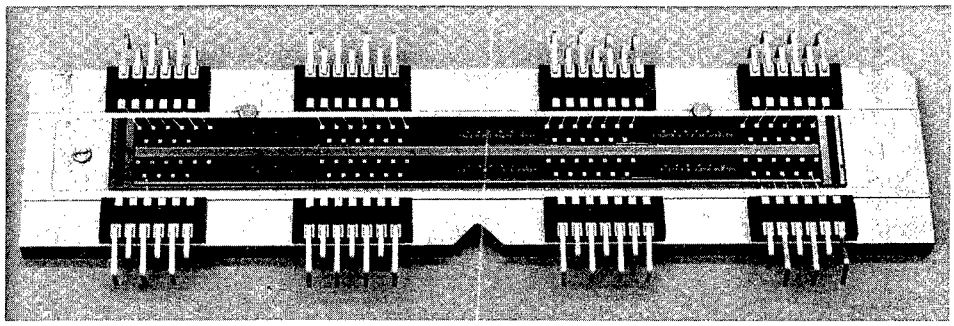


Fig. 4.40 Monolithic PtSi CCD array.

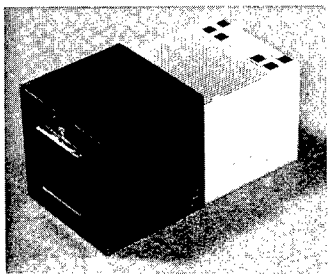


Fig. 4.41 Z-technology module.

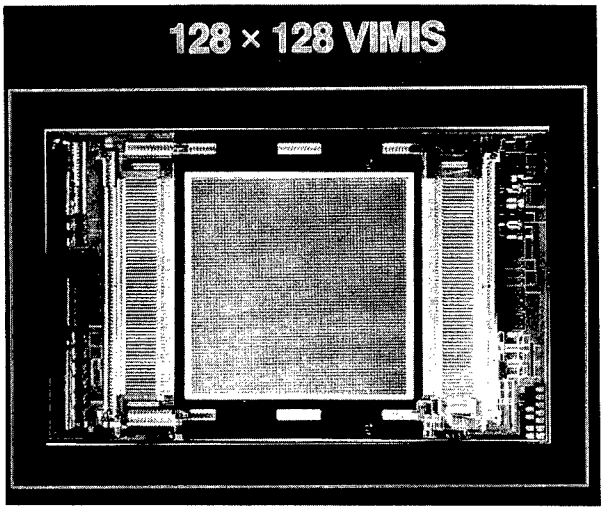


Fig. 4.42 A 128 × 128 LWIR HgCdTe VIMIS array.

Loophole,[79–83] vertically integrated metal insulator semiconductor (VIMIS),[84] and vertically integrated photodiode (VIP) approaches, as illustrated in Fig. 4.37(e), rely on thinning the detector material after adhesively bonding it to the silicon readout. Detector elements, either diodes or MIS devices, are connected to the underlying readout with vias, which are etched through the detector material to contact pads on the readout and metallized. Figure 4.42 shows a 128 × 128 LWIR HgCdTe VIMIS array.

Maturity and Cost. Detector maturity is a function of the accumulated experience and development effort, the complexity of the device required (or desired), and the inherent difficulty presented by the material technology. At the present time, PV HgCdTe and IBC extrinsic silicon detectors are not fully mature. Still, they are quite far along in terms of our ability to fabricate large 2-D arrays and successfully demonstrate them in development systems. In comparison, III-V quantum well,[85] III-V superlattice,[86,87] and superconductor[88,89]

detector technologies are several new concepts that are still in the feasibility demonstration phase.[90]

Mature detector technologies such as InSb and PtSi are still evolving significantly as applications for larger array configurations and smaller pixel sizes continue to push the technology. Other mature technologies such as PbS, PbSe, and PC HgCdTe have been significantly enhanced in their performance in the past 10 to 20 years in response to demands for higher responsivity, lower $1/f$ noise, better uniformity, and greater producibility.

The cost of an infrared image sensor is highly dependent on numerous factors, the most important being:

- D^* required compared to the background limit and margin of cooling
- requirement for detailed characterization and testing
- configuration, packaging, and specification uniqueness
- level of documentation, inspection, and testing
- array perfection.

Inexpensive sensors have nominal specifications and are tested for functionality rather than for complete parametric characterization. An exception to this rule occurs for specific configurations produced in large quantities, or variations that can utilize major portions of an existing high-volume product line.

In general, if an established technology that is already in production can do the job, it will be very cost effective to use. If the application requires a newer detector technology, such as PV HgCdTe, and is within the currently developed performance envelope, fabrication is relatively straightforward even though yield may be limited at present. Expanding the performance envelope in one or more parameters can be both exciting and expensive.

Performance and Configurations of IR Image Sensors

HgCdTe. HgCdTe detectors are available to cover the spectral range from 1 to 25 μm. Figure 4.32 illustrates representative spectral response from photovoltaic devices. The versatility of HgCdTe detector material is directly related to being able to grow a broad range of alloy compositions in order to optimize the response at a particular wavelength.

Photovoltaic detector array performance[69,75,91–100] is generally characterized in terms of the diode R_0A product. Figure 4.43 illustrates the trend of R_0A product for small diodes as a function of wavelength at 80 K. This trend is for generation-recombination-limited diodes at 5 μm and diffusion-limited diodes at the longer wavelengths. The data set in Fig. 4.43 comprises *p*-on-*n* polarity devices.[101] MWIR data for *n*-on-*p* devices is comparable to *p*-on-*n*, while for LWIR devices *n*-on-*p* polarity R_0A is typically lower.

For applications at 80 K, PV HgCdTe is generally limited to wavelengths of ~12 μm or less in order to maintain a high enough impedance to interface with on-focal-plane complementary metal-oxide semiconductor (CMOS) readouts.

For MWIR applications with spectral cutoffs of 3 to 4.7 μm, operation is possible at temperatures in the range of 175 to 220 K, which can be achieved with thermoelectric cooling. SWIR applications can operate at correspondingly higher temperatures, up to and above room temperature.

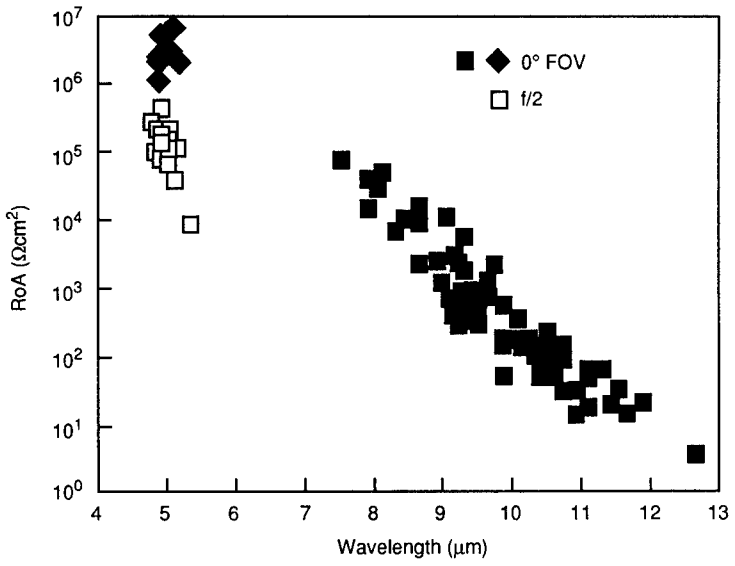


Fig. 4.43 R_0A product of HgCdTe PV devices at 80 K. Each data point represents the average R_0A product of an array of small (approximate range of 10^{-5} to 10^{-3} cm^2) p-on-n polarity diodes. The R_0A product depends on background flux and field of view for each data point as noted.

PV HgCdTe arrays have been made in linear [240, 288 (European), 480, and 960 elements), 2-D scanning with TDI, and 2-D staring formats from 32 × 32 up to 480 × 640. Pixel sizes ranging from 20 μm^2 to more than 1 mm have been demonstrated. These devices have applications for push-broom scanning systems for Landsat earth resource mapping as well as thermal imaging and search and track applications in the SWIR, MWIR, and LWIR regions. Table 4.8 lists typical performance specifications for an LWIR 2-D PV HgCdTe scanning array with four elements in TDI. Figure 4.44 shows some PV HgCdTe array formats that have been demonstrated.

Indirect and direct hybrid backside-illuminated configurations are most commonly used, although frontside-illuminated, wire-bonded devices can also be made for linear arrays. Both n-p and p-n diode polarity junction technologies have been developed, depending on the application. Liquid-phase epitaxial material is standard for most applications, although MBE and MOCVD materials are beginning to contribute to technology demonstrations and to be used in specialized products. Antireflection coatings are available as an option for all spectral bands.

PV HgCdTe laser detectors, specialized for use with CO_2 lasers at 10.6 μm and 80 K operation, are available with response speeds up to 1 GHz and higher. Performance is commonly measured in the heterodyne mode where the CO_2 laser provides the local oscillator frequency. Detector performance can be compared with the quantum efficiency limit for a heterodyne receiver under these conditions, namely[102]:

Table 4.8 Typical Performance Specifications for an LWIR PV HgCdTe Array

Array format	240 × 4
Pixel size	40 × 40 μm
Spectral response cutoff	$10.0 < \lambda < 10.5$ μm
Average D^* at 77 K and 30-deg FOV	$> 1.2 \times 10^{11}$ Jones
D^* standard deviation	< 15%
D^* defects below 0.6×10^{11} Jones	< 4 pixels
Quantum efficiency (without antireflection coating)	> 65%

$$\text{NEP} = h\nu\sqrt{B}/\eta \ , \tag{4.199}$$

where NEP is the noise equivalent power, $h\nu$ is photon energy, B is the IF amplifier bandwidth, and η is the heterodyne quantum efficiency. Experimental heterodyne quantum efficiencies exceed 50% at frequencies less than 500 MHz, and are approximately 30% at higher frequencies.[103] These detectors are useful for laser radar imagery and can be made as single elements or in small arrays. Figure 4.45 shows a four-element, gigahertz, quadrant array device indium bump bonded to a fanout.

The technology base developed for CO_2 laser detection can be readily extended to shorter wavelength laser system applications, including laser radar imaging. For applications at 1.5 μm, a unique feature of the valence band of HgCdTe, in which the split-off valence band is below the top of the valence band by the same amount as the conduction band is above the top of the valence band, leads to high-performance avalanche photodiodes (APDs).[104,105] Devices with hole-to-electron avalanche coefficient ratios (β/α) of 25 have been demonstrated, leading to low-noise (excess noise factor of ~4) APDs with high gain. Figure 4.46 illustrates dark current and photocurrent as a function of bias voltage for such a device.[106] Avalanche gain in excess of 10 is achieved at 20 V.

Photoconductive HgCdTe technology is limited at the present time to linear arrays, although custom 2-D arrays up to 10 × 10 have been made for unique applications. Production products include 12-μm cutoff arrays of 30, 60, 120, 160, and 180 elements for operation at 80 K, as well as 5-μm cutoff arrays of 30 elements for operation at 190 K. In the SPRITE configuration, linear arrays of about 10 elements are available, which provide the signal-to-noise enhancement of several detectors in a TDI function when combined with a synchronized scanned imaging system.[107] In all the PC HgCdTe linear array configurations, the signal from each detector is brought outside the dewar for preamplification and multiplexing. The development of on-focal-plane multiplexing technologies capable of handling the low impedance of photoconductive devices has not yet been demonstrated.

Significant improvements in the gain of photoconductive devices has been realized in the past decade with the development of trapping-mode detectors

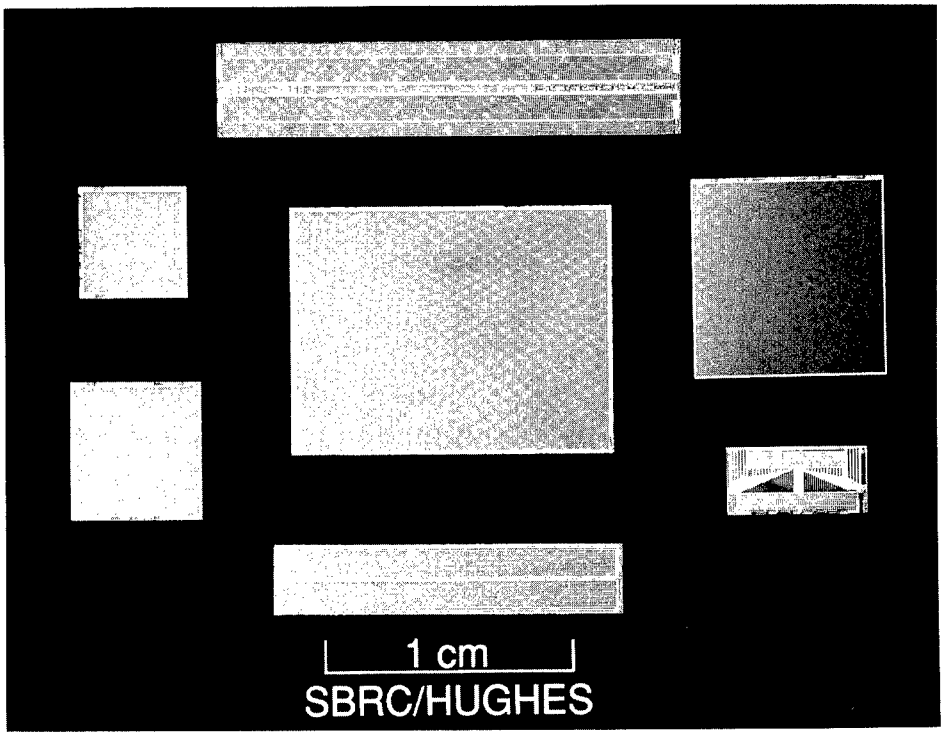


Fig. 4.44 HgCdTe arrays shown in a variety of formats. The smallest combines both 60 × 4 and 128 × 4 arrays. Scanning format arrays (with TDI) of 480 × 4 and 960 × 4 are shown along with staring arrays of 64 × 64, 128 × 128, 256 × 256, and 480 × 640 formats.

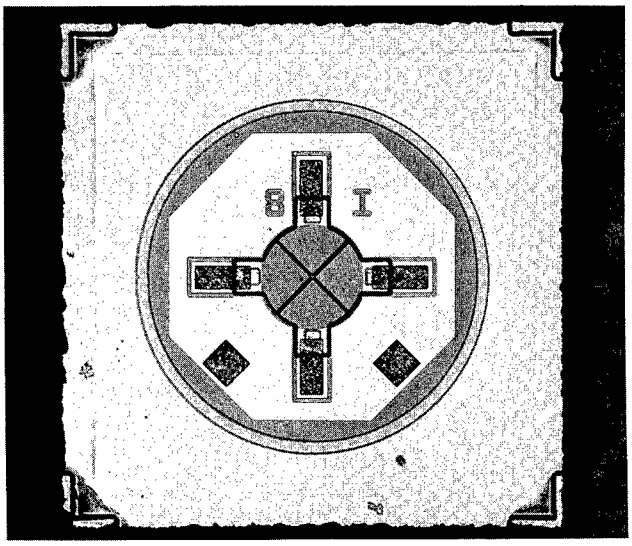


Fig. 4.45 HgCdTe CO_2 laser detector quadrant array having frequency response in excess of 1 GHz.

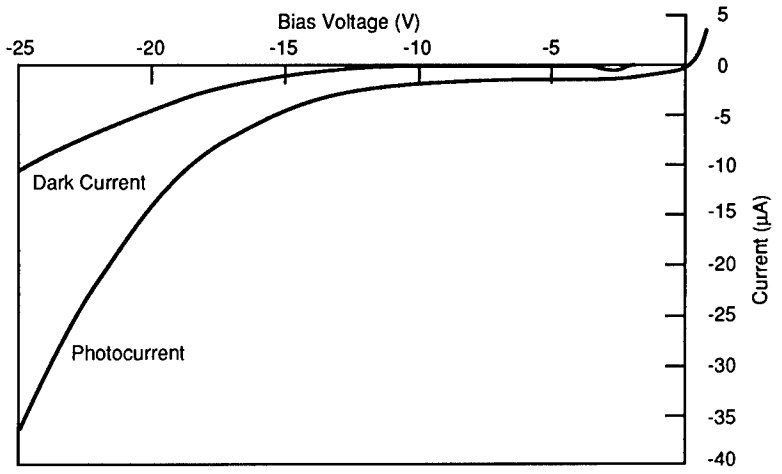


Fig. 4.46 Dark and photocurrent as a function of bias for a PV HgCdTe avalanche photodiode having peak spectral sensitivity at 1.5 μm.

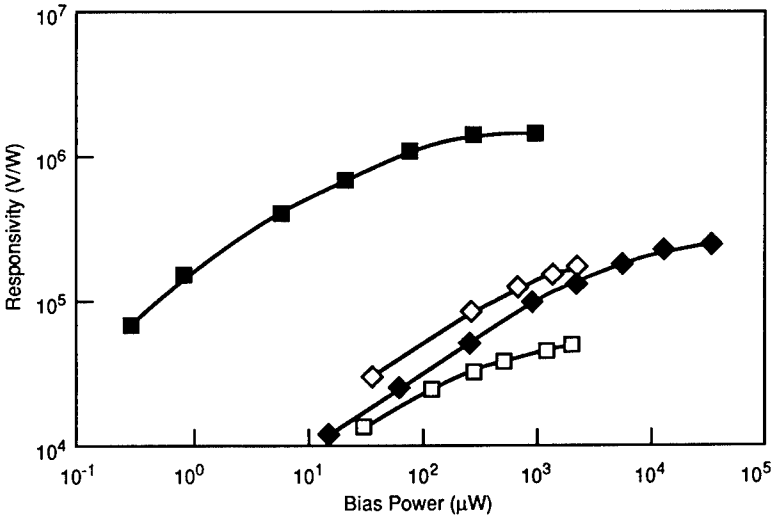


Fig. 4.47 Comparison of the bias dependence of responsivity of conventional photoconductive HgCdTe detectors (◇□♦) with trapping-mode devices (■). Data are for devices with dimensions of about 50 × 50 μm with a 12-μm spectral cutoff at 80 K. Trapping-mode devices have 3 orders of magnitude lower bias power requirements to achieve 10^5 V/W responsivity. The impedance of all devices is of the order of 100 Ω.

and detectors with blocking contacts.[108] This is illustrated in Fig. 4.47, which compares the responsivity of 12-μm cutoff conventional[109] and trapping-mode PC HgCdTe devices[110] at 80 K. The improved gain can be used to either reduce bias power and/or raise the detector noise levels so that preamplifier noise is less critical in the imaging system electronics.

A further benefit of the improved gain is that $1/f$ noise is significantly reduced. The $1/f$ noise knees, in which the $1/f$ noise component is equal to the white noise (generation-recombination noise) level, are typically 1000 Hz in ordinary PC HgCdTe devices, but only of the order of a few hundred hertz or less in their high-gain counterparts at 80 K and for $f/2$ background flux conditions.[111]

Both conventional and trapping-mode PC HgCdTe devices achieve[112] D^* performance at about 80% to 90% of the background-limited value at 80 K, with $f/2$ conditions and with spectral response out to about 12 to 13 μm.

Figure 4.48 shows the spectral characteristics of several PC HgCdTe detectors. At 80 K the response of the HgCdTe photoconductors can be extended to as much as 25 μm. Of course, longer wavelength devices are not background limited to as low a background as devices having shorter wavelength response at 80 K. At lower temperatures the performance of longer wavelength devices improves as the thermal noise is reduced, at least as long as the detector noise remains greater than the preamplifier noise.

Standard linear arrays are made in a frontside-illuminated configuration with wire-bonded leads. However, backside-illuminated configurations have been demonstrated. Detector elements are typically antireflection coated with quarter-wave ZnS.

Photoconductive laser detectors for 10.6-μm applications are available for operation at both room temperature and with thermoelectric cooling. These detectors feature bandwidths in the range of 50 to 100 MHz. Heterodyne quantum efficiencies of 0.5% to 4% for operation at 180 K can be produced. As with photovoltaic laser detectors, these devices may be used to generate laser radar

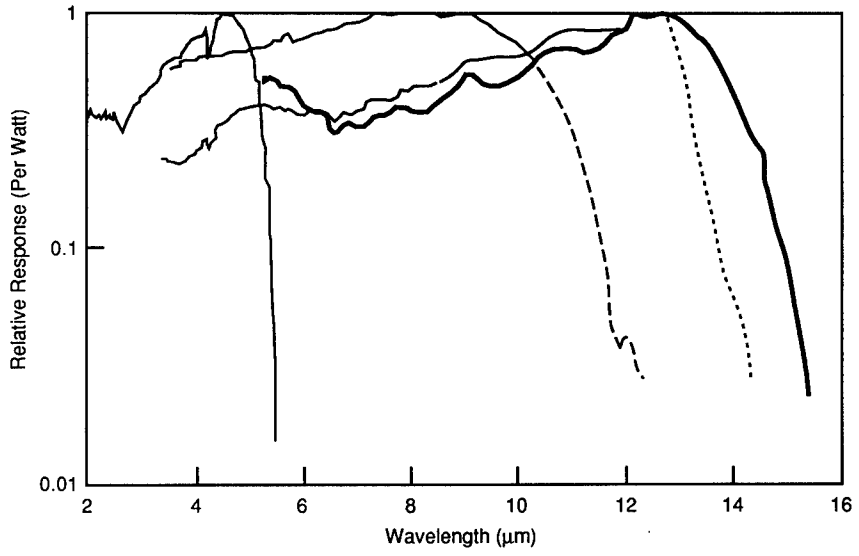


Fig. 4.48 Spectral response examples of PC HgCdTe. The three longest cutoff devices were measured at 80 K and the shorter cutoff sample at 194 K. Samples are antireflection coated, which results in some spectral structure at the shorter wavelengths.

imagery, which can be processed to reveal features along with object range and velocities.

PtSi. PtSi detectors offer the largest IR image sensor formats currently available. A partial list of configurations demonstrated include square formats of 128 × 128, 256× 256, 512 × 512, and 1024 × 1024; rectangular formats of 244 × 320, 280 × 340, 244 × 512, and 480 × 640; and long linear arrays of 2048 and 4096 elements with 16 and 4 TDI columns, respectively.[113] Both monolithic[114] and hybrid[115] PtSi array configurations are made, with the hybrid structure offering a nearly 100% fill factor, while the monolithic designs are generally limited to ~30% to 55% fill factor for small pixel dimensions. Figure 4.40 shows a monolithic PtSi CCD array in a scanning format having 2048 × 16 elements. Figure 4.38 shows a direct hybrid PtSi array in a staring format with 488 × 640 elements on a 20-μm pitch. The combination of large array formats and excellent array responsivity uniformity makes PtSi attractive for a variety of high-background-flux applications.[116–118]

The spectral response or quantum efficiency of PtSi detectors is unusual and related to the photodetection mechanism. Infrared photons energize electrons from the PtSi layer, which then have a probability of tunneling through the PtSi-Si Schottky barrier.[119] Since the tunneling probability is an exponential function of the photon energy, the spectral response or quantum efficiency decays exponentially with wavelength before falling off more steeply as the cutoff threshold is approached on a "per photon" scale. Figure 4.49 illustrates this characteristic in the quantum efficiency.[120] As a consequence, quantum efficiency is quite low for PtSi in the 4- to 5-μm spectral region,

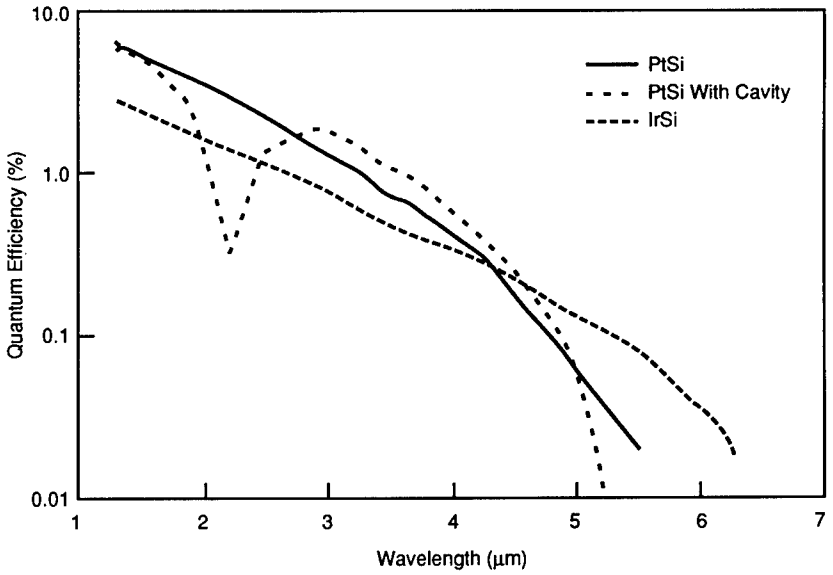


Fig. 4.49 Quantum efficiency for PtSi and IrSi detectors as a function of wavelength. A reflecting cavity peaks the response of PtSi over the spectral band of 3.0 to 4.5 μm in one PtSi example. IrSi extends the spectral sensitivity to longer wavelengths and requires a lower operating temperature.

typically of the order of 0.1% to 1%. Imagery is nevertheless very good under high-background conditions due to the large number of pixels available, combined with the excellent operability and uniformity of PtSi. Responsivity uniformity one-sigma values as low as 0.2% have been reported. A number of PtSi camera systems are now available commercially. Specifications for typical PtSi imaging arrays are summarized in Table 4.9.

PtSi detectors need to be cooled to at least about 80 K, and 70 K is even better. This is compatible with both single-stage closed-cycle coolers and pour-filled liquid N_2 dewar operation.

Development of longer wavelength silicide detectors is under way to extend the sensitivity of silicide devices into the 8- to 12-μm spectral window. IrSi devices have been explored for this purpose,[121] as shown in Fig. 4.49. IrSi devices require correspondingly lower operating temperatures. Other approaches to longer wavelength silicide detectors are also under development.

InSb. Photovoltaic InSb remains a popular detector for the MWIR spectral band at 80 K. Its spectral response at 80 K is shown in Fig. 4.50. InSb material is highly uniform and, combined with a planar-implanted process in which the device geometry is precisely controlled, the resulting detector array responsivity uniformity is good to excellent. Devices are usually made with a *p-n* diode polarity using diffusion or ion implantation.[122] Staring arrays of backside-illuminated, direct hybrid InSb detectors in 58 × 62, 128 × 128, 200 × 200, 256 × 256, and 640 × 480 formats[123–127] are available with readouts suitable for both high-background *f*/2 operation and for low-background astronomy applications. Specifications for astronomy array devices are summarized in Table 4.10. Linear array formats of 64 and 128 elements[125,127] are produced with frontside-illuminated detectors for both high-background and astronomy applications as well. Linear and 2-D arrays based on charge injection devices have also been developed[128] in InSb.

Other array configurations, both front and backside illuminated for staring and scanning applications can be made as desired. Figure 4.51 shows a variety of both scanning and staring InSb products. High-efficiency quarter-wavelength

Table 4.9 Typical Performance of Hybrid PtSi Arrays at 77 K

	Configuration	
	256 × 256	488 × 640
Elements	65,536	312,320
Spacing (μm)	30	20
Fill factor (%)	>88	>80
Emission factor	>0.3	>0.3
Responsivity (mV/K at *f*/2, 60 fps)	>10	>10
Operability (%)	≥96	≥96
Dynamic range (dB)	64	64
Noise floor (electrons)	≤200	≤200
Frame rate (frames/s)	60	60
NEΔT (°C at *f*/2)	<0.09	<0.09

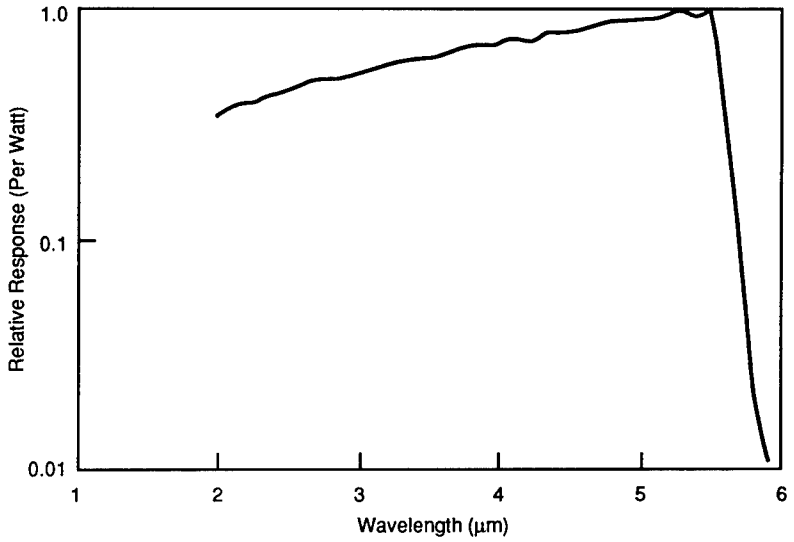


Fig. 4.50 Spectral response of InSb at 80 K. Quantum efficiency without antireflection coating as shown is approximately 65%. With antireflection coating, the quantum efficiency can exceed 90% over a portion of the spectral band.

Table 4.10 Typical Performance of InSb Astronomy Arrays at 50 K and Background Flux Levels Less Than 10^9 photons/cm^2 s^{-1}

	Configuration	
	58 × 62	256 × 256
Elements	3596	65,536
Spacing (μm)	76	30
Fill factor (%)	>90	>90
Peak quantum efficiency (%)	>90	>90
Dark current (fA)	≤2.5	≤1
NEP (aW) at 3 μm, 100 s at 2.2 μm, 1 s	≤10	≤20
Operability (%)	≥96	≥96
Integration capacity (q)	10^6	5×10^5
Mean readout noise (q) at 260-ms integration at 1-s integration	≤400	≤75

Table 4.11 Common Impurity Levels Used in Extrinsic Si IR Detectors. Operating Temperature Depends on Background Flux Level

Impurity	Energy (meV)	Cutoff (μm)	Temp (K)
Indium	155	8	40–60
Bismuth	69	18	20–30
Gallium	65	19	20–30
Arsenic	54	23	13
Antimony	39	32	10

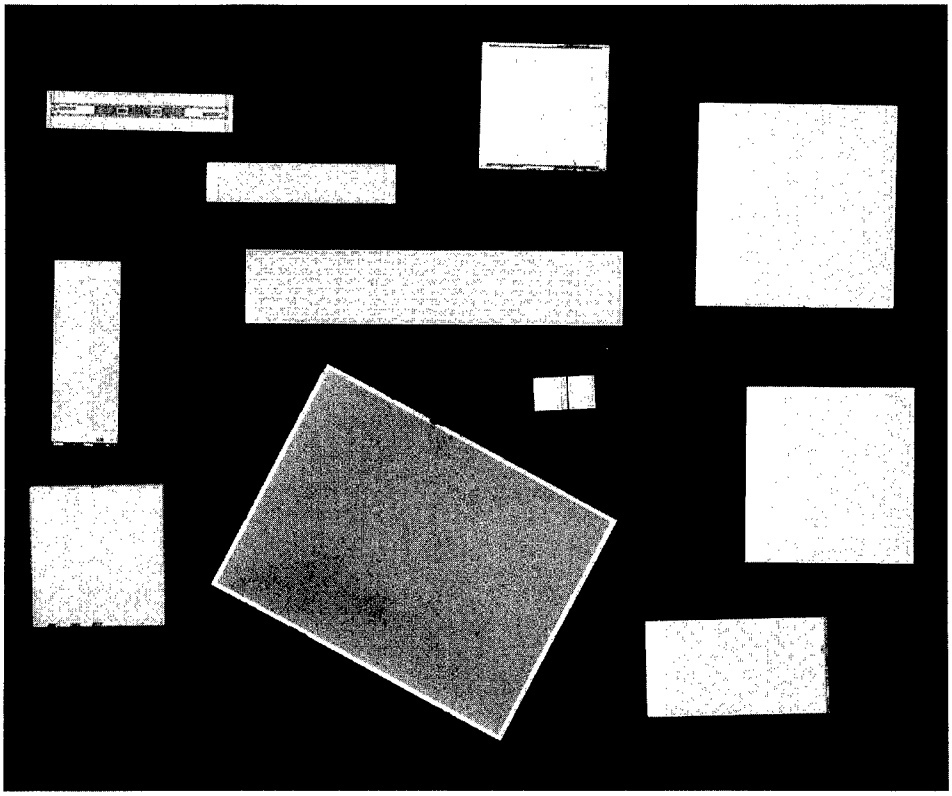


Fig. 4.51 InSb is produced in a wide variety of array configurations for both scanning and staring applications. Shown here are scanning and staring formats including multiple 16-element linear, 16 × 2, 64 × 12 × 2, 64 × 16 × 2, 128 × 5 × 2, 134 × 30, 58 × 62, 128^2, 256^2, and 480 × 640. Element sizes range from 20 × 20 to 200 × 200 μm.

antireflection coating is available to minimize reflection loss over a portion of the spectral response band for specific applications.

Since the spectral response of InSb shifts to longer wavelengths as the temperature increases, thermally generated noise increases rapidly with higher operating temperature for InSb devices. Nevertheless, operation up to at least 145 K is possible at high-background-flux levels, making these devices useful for satellite applications such as Landsat, which rely on radiative coolers.

Extrinsic Silicon Detectors. Extrinsic silicon detectors rely on photoexcitation of impurity levels within the bandgap of silicon. The spectral response of the detector depends on the energy level of the particular impurity state and the density of states as a function of energy in the band to which the bound charge carrier is excited. Table 4.11 lists some of the common impurity levels and the corresponding long-wavelength cutoff of the extrinsic silicon detector based on them. Note that the exact long-wavelength spectral cutoff is a function of the impurity doping density, with higher densities giving slightly longer spectral

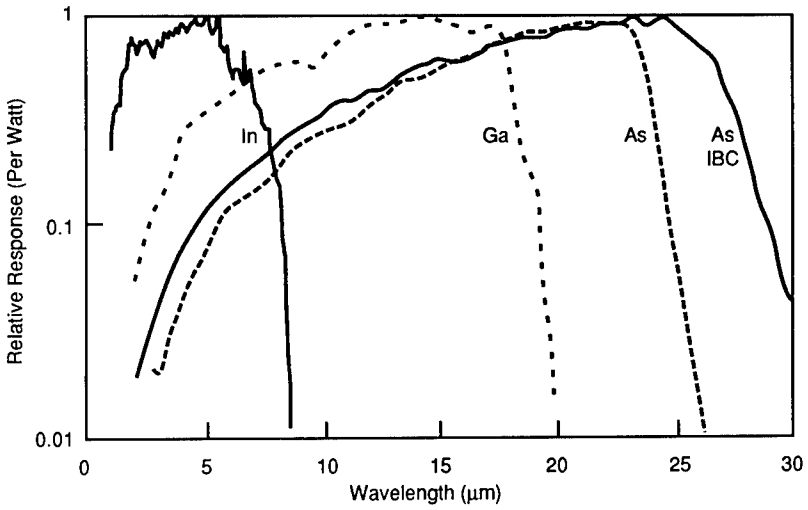


Fig. 4.52 Examples of extrinsic silicon detector spectral response. Shown are Si:In, Si:Ga, and Si:As bulk detectors and a Si:As IBC device.

response. Figure 4.52 illustrates the spectral response for several extrinsic silicon detectors. The longer spectral response of the IBC Si:As device compared with the bulk Si:As device is due to the higher doping level in the former, which reduces the binding energy for an electron.

The performance of extrinsic silicon detectors is generally background limited with a quantum efficiency that varies with the specific dopant and dopant concentration, wavelength, and device thickness. Typical quantum efficiencies are in the range of 10% to 50% at the response peak.

Extrinsic silicon detectors operate either as high-impedance photoconductors[129] or as IBC devices. Both differ from intrinsic photoconductive devices. Conventional extrinsic photoconductors have only one type of mobile charge carrier, unlike intrinsic devices. The low impedance of the intrinsic photoconductive devices results in space charge neutrality being relatively easy to maintain, and hence the excess distribution of both electrons and holes moves in one direction under an applied bias. Space charge neutrality can be violated in high-impedance photoconductors, which can lead to unusual effects as noted later. IBC detectors are unique in having some properties similar to ordinary extrinsic photoconductors, such as photoexcitation from impurity levels in the bandgap, but for also being able to collect both types of carriers, namely, those in the continuum and those in the "hopping" impurity band. This property gives the IBC devices some aspects of photovoltaic detectors, namely, reduced recombination noise.

The high impedance of conventional extrinsic photoconductors can lead to peculiar dielectric relaxation effects in these devices.[130,131] This occurs when the dielectric relaxation time $\tau = \varepsilon\varepsilon_0\rho$ becomes longer than the charge-carrier lifetimes,[132–138] where ρ is the resistivity of the sample, and $\varepsilon\varepsilon_0$ for silicon is 1×10^{-12} s/Ω cm. IBC detectors have been developed to overcome this

problem.[48,139–141] In these devices, the photosensitive layer is heavily doped so that hopping-type conduction[142] occurs and the impedance remains low enough so that dielectric relaxation can occur in a short time. The hopping current must be blocked before it reaches the device electrode to prevent excess noise in the detector. This is accomplished with a thin, lightly doped layer of silicon material.[48]

Extrinsic silicon detectors are frequently cooled with liquid He for applications such as ground- and space-based astronomy.[143–145] Closed-cycle two- and three-stage refrigerators are available for use with these detectors for cooling to 20 to 60 and 10 to 20 K, respectively.

Extrinsic silicon detector arrays have been made in 58 × 62 element formats for low-background astronomy applications.[146,147] Table 4.12 lists the specifications of Ga-doped detectors in this format. Other impurity dopants such as Sb or As can be substituted for Ga. Both linear scanning and 2-D arrays can readily be produced. Linear arrays more than 2.5 cm in length have been demonstrated.

IBC As-doped arrays in a 10 × 50 format have been reported.[139] These have demonstrated excellent responsivity uniformity with standard deviations as low as 1.5%. At a background flux level of 5×10^{12} photons/cm^2 s^{-1} these arrays achieved average D^* values of 6.7×10^{12} Jones. Table 4.13 summarizes the performance of these IBC detectors. Figure 4.53 shows a 2 × 192 × 6 arsenic-doped IBC detector.

The IBC structure can be specially doped and biased to achieve operation as a solid-state photomultiplier.[148] In this mode in which photoexcited photons are amplified by impact-ionization of impurity-bound carriers, individual photons can be counted at low flux levels. The solid-state photomultiplier has been found[149] to respond to photons across both the intrinsic and extrinsic spectral ranges, from 0.4 to 28 μm. Only test chips that include small linear array configurations of 10 elements have been reported.[150]

PbS and PbSe. PbS and PbSe detector materials may be chemically deposited as polycrystalline thin films on insulating substrates.[151] Both are employed

Table 4.12 Typical Performance of Si:Ga Astronomy Arrays at 4 K and Background Flux Levels Less Than 10^9 photons/cm^2 s^{-1}

	Configuration 58 × 62
Elements	3596
Spacing (μm)	76
Fill factor (%)	> 90
Peak responsivity (A/W)	> 1
Dark current (fA)	≤ 0.1
NEP (aW) at 15 μm, 1 s	≤ 40
Operability (%)	≥ 98
Integration capacity (q)	~2×10^6
Mean readout noise (q)	≤ 300

Table 4.13 Performance of Si:As IBC Arrays at 12 K and Background Flux Levels of 10^{12} photons/cm^2 s^{-1}

	Configuration 10 × 50
Elements	500
Responsivity (A/W at 10.6 μm)	> 3
Dark current (pA)	≤ 55
D^*	> 5 × 10^{12}
Operability (%)	≥ 98
Integration capacity (q)	> 2 × 10^6
Mean readout noise (q)	≤ 180
Integration time (μs)	62 – 3 × 10^8

as photoconductors and can operate at any temperature between 300 and 77 K. Table 4.6 summarizes how the spectral response long-wavelength cutoff of these materials shifts with temperature. Figure 4.54 illustrates typical spectral response curves for PbS at 193 K and PbSe at 77 K.

Depending on operating temperature, background flux, and chemical additives, the detector impedance per square can be adjusted in the range of 10^6 to 10^9 Ω/square. As shown in Fig. 4.35, the responsivity uniformity of PbS and PbSe is generally of the order of 3% to 10%. When sufficiently cooled to eliminate thermal noise, the D^* performance of PbS and PbSe at high-background-flux levels comes within about a factor of 2 of the background limit, implying a quantum efficiency of about 30%. Quantum efficiency is probably limited by incomplete absorption of the incident flux in the relatively thin (1- to 2-μm) detector material deposited by the chemical process.

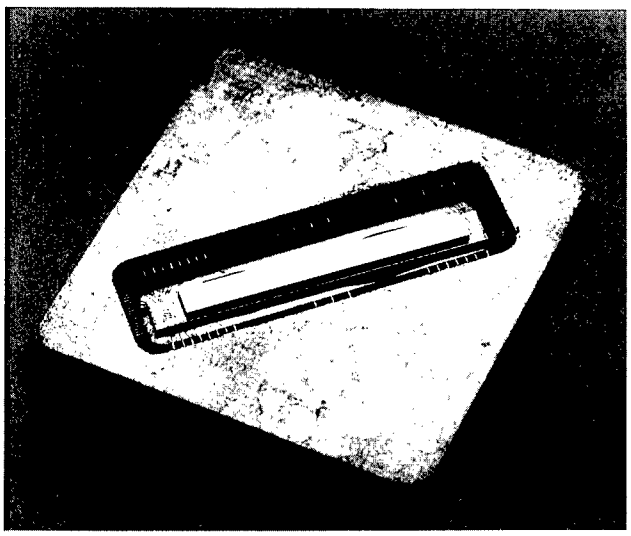


Fig. 4.53 An arsenic-doped silicon IBC direct hybrid array in a format having two subarrays of 192 × 6 pixels each.

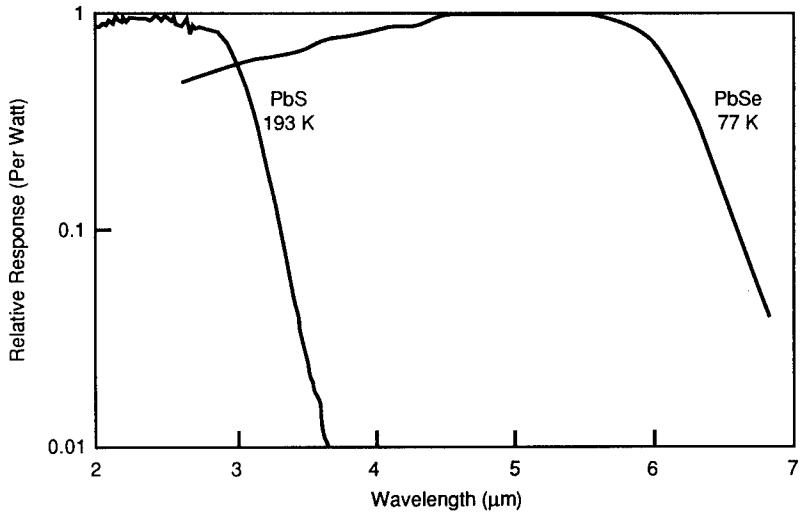


Fig. 4.54 Spectral response of a PbS detector at 193 K and a PbSe detector at 77 K. The spectral response for these materials varies with temperature as indicated in Table 4.6.

PbS and PbSe arrays have been made in a variety of linear array formats for use in focal planes not having cooled readouts. Operability of 99% to 100% is readily achieved for these arrays having 100 or fewer elements. Large arrays of between 1000 and 2000 elements on a single substrate have also been produced with operability exceeding 98%.

The high impedance of PbS and PbSe photoconductive devices allows them to be interfaced with CMOS readout circuits. Linear array formats with CMOS readouts are currently available in 64-, 128-, and 256-element configurations,[152] as illustrated in Fig. 4.55. Table 4.14 summarizes the performance of PbSe arrays in these configurations.[153]

Figure 4.56 illustrates typical peak D^* values[154] for PbSe. Note that the decrease in D^* between 150 and 77 K for the high-background condition occurs because the background flux increases substantially as the spectral cutoff moves to longer wavelengths with cooling. PbSe has significant $1/f$ noise, with a knee frequency of the order of 300 Hz at 77 K, 750 Hz at 200 K, and 7 kHz at 300 K. This generally limits this material to use in scanning imagers.

Table 4.14 Typical Performance of PbSe Linear Array with CMOS Multiplexed Readout

	Configuration 64, 128, 256 Linear
Pixel size (μm)	38 × 56
Spacing (μm)	51
D^* (peak, 1400 Hz) (Jones)	$> 3 \times 10^{10}$
Operability (%)	≥ 98
Dynamic range	2000
Uniformity	< 20%

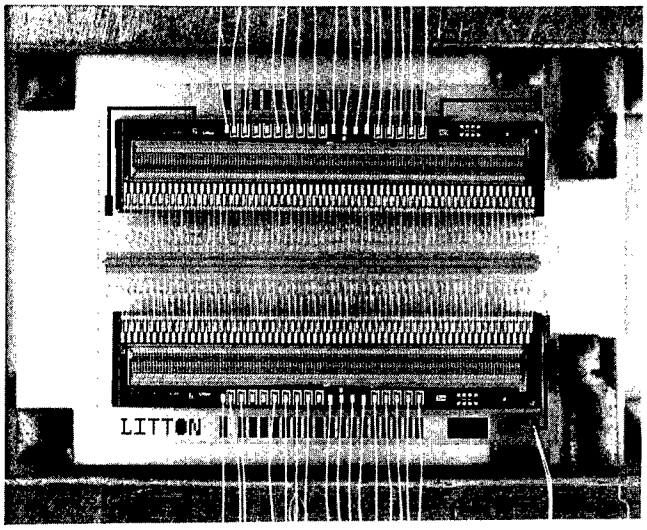


Fig. 4.55 PbSe focal plane in a 256 × 1 format with dual readouts. (Courtesy of Litton Electron Devices)

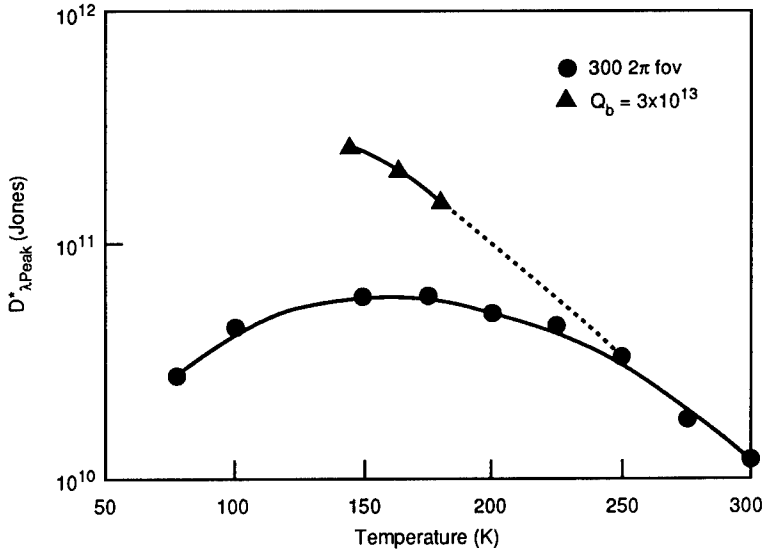


Fig. 4.56 Detectivity of PbSe at two background flux levels as a function of temperature. Noise was measured at 560 Hz for the low-background conditions and at the 1/f knee frequency at the high-background condition.

4.6 CONCLUSION

Second-generation infrared image sensors have been under intensive development for nearly two decades since the invention of the CCD. Two-dimensional arrays of high-performance detectors have now been demonstrated in a variety of scanning and staring formats using PV HgCdTe, PtSi, InSb, extrinsic silicon, PbS, and PbSe. A major effort is under way at the present time to address the issues of producibility, production readiness, and cost of these sensors in high-volume production.

Acknowledgments

This chapter could not have been written without extensive help from colleagues at Santa Barbara Research Center (SBRC) and in the industry. Some of these include Ken Ando and Bill Rogatto for inspiring the work; Thomas Johnson for references on the early history of IR detectors; William Radford for data on LWIR PV HgCdTe; Mark Langell for data on MWIR HgCdTe; Peter Bratt for data on HgCdTe laser detectors; Michael Jack for data on HgCdTe avalanche photodiodes; Michael Moroz, Chris Tacelli, and Chao Huang for data on PC HgCdTe; James Gates, Walter Kosonocky, Jon Mooney, and Edward Nelson for data on PtSi; John Toman and David Randall for InSb data; Grant Albright, Ramesh Bharat, and John Sheppard for extrinsic silicon data; and Walter Murphy for PbS data. Photographs were supplied through the courtesy of Irvine Sensors (John Carson), Texas Instruments (Julie England and Grady Roberts), Eastman Kodak (Edward Nelson), Litton Electron Devices (James Kreider), and Hughes Aircraft/SBRC (Peter Bratt, George Domingo, James Gates, James Myrosznyk, John Toman, and Devin Walsh).

References

1. R. Clark Jones, D. Goodwin, and G. Pullan, "Standard procedures for testing infrared detectors and for describing their performance," AD 257597, Office of Defense and Development Research and Engineering, Washington, D.C., pp. 1–45 (Sep. 1960).
2. R. A. Smith, F. E. Jones, and R. P. Chasmar, *The Detection and Measurement of Infrared Radiation,* Clarendon Press, Oxford, England (1968).
3. M. J. E. Golay, "Theoretical considerations in heat and infra-red detection, with particular reference to the pneumatic detector," p. 347, and "A pneumatic infra-red detector," p. 357, *Review of Scientific Instruments* **18**(5) (May 1947). Also M. J. E. Golay, "The theoretical and practical sensitivity of the pneumatic infra-red detector," *Review of Scientific Instruments* **20,** 816 (1949).
4. E. H. Putley, "The pyroelectric detector," in *Semiconductors and Semimetals,* R. K. Willardson and A. C. Beer, Eds., Academic Press, New York, Vol. 5, pp. 259–285 (1970).
5. R. Havens, "Theoretical comparison of heat detectors," *Journal of the Optical Society of America* **36,** 355 (1946).
6. R. D. Hudson, *Infrared System Engineering,* John Wiley & Sons, New York, p. 357 (1969).
7. R. L. Petritz, "Fundamentals of infrared detection," pp. 1459–1467; also G. R. Pruett and R. L. Petritz, "Detectivity and preamplifier considerations for indium antiomonide photovoltaic detectors," *Proceedings of the IRE* **47,** 1524–1529 (Sep. 1959).
8. R. H. Haitz, A. Goetzberger, R. M. Scarlett, and W. Schockley, *Journal of Applied Physics* **34,** 1581 (1963).
9. R. J. McIntyre, "Multiplication noise in uniform avalanche diodes," *IEEE Transactions on Electron Devices* **13,** 164 (1966).

10. T. S. Moss, G. J. Burrell, and B. Ellis, *Semiconductor Opto-Electronics,* John Wiley & Sons, New York (1973).

11. C. J. Summers and K. F. Brennan, "Variably spaced superlattice energy filter, a new device design concept for high-energy electron injection," *Applied Physics Letters* **48**(12), 806–808 (Mar. 24, 1986).

12. C. J. Summers and K. F. Brennan, "New resonant tunneling superlattice avalanche photodiode device structure for long-wavelength infrared detection," *Applied Physics Letters* **51**(4), 276–278 (July 27, 1987).

13. K. F. Brennan and C. J. Summers, "The variably spaced superlattice energy filter quantum well avalanche photodiode: a solid-state photomultiplier," *IEEE Journal of Quantum Electronics* **QE-23**(3), 320–327 (Mar. 1987).

14. B. F. Levine, C. G. Bethea, G. Hasnian, V. O. Shen, E. Pelve, R. R. Abott, and Hseih, "High sensitivity low dark current 10 μm GaAs quantum well infrared photodetectors," *Applied Physics Letters* **56**(9), 851–853 (Feb. 26, 1990).

15. G. Vemuri and R. Roy, "Super-regenerative laser receiver: transient dynamics of a laser with an external signal," *Physics Review A* **39**(5), 2539 (Mar. 1, 1989).

16. D. J. Lovell, "Some early lead salt detector developments," AFOSR Report 68-0264; "The development of lead salt detectors," *American Journal of Physics* **37,** 467–478 (1969); D. J. Lovell, "Pioneers in infrared detection," *Optical Spectra,* pp. 62–63 (Apr. 1974); H. Levinstein, "Infrared detectors," *Physics Today,* 23–28 (Nov. 1977).

17. I. Littler, S. Balle, K. Bergmann, G. Vemuri, and R. Roy, "Detection of weak signals via the decay of an unstable state: Initiation of an injection seeded laser," to be published.

18. W. L. Wolfe, "Photon number D* figure of merit," *Applied Optics* **12**(3), 619–621 (Mar. 1973).

19. K. M. Van Vliet, "Noise in semiconductors and photoconductors," *Proceedings of the IRE* **46,** 1004 (1958).

20. P. W. Kruse, L. D. McGlauchlin, and R. B. McQuisten, *Elements of Infrared Technology,* John Wiley & Sons, New York (1962).

21. J. B. Johnson, "Thermal agitation of electricity in conductors," *Physical Review* **32,** 97 (1928).

22. A. van der Ziel, "Noise in junction transistors," *Proceedings of the IRE* **46,** 1019 (1958).

23. R. L. Williams, "Speed and sensitivity limitations of extrinsic photoconductors," *Infrared Physics* **9,** 37–40 (1969).

24. S. R. Borrello, "Detection uncertainty," *Infrared Physics* **12,** 267–270 (1972).

25. J. A. Jamieson, "Preamplifiers for nonimage-forming infrared systems," *Proceedings of the IRE* **47,** 1522 (1959).

26. P. C. Caringella and W. L. Eisenman, "System for low-frequencies noise measurements," *Review of Scientific Instruments* **33,** 654 (1962).

27. R. F. Potter, J. M. Pernett, and A. B. Naugle, "The measurement and interpretation of photodetector parameters," *Proceedings of the IRE* **47,** 1503 (1959).

28. A. E. Martin, *Infrared Instrumentation and Techniques,* Elsevier, New York (1966).

29. W. L. Eisenman, R. L. Bates, and J. D. Merriam, "Black radiation detector," *Journal of the Optical Society of America* **53,** 729 (1963).

30. W. L. Eisenman and R. L. Bates, "Improved black radiation detector," *Journal of the Optical Society of America* **54,** 1280 (1964).

31. R. Stair, W. E. Schneider, W. R. Walters, and J. K. Jackson, "Some factors affecting the sensitivity and spectral response of thermoelectric (radiometric) detectors," *Applied Optics* **4,** 703 (1965).

32. *Properties of Photoconductive Detectors,* NOLC Report 564, Naval Ocean Systems Center, San Diego (a continuing series begun 30 June 1952).

33. Data Sheet No. 698-A, Cell D52-9P, PbSe, Infrared Industries, Santa Barbara, CA (Feb. 1961).

34. R. B. Emmons, Sylvania Electronics System, Western Division, Mountain View, CA, private communication (1970).

35. A. F. Milton, Institute for Defense Analyses, Arlington, VA, private communication (1970).

36. M. M. Blouke, C. B. Burgett, and R. L. Williams, "Sensitivity limits for extrinsic and intrinsic infrared detectors," *Infrared Physics* **13**(1), 61–72 (Jan. 1973).

37. M. A. Kinch and S. R. Borrello, "0.1 eV HgCdTe photodetectors," *Infrared Physics* **15**(2), 111–124 (May 1975).

38. N. Sclar, G. J. Hoover, W. C. Milo, and R. L. Pierce, Rockwell International, Anaheim, CA, private communication (1974).

39. H. Macurda and R. Baxter, Philco-Ford Corporation, Aeronutronic Division, Newport Beach, CA, private communication (1974).

40. Many relevant papers were published in the proceedings of the first two international conferences on photoconductivity: R. G. Breckenridge et al., Eds., *Proceedings First International Photoconductivity Conf.*, John Wiley & Sons, New York (1956); H. Levinstein, Ed., *Proceedings Second International Photoconductivity Conf.*, Pergamon Press, New York (1962).

41. W. D. Lawson, S. Nielson, E. H. Putley, and A. S. Young, "Preparation and properties of HgTe and mixed crystals of HgTe-CdTe," *J. Phys. Chem. Solids* **9,** 325–329 (1959).

42. S. Borrello and H. Levinstein, "Preparation and properties of mercury-doped germanium," *Journal of Applied Physics* **33,** 2947–2950 (1962).

43. The photoconductive and photovoltaic detector technology of HgCdTe is summarized in the following references: D. Long and J. L. Schmidt "Mercury-cadmium telluride and closely related alloys," in *Semiconductors and Semimetals,* R. K. Willardson, A. C. Beer, Eds., Academic Press, New York, Vol. 5, pp. 175–255 (1970); R. A. Reynolds, C. G. Roberts, R. A. Chapman, and H. B. Bebb, "Photoconductivity processes in 0.09 eV bandgap HgCdTe," in *Proceedings of the Third International Conference on Photoconductivity,* Stanford, California, August 12–15, 1969, E. M. Pell, Ed., Pergamon Press, New York, p. 217 (1971); P. W. Kruse, D. Long, and O. N. Tufte, "Photoeffects and material parameters in HgCdTe alloys," in *Proceedings of the Third International Conference on Photoconductivity,* Stanford, California, August 12–15, 1969, E. M. Pell, Ed., Pergamon Press, New York, p. 223 (1971); R. M. Broudy and V. J. Mazurczyk, "(HgCd)Te photoconductive detectors," Chap. 5 in *Semiconductors and Semimetals,* R. K. Willardson, A. C. Beer, Eds., Academic Press, New York, Vol. 18, pp. 157–199 (1981); M. B. Reine, A. K. Sood, and T. J. Tredwell, "Photovoltaic infrared detectors," Chap. 6 in *Semiconductors and Semimetals,* R. K. Willardson, A. C. Beer, Eds., Academic Press, New York, Vol. 18, pp. 201–311 (1981); D. Long, "Photovoltaic and photoconductive infrared detectors," in *Topics in Applied Physics, Optical and Infrared Detectors,* R. J. Keyes, Ed., Springer-Verlag, Berlin, Vol. 19, pp. 101–147 (1970); C. T. Elliot, "Infrared detectors," Chap. 6B in *Handbook on Semiconductors,* C. Hilsum, Ed., North Holland, New York, Vol. 4, pp. 727–798 (1981).

44. R. A. Soref, "Extrinsic IR photoconductivity of Si doped with B, Al, Ga, P, As or Sb," *Journal of Applied Physics.* **38,** 5201–5209 (1967); R. A. Soref, "Monolithic silicon mosaics for far infrared imaging," *IEEE Transactions on Electron Devices* **ED-15,** 209–214 (1968).

45. P. Bratt, "Impurity germanium and silicon infrared detectors," in *Semiconductors and Semimetals,* R. Willardson, A. Beer, Eds., Academic Press, New York, Vol. 12, pp. 39–142 (1977).

46. J. Leotin, C. Lavernyh, M. Goiran, S. Askenazy, and J. Birch, "Stress tunable gallium doped germanium infrared detector system," *International Journal of Infrared and Millimeter Waves* **6,** 323–337 (May 1985).

47. D. Lutz, D. Lemke, and J. Wolf, "Stressed Ge:Ga infrared detectors: performance and operational parameters," *Applied Optics* **25,** 1698–1700 (1986).

48. M. D. Petroff and M. G. Stapelbroek, "Blocked impurity band detectors," U.S. Patent 4,568,960, filed Oct. 23, 1980.

49. I. C. Wu, J. W. Beeman, P. N. Luke, W. L. Hansen, and E. E. Haller, "Ion-implanted extrinsic Ge photodetectors with extended cutoff wavelength," *Applied Physics Letters* **58,** 1431–1433 (1991).

50. F. Shepherd and A. Yang, "Silicon Schottky retinas for infrared imaging," *IEDM Technical Digest,* pp. 310–313 (1973).

51. I. Melngalilis and T. C. Harman, "Single-crystal lead-tin chalcogenides," in *Semiconductors and Semimetals,* R. K. Willardson, A. C. Beer, Eds., Academic Press, New York, pp. 111–174 (1970).

52. A recent publication on this effort is H. Zogg, C. Maissen, J. Masek, S. Blunier, and T. Hosino, "Monolithic IR sensor arrays in heteroepitaxial narrow gap lead chalcogenides on Si for the SWIR, MWIR and LWIR range," *Proceedings of the SPIE* **1308,** 169–177 (1990). This reports on a 66-element linear array configuration.

53. C. T. Elliot, D. Day, and B. J. Wilson, "An integrating detector for serial scan thermal imaging," *Infrared Physics* **22,** 31–42 (1982); A. Blackburn, M. V. Blackman, D. E. Charlton, W. A. E. Dunn, M. D. Jennr, K. J. Oliver, and J. T. M. Wotherspoon, "The practical realization and performance of SPRITE detectors," *Infrared Physics* **22,** 57–64 (1982).

54. W. S. Boyle and G. E. Smith, "Change coupled semiconductor devices," *Bell System Technical*

Journal, pp. 587–593 (1970).

55. J. C. Fraser, D. H. Alexander, R. M. Finnila, and S. C. Su, "An extrinsic SiCCD for detecting infrared radiation," *Digest of Technical Papers, International Electron Device Meeting,* Washington, D.C., December 1973, Institute of Electrical and Electronic Engineers, New York, pp. 442–445 (1973).
56. K. Nummedal, J. C. Fraser, S. C. Su, R. Baron, and R. M. Finnila, "Extrinsic silicon monolithic focal plane array technology and applications," *Proceedings 1975 CCD Applications International Conf.,* San Diego, California, October 1975, Naval Ocean Systems Center, San Diego, pp. 19–30 (1975).
57. N. Sclar, R. L. Maddox, and R. A. Florence," Silicon monolithic infrared detector array," *Applied Optics* **16,** 1525–1532 (1977).
58. W. Parrish, F. Renda, D. Maeding, J. Toman, C. Burgett, R. E. Eck, and N. L. Ray, "Characterization of a 32 × 32 InSb hybrid focal plane," *IEDM Technical Digest,* pp. 513–516 (Dec. 1978).
59. I. S. McLean, "Results with the UKIRT infrared camera," *Proceedings of the SPIE* **782,** 138–141 (1987).
60. I. S. McLean, "Infrared astronomy's new image," *Sky and Telescope,* pp. 254–258 (Mar. 1988).
61. I. Gatley, D. L. DePoy, and A. M. Fowler, "Astronomical imaging with infrared detector arrays," *Science* **242,** 1264–1270 (1988).
62. B. Ewing, "Optical solutions for the army's new light helicopter," *Photonics Spectra,* pp. 85–92 (July 1990).
63. "Detector growth will be in IR arrays," *Laser Focus World* **26,** 49–59 (Nov. 1990).
64. H. W. Messenger, "Commercial detectors see the light," *Laser Focus World* **26,** 133–141 (Nov. 1990).
65. D. A. Scribner, M. R. Kruer, and J. M. Killiany, "Infrared focal plane array technology," *Proceedings of the IEEE* **79,** 66–82 (1991).
66. Some papers discussing the trade-off between temperature, background, and extrinsic versus intrinsic as well as some of the recent detector innovations such as quantum well and superlattice detectors are as follows: M. M. Blouke, C. B. Burgett, and R. L. Williams, "Sensitivity limits for extrinsic and intrinsic infrared detectors," *Infrared Physics* **13,** 61–77 (1973); N. Sclar, "Temperature limitations for IR extrinsic and intrinsic photodetectors," *IEEE Transactions on Electron Devices* **ED-27,** 109–118 (1980); M. A. Kinch and A. Yariv, "Performance limitations of GaAs/AlGaAs infrared superlattices," *Applied Physics Letters* **55,** 2093–2095 (1989); B. F. Levine, "Comment on 'Performance limitations of GaAs/AlGaAs infrared superlattices,' " *Applied Physics Letters* **56,** 2354–2355 (1990); M. A. Kinch and A. Yariv, "Response to 'Comment on Performance limitations of GaAs/AlGaAs infrared superlattices,' " *Applied Physics Letters* **56,** 2355–2356 (1990).
67. After R. Clark Jones who defined this figure of merit: R. C. Jones, "A method of describing the detectivity of photoconductive cells," *Review of Scientific Instruments* **24,** 1035–1040 (1953); R. C. Jones, "Performance of detectors for visible and infrared radiation," in *Advances in Electronics,* L. Marton, Ed., Academic Press, New York, Vol. 5, pp. 2–6 (1953); R. C. Jones, "Phenomenological description of the response and detecting ability of radiation detectors," *Proceedings of the Institute of Radio Engineers* **47,** 1495–1502 (1959).
68. References discussing the details of detector performance measurements include R. D. Hudson, Jr., "Infrared systems engineering," John Wiley & Sons, New York, Chap. 7, pp. 264–303 (1969); W. L. Wolfe and G. J. Zissis, Eds., *The Infrared Handbook,* Environmental Research Institute of Michigan, Ann Arbor, MI, Chap. 11, pp. 11-1–11-104 (Revised 1985); R. W. Boyd, *Radiometry and the Detection of Optical Radiation,* John Wiley & Sons, New York (1983); E. L. Derniak and D. G. Crowe, *Optical Radiation Detectors,* John Wiley & Sons, New York (1984); and J. D. Vincent, *Fundamentals of Infrared Detector Operation and Testing,* John Wiley & Sons, New York (1989).
69. J. P. Rosbeck, R. E. Starr, S. L. Price, and K. J. Riley, "Background and temperature dependent current-voltage characteristics of HgCdTe photodiodes," *Journal of Applied Physics* **53,** 6430–6440 (1982). The paper summarizes experimental data on this effect and discusses a model relating the reduction in R_0A product to a bias dependence of the quantum efficiency due to changes in the depletion layer width. The background dependence of R_0A product occurs as a result of several second-order effects. One of these is a reduction in the depletion layer width as more background current flows through the diode. This can affect quantum efficiency and effective diode area. In the case of heterojunction diodes, junction barriers may also be modulated by the photocurrent, which affects the quantum efficiency. A detailed

understanding of this effect is still under investigation. At the present time, Richard Schoolar of the Aerospace Corporation has been analyzing these effects. K. Kosai, private communication.

70. These uniformity data are unpublished Santa Barbara Research Center (HgCdTe, InSb, PbS, and PbSe) and Hughes Technology Center (PtSi) results with two exceptions. Edward Nelson of Kodak supplied the data for the visible silicon sensor and two PtSi data points.
71. As appears to be the case in Fig. 4.35 for LWIR PV HgCdTe. Paul Norton and William Radford, unpublished results.
72. D. Long, "Uniformity of infrared detector parameters in alloy semiconductors," *Infrared Physics* **12,** 115–124 (1972).
73. For a discussion of this topic in greater detail, see J. M. Mooney and E. L. Dereniak, "Comparison of the performance limit of Schottky-barrier and standard infrared focal plane arrays," *Optical Engineering* **26,** 223 (1987); J. M. Mooney, F. Shepherd, W. S. Ewing, J. E. Murguia, and J. Silverman, "Responsivity nonuniformity limited performance of infrared staring cameras," *Optical Engineering* **28,** 1151–1161 (1989).
74. D. A. Scribner, K. A. Sarkady, J. T. Caulfield, M. R. Kruer, G. Katz, and C. J. Gridley, "Nonuniformity correction for staring IR focal plane arrays using scene-based techniques," *Proceedings of the SPIE* **1308,** 224–233 (1990).
75. S. M. Johnson, M. H. Kalisher, W. L. Ahlgren, J. B. James, and C. A. Cockrum, "HgCdTe 128 × 128 infrared focal plane arrays on alternative substrates of CdZnTe/GaAs/Si," *Applied Physics Letters* **56,** 946 (1990); K. Zanio, R. Bean, R. Mattson, P. Vu, and S. Taylor, "HgCdTe on GaAs/Si for mid-wavelength infrared focal plane arrays," *Applied Physics Letters* **56,** 1207–1209 (1990).
76. R. A. Chapman, S. R. Borrello, A. Simmons, J. D. Beck, A. J. Lewis, M. A. Kinch, J. Hynecek, and C. G. Roberts, "Monolithic HgCdTe charge transfer device infrared imaging arrays," *IEEE Transactions on Electron Devices* **ED-27,** 134–145 (1980); R. D. Thom, T. L. Koch, J. D. Langan, and W. J. Parrish, "A fully monolithic InSb infrared CCD array," *IEEE Transactions on Electron Devices* **ED-27,** 160–170 (1980); C-Y. Wei, K. L. Wang, E. A. Taft, J. M. Swab, M. D. Gibbons, W. E. Davern, and D. M. Brown, "Technology development for InSb infrared imagers," *IEEE Transactions on Electron Devices* **ED-27,** 170–175 (1980).
77. T. T. Schaefer, "The RM 20B mosaic measurement experiment," *Proceedings of the SPIE* **62,** 36–45 (1975); J. Carson, "Infrared mosaic technology," *Proceedings of the SPIE* **62,** 3–6 (1975); J. Carson, "Infrared mosaic technology," *Proceedings of the SPIE* **510,** 79–81 (1984).
78. A recent publication containing a dozen or so articles on Z technology may be found in *Proceedings of the SPIE* **1097**; additional articles are in *Proceedings of the SPIE* **930** and **1339.**
79. I. M. Baker and R. A. Ballingall, "Photovoltaic CdHgTe-silicon hybrid focal planes," *Proceedings of the SPIE* **510,** 121–129 (1985).
80. I. M. Baker, J. E. Parsons, J. H. W. Lewis, R. A. Lockett, J. T. M. Wotherspoon, R. A. Ballingall, and I. Blenkinsop, "Recent developments in CdHgTe-silicon hybrid focal planes," *Proceedings of the SPIE* **588,** 16–23 (1986).
81. M. A. Keenan, I. M. Baker, J. E. Parsons, R. A. Ballingall, P. N. J. Dennis, and T. W. Ridler, "Advances in linear and two dimensional CdHgTe-Si hybrid focal plane arrays," *Third International Conference on Advanced Infrared Detectors and Systems,* IEE, London, Vol. 263, pp. 54–59 (1986).
82. G. Finger, M. Meyer, and A. F. M. Moorwood, "Test results with Mullard CMT-CCD hybrid focal plane arrays," *Proceedings of the SPIE* **865,** 94–101 (1988).
83. C. T. Elliot, N. T. Gordon, R. S. Hall, and G. Crimes, "Reverse breakdown in long wavelength lateral collection CdHgTe diodes," *Journal of Vacuum Science and Technology* **A8,** 1251–1253 (1990).
84. R. L. Smythe, "Monolithic HgCdTe focal plane arrays," *GOMAC Conference Proceedings,* pp. 289–292, U.S. Army ERADCOM, Ft. Monmouth, NJ (Nov. 1982).
85. B. F. Levine, C. G. Bethea, G. Hasnain, V. O. Shen, E. Pelve, R. R. Abbott, and S. J. Hsieh, "High sensitivity low dark current 10 μm GaAs quantum well infrared photodetectors," *Applied Physics Letters* **56,** 851–853 (1990).
86. Examples of papers on this approach are D. L. Smith and C. Mailhiot, "Proposal for strained type II superlattice infrared detectors," *Journal of Applied Physics* **62,** 2545–2548 (1987); S. R. Kurtz, L. R. Dawson, T. E. Zipperian, and S. R. Lee, "Demonstration of an InAsSb strained-layer superlattice photodiode," *Applied Physics Letters* **52,** 1581–1583 (1988); R. H. Miles, D. H. Chow, J. N. Schulman, and T. C. McGill, "Infrared optical characterization of InAs/GaInSb superlattices," *Applied Physics Letters* **57,** 801–803 (1990).

87. O. Byungsung, J.-W. Choe, M. H. Francombe, K. M. S. V. Bandara, D. D. Coon, Y. F. Lin, and W. J. Takei, "Long-wavelength infrared detection in a Kastalsky-type superlattice structure," *Applied Physics Letters* **57,** 503–505 (1990).

88. T. G. Stratton, B. E. Cole, P. W. Kruse, R. A. Wood, K. Beauchamp, T. F. Wang, B. Johnson, and A. M. Goldman, "High-temperature superconducting microbolometer," *Applied Physics Letters* **57,** 99–100 (1990).

89. A number of relevant papers on this subject are in *Proceedings of the SPIE* **1292** and **1447**.

90. A number of interesting papers on the superlattice and quantum well detector concepts, as well as some other novel detector concepts, are presented in *Proceedings of the Innovative Long Wavelength Infrared Detector Workshop,* Pasadena, California, April 24–26, 1990, Jet Propulsion Laboratory.

91. M. Lanir, C. C. Wang, and A. H. B. Vanderwyck, "Backside illuminated HgCdTe/CdTe photodiodes," *Applied Physics Letters* **34,** 50–52 (1979).

92. C. C. Wang, M. Chu, S. H. Shin, W. E.Tennant, J. T. Cheung, M. Lanir, A. H. B. Vanderwyck, G. M. Williams, L. O. Bubulack, and R. J. Eisel, "HgCdTe/CdTe heterostructure diodes and mosaics," *IEEE Transactions on Electron Devices* **ED-27,** 154–160 (1980).

93. M. Chu, A. H. B. Vanderwyck, and D. T. Cheung, "High performance backside-illuminated HgCdTe/CdTe (10 μm) planar diodes," *Applied Physics Letters* **37,** 486–488 (1980).

94. P. Becla and E. Placzek-Popko, "Electrical properties of infrared photovoltaic CdHgTe detectors," *Infrared Physics* **21,** 323–332 (1981).

95. P. Migliorato, R. F. C. Farrow, A. B. Dean, and G. M. Williams, "CdTe/HgCdTe indium-diffused photodiodes," *Infrared Physics* **22,** 331–336 (1982).

96. M. Lanir and K. J. Riley, "Performance of PV HgCdTe arrays for 1-14 μm applications," *IEEE Transactions on Electron Devices* **ED-29,** 274–279 (1982).

97. J. M. Arias, S. H. Shin, J. G. Pasko, R. E. DeWames, and E. R. Gertner, "Long and middle wavelength infrared photodiodes fabricated with HgCdTe grown by molecular beam epitaxy," *Journal of Applied Physics* **65,** 1747–1753 (1989).

98. M. B. Reine, "Status of LWIR HgCdTe infrared detector technology," in *Proceedings of the Innovative Long Wavelength Infrared Detector Workshop,* Pasadena, California, April 24–26, 1990, Jet Propulsion Laboratory, pp. 61–77.

99. L. J. Kozlowski, K. Vural, V. H. Johnson, J. K. Chen, R. B. Bailey, D. Bui, M. J. Gubala, and J. R. Teague, "256 × 256 PACE-1 PV HgCdTe focal plane arrays for medium and short wavelength IR applications," *Proceedings of the SPIE* **1308,** 202–208 (1990).

100. W. E. Tennant, "LWIR HgCdTe—innovative detectors in an incumbent technology," *Proceedings of the Innovative Long Wavelength Infrared Detector Workshop,* Pasadena, California, April 24–26, 1990, Jet Propulsion Laboratory, pp. 79–91.

101. Santa Barbara Research Center, unpublished data.

102. M. C. Teich, "Infrared heterodyne detection," *Proceedings of the IEEE* **56,** 37–46 (1968).

103. Early HgCdTe CO_2 laser detectors were displayed at the French exhibit of the Montreal Expo in 1967. Some publications on CO_2 laser detectors include C. Vérié and M. Sirieix, "Gigahertz cutoff frequency capabilities of CdHgTe photovoltaic detectors at 10.6 μ," *IEEE Journal of Quantum Electronics* **QE-8,** 180–184 (1972); D. L. Spears, "Planar HgCdTe quadrantal heterodyne arrays with GHz response at 10.6 μm," *Infrared Physics* **17,** 5–8 (1977); E. Igras, J. Piotrowski, and T. Piotrowski, "Ultimate detectivity of (CdHg)Te infrared photoconductors," *Infrared Physics* **19,** 143–149 (1979); D. L. Spears, "Wide bandwidth CO_2 laser photomixers," *Proceedings of the SPIE* **227,** 108–116 (1980); D. L. Spears, "IR detectors: heterodyne and direct," in *Optical and Laser Remote Sensing,* D. K. Killinger, A. Mooradian, Eds., Springer-Verlag, Berlin (1980); M. C. Wilson and D. J. Dinsdale, "CO_2 laser detection with cadmium mercury telluride," in *Third International Conference on Advanced Infrared Detectors and Systems,* IEE, London, Vol. 263, pp. 139–145 (1986); M. C. Wilson, W. A. E. Dunn, and D. J. Wilson, "Infrared heterodyne detectors using cadmium mercury telluride at intermediate temperatures," *Proceedings of the SPIE* **588,** 29–31 (1986); P. R. Bratt, "Development of a P-I-N HgCdTe photomixer for laser heterodyne spectroscopy," NASA Contract Report 4096 (Sep. 1987).

104. B. Orsal, R. Alabedra, M. Valenza, G. Lecoy, J. Meslage, and C. Y. Boisrobert, "HgCdTe 1.55 μm avalanche photodiode noise analysis in the vicinity of resonant impact ionization connected with the spin-orbit split-off band," *IEEE Transactions on Electron Devices* **ED-35,** 101–107 (1988).

105. S. H. Shin, J. G. Pasko, H. D. Law, and D. T. Cheung, "1.22 μm HgCdTe/CdTe avalanche photodiodes," *Applied Physics Letters* **40,** 965–967 (1982).

106. G. R. Chapman and M. D. Jack, Santa Barbara Research Center, unpublished data.

107. A. P. Davis, "UK thermal imaging common modules class II—an update on detector and related component enhancements," *Proceedings of the SPIE* **1157,** 176–184 (1989).

108. D. L. Smith, "Effects of blocking contacts on generation-recombination noise and responsivity in intrinsic photoconductors," *Journal of Applied Physics* **56,** 1663–1669 (1984); D. K. Arch, R. A. Wood, and D. L. Smith, "High responsivity HgCdTe heterojunction photoconductor," *Journal of Applied Physics* **58,** 2360–2370 (1985).

109. Data are summarized from unpublished Santa Barbara Research Center results and from M. A. Kinch, S. R. Borrello, B. H. Breazeale, and A. Simmons, "Geometrical enhancement of HgCdTe photoconductive detectors," *Infrared Physics* **17,** 137–145 (1977).

110. P. Norton, "HgCdTe for NASA EOS missions and detector uniformity benchmarks," *Proceedings of the Innovative Long Wavelength Infrared Detector Workshop,* Pasadena, California, April 24–26, 1990, Jet Propulsion Laboratory, pp. 93–94.

111. For a recent discussion of white noise in conventional photoconductors see D. L. Smith, "Theory of generation-recombination noise in intrinsic photoconductors," *Journal of Applied Physics* **53,** 7051–7060 (1982). The $1/f$ noise is discussed in M. A. Kinch, S. R. Borrello, B. H. Breazeale, and A. Simmons, "Geometrical enhancement of HgCdTe photoconductive detectors," *Infrared Physics* **17,** 137–145 (1977) and in R. M. Broudy and V. J. Mazurczyk, "(HgCd)Te photoconductive detectors," Chap. 5 in *Semiconductors and Semimetals,* R. K. Willardson, A. C. Beer, Eds., Academic Press, New York, Vol. 18, pp. 157–199 (1981).

112. Readers interested in the performance limits of conventional photoconductors may wish to consult D. Long, "On generation-recombination noise in infrared detector materials," *Infrared Physics* **7,** 169–170 (1967); R. L. Williams, "Sensitivity limits of 0.1 eV intrinsic photoconductors," *Infrared Physics* **8,** 337–343 (1968).

113. W. F. Kosonocky, "Review of Schottky-barrier imager technology," *Proceedings of the SPIE* **1308,** 2–26 (1990).

114. E. Kohn, S. Roosild, F. Shepherd, and A. Yang, "Infrared imaging with monolithic CCD-addressed Schottky-barrier detector arrays, theoretical and experimental results," *International Conf. on Application of CCDs,* October 29–31, 1975, pp. 59–69.

115. J. Edwards, J. Gates, H. Altin-Mees, W. Connelly, and A. Thompson, "244 $\times$400 element hybrid platinum silicide Schottky focal plane array," *Proceedings of the SPIE* **1308,** 99–110 (1990).

116. B. Capone, L. Skolnik, R. Taylor, F. Shepherd, S. Roosild, W. Ewing, W. Kosonocky, and E. Kohn, "Evaluation of a Schottky IRCCD staring mosaic focal plane," *Proceedings of the SPIE* **156,** 120–131 (1978); see also *Optical Engineering* **18**(5), 535–541 (1979).

117. J. Silverman, J. M. Mooney, and V. E. Vickers, "Display of wide dynamic range infrared images from PtSi Schottky barrier cameras," *Optical Engineering* **29,** 97–104 (1990).

118. J. E. Murguia, J. M. Mooney, and W. E. Ewing, "Evaluation of a PtSi infrared camera," *Optical Engineering* **29,** 786–794 (1990).

119. Some interesting details of this process have been discussed recently in J. M. Mooney, "The dependence of the Schottky emission coefficient on reverse bias," *Journal of Applied Physics* **65,** 2869–2871 (1989).

120. Jon Mooney, unpublished data.

121. P. Pellegrini, A. Golubovic, and C. Ludington, "IrSi Schottky barrier diodes for infrared systems," *IEDM Technical Digest,* pp. 157–160 (1982); B. Y. Tsaur, M. M. Weeks, R. Trubiano, P. W. Pellegrini, and T. R. Yew, "IrSi Schottky-barrier infrared detectors with 10 μm cutoff wavelength," *IEEE Electron Device Letters* **9**(12), 650–653 (1988).

122. J. Rosbeck, I. Kosai, R. Hoendervoogt, and M. Lanir, "High performance Be implanted photodiodes," *IEDM Technical Digest,* pp. 161–164 (Dec. 1981).

123. G. Orias, A. Hoffman, and M. Casselman, "58 $\times$ 62 indium antimonide focal plane array for infrared astronomy," *Proceedings of the SPIE* **627,** 408–417 (1986).

124. J. Blackwell, S. Botts, A. Laband, and H. Arnold, "An affordable 128 $\times$ 128 indium antimonide hybrid focal plane array," *Proceedings of the SPIE* **1157,** 243–249 (1989).

125. Litton Electron Devices, Tempe, AZ.

126. Santa Barbara Research Center, unpublished data.

127. J. Wimmers, R. Davis, C. Niblack, and D. Smith, "Indium antimonide detector technology at Cincinnati Electronics," *Proceedings of the SPIE* **930,** 125–138 (1988).

128. M. Gibbons, S. Wang, S. Jost, V. Meikleham, T. Myers, and A. Milton, "Developments in

indium antimonide material and charge injection devices," *Proceedings of the SPIE* **865,** 52–58 (1988).

129. N. Sclar, "Development status of silicon IR detectors," *Proceedings of the SPIE* **409,** 53–61 (1983).
130. A. Zachor, E. Huppi, and E. Ray, "Nonlinear response of low-background extrinsic silicon detectors—a phenomenological model," *Applied Optics* **20,** 1000–1004 (1981).
131. A. Zachor, E. Huppi, I. Coleman, and D. Frodsham, "Nonlinear response of low-background extrinsic silicon detectors—a revised model," *Applied Optics* **21,** 2027–2035 (1982).
132. A. Rose, "Space-charge limited currents in solids," *Physics Review* **97,** 1538–1544 (1955).
133. W. van Roosbroeck and H. C. Casey, Jr., "A new regime of semiconductor behavior: carrier transport when dielectric relaxation time exceeds lifetime," *Proceedings of the Tenth International Conference on the Physics of Semiconductors,* Cambridge, Massachusetts, August 17–21, 1970, pp. 832–838.
134. H. J. Queisser, H. C. Casey, Jr., and W. van Roosbroeck, "Carrier transport and potential distributions for a semiconductor *p-n* junction in the relaxation regime," *Physics Review Letters* **26,** 551–554 (1971).
135. A. F. Milton and M. M. Blouke, "Sweepout and dielectric relaxation in compensated extrinsic photoconductors," *Physics Review* **B3,** 4312–4330 (1971).
136. M. M. Blouke, E. E. Harp, C. R. Jeffus, and R. L. Williams, "Gain saturation in extrinsic germanium photoconductors operating at low temperatures," *Journal of Applied Physics* **43,** 188–194 (1972).
137. M. M. Blouke and R. L. Williams, "Gain saturation in high-resistivity Si:B photoconductors," *Applied Physics Letters* **20,** 25–27 (1972).
138. W. van Roosbroeck and H. C. Casey, Jr., "Transport in relaxation semiconductors," *Physics Review* **B5,** 2154–2175 (1972).
139. S. B. Stetson, D. B. Reynolds, M. G. Stapelbroek, and R. L. Sermer, "Design and performance of blocked-impurity-band detector focal plane arrays," *Proceedings of the SPIE* **686,** 48–65 (1986).
140. D. M. Watson and J. E. Huffman, "Germanium blocked-impurity-band far-infrared detectors," *Applied Physics Letters* **52,** 1602–1604 (1988).
141. J. Monroy, R. Baron, G. Albright, J. Boisvert, and L. Flesner, "Energetic electron-induced impurity ionization in Si:As IBC detectors," *IEEE Transactions on Nuclear Science* **35,** 1307–1312 (1988).
142. A few key papers that describe the experimental data and the physics of hopping-conduction are as follows: H. Fritzsche, "Electrical properties of germanium semiconductors at low temperatures," *Physics Review* **99,** 406–419 (1955); E. M. Conwell, "Impurity band conduction in germanium and silicon," *Physics Review* **103,** 51–61 (1956); S. H. Koenig and G. R. Gunther-Mohr, "The low temperature electrical conductivity of *n*-type germanium," *J. Phys. Chem. Solids* **2,** 268–283 (1957); A. Miller and Elihu Abrahams, "Impurity conduction at low concentrations," *Physics Review* **120,** 745–755 (1960).
143. J. Wolf and D. Lemke, "Performance of Si:Ga infrared detectors under reduced background fluxes," *Astronomy and Astrophysics* **119,** 294–296 (1983).
144. J. Goebel, J. McKelvey, C. McCreight, and G. Anderson, "Low background direct readout array performance," *Proceedings of the SPIE* **627,** 418–429 (1986).
145. M. McKelvey, C. McCreight, J. Goebel, N. Moss, and M. Savage, "Characterization of direct readout Si:Sb and Si:Ga infrared detector arrays for space-based astronomy," *Proceedings of the SPIE* **868,** 73–80 (1988).
146. D. Y. Gezari, W. C. Folz, L. A. Woods, and J. B. Wooldridge, "A 58 × 62 pixel Si:Ga array camera for 5-14 μm astronomical imaging," *Proceedings of the SPIE* **973,** 287–298 (1988).
147. D. Gezari, W. Folz, and L. Woods, "Initial astronomical results with a new 5-14 μm Si:Ga 58 × 62 DRO array camera," *Proceedings of the 3rd Ames Detector Workshop,* pp. 267–282 (Feb. 1989).
148. M. D. Petroff, M. G. Stapelbroek, and W. A. Kleinhans, "Solid state photomultiplier," U.S. Patent 4,586,068, filed Oct. 7, 1983.
149. M. D. Petroff, M. B. Stapelbroek, and W. A. Kleinhans, "Detection of individual 0.4-28 μm wavelength photons via impurity-impact ionization in a solid-state photomultiplier," *Applied Physics Letters* **51,** 406–408 (1987).
150. R. Bharat, M. D. Petroff, and M. G. Stapelbroek, "Solid-state photomultiplier—a high per-

formance detector for astronomy," *Proceedings of the Workshop on Ground-based Astronomical Observations with Infrared Array Detectors,* C. G. Wynn-Williams, E. E. Becklin, Eds., Hilo, Hawaii, March 24–26, 1987.

151. Reviews of these devices include J. N. Humphrey, "Optimization of lead sulfide infrared detectors under diverse operating conditions," *Applied Optics* **4,** 665–675 (1965); D. Bode, *Lead Salt Detectors,* Academic Press, New York (1966); T. H. Johnson, "Lead salt detectors and arrays; PbS and PbSe," *Proceedings of the SPIE* **443,** 60–94 (1984).
152. "Second-generation PbSe arrays demonstrate scanning imagery," *Laser Focus World* **26,** 11–13 (Nov. 1990).
153. Litton Electron Devices, Tempe, AZ.
154. Santa Barbara Research Center, unpublished data.

Bibliography

Andrews, D. H., R. M. Milton, and W. DeSorbo, "A fast superconducting bolometer," *Journal of the Optical Society of America* **36**(9), 518–524 (1946).

Bell, R. L., and W. E. Spicer, "3-5 compound photocathodes: a new family of photoemitters with greatly improved performance," *Proceedings of the IEEE* **58,** 1788 (1970).

Biard, J. R., and W. E. Spicer, "A model of the avalanche photodiode," *IEEE Transactions on Electron Devices* **14,** 233–238 (1967).

Blouke, M. M., C. B. Burgett, and R. L. Williams, "Sensitivity limits for extrinsic and intrinsic infrared detectors," *Infrared Physics* **13**(1), 61–72 (1973).

Borrello, S. R., "Detection uncertainty," *Infrared Physics* **12,** 267–270 (1972).

Boyle, W. S., and K. F. Rodgers, Jr., "Performance characteristics of a new low-temperature bolometer," *Journal of the Optical Society of America* **49**(1), 66–69 (1959).

Burstein, E., G. Pines, and N. Sclar, "Optical and photoconductive properties of silicon and germanium," *Photoconductive Conference,* R. G. Breckenridge et al., Eds., John Wiley & Sons, New York, pp. 353–413 (1956).

Coringella, P. C., and W. L. Eisenman, "System for low-frequencies noise measurements," *Review of Scientific Instruments* **33,** 654 (1962).

Dereniak, E. L., and D. G. Crowe, *Optical Radiation Detectors,* John Wiley & Sons, New York (1984).

Eisenman, W. L., and R. L. Bates, "Improved black radiation detector," *Journal of the Optical Society of America* **54,** 1280 (1964).

Eisenman, W. L., R. L. Bates, and J. D. Merriam, "Black radiation detector," *Journal of the Optical Society of America* **53,** 729 (1963).

Emmons, R. B., and G. Lucovsky, "The frequency response of avalanching photodiodes," *IEEE Transactions on Electron Devices* **13,** 297 (1966).

Golay, M. J. E., "Theoretical considerations in heat and infra-red detection, with particular reference to the pneumatic detector," p. 347, and "A pneumatic infra-red detector," p. 357; *Review of Scientific Instruments* **18**(5) (May 1947); M. J. E. Golay, "The theoretical and practical sensitivity of the pneumatic infra-red detector," *Review of Scientific Instruments* **20,** 816 (1949).

Holter, M., S. Nudelman, G. Suits, W. Wolfe, and G. Zissis, *Fundamentals of Infrared Technology,* Macmillan, New York (1962).

Infrared Industries, Santa Barbara, CA, Data Sheet No. 698-A, Cell D52-9P, PbSe (Feb. 1961).

Jamieson, J. A., "Preamplifiers for nonimage-forming infrared systems," *Proceedings of the IRE* **47,** 1522 (1959).

Johnson, J. B., "Thermal agitation of electricity in conductors," *Physical Review* **32,** 97 (1928).

Johnson, K. M., "High-speed photodiodes signal enhancement at avalanche breakdown voltage," *IEEE Transactions on Electron Devices* **12,** 55 (1965).

Jones, R. Clark, D. Goodwin, and G. Pullan, "Standard procedures for testing infrared detectors and for describing their performance," AD No. 257597, Office of Defense and Development Research and Engineering, Washington, DC, pp. 1–45 (Sep. 1960).

Kinch, M. A., and S. R. Borrello, "0.1 eV HgCdTe photodetectors," *Infrared Physics* **15**(2), 111–124 (May 1975).

Kruse, P. W., L. D. McGlauchlin, and R. B. McQuistan, *Elements of Infrared Technology,* John Wiley & Sons, New York (1962).

Long, D., "Generation-recombination noise limited detectivities of impurity and intrinsic photoconductive 8-14 μ infrared detectors, *Infrared Physics* **7,** 121–128 (1967).

Long, D., and J. L. Schmit, "Mercury-cadmium telluride and closely related alloys," in *Semiconductors and Semimetals,* R. K. Willardson, A. C. Beer, Eds., Academic Press, New York, Vol. 5, pp. 175–255 (1970).

Low, F. J.,"Low-temperature germanium bolometer," *Journal of the Optical Society of America* **51**(11), 1300–1304 (1961).

Low, F. J., and A. R. Hoffman, "The detectivity of cryogenic bolometers," *Applied Optics* **2**(6), 649–650 (1963).

Martin, A. E., *Infrared Instrumentation and Techniques,* Elsevier Science Publishing, New York (1966).

McIntyre, R. J., "Multiplication noise in uniform avalanche diodes," *IEEE Transactions on Electron Devices* **13,** 164 (1966).

Melngailis, I., and T. Harman, "Single-crystal lead-tin chalcogenides," in *Semiconductors and Semimetals, Infrared Detectors,* R. K. Willardson, A. C. Beer, Eds., Academic Press, New York, Vol. 5, p. 113 (1970).

Miller, S. L., "Ionization rates for holes and electrons in silicon," *Physical Review* **105,** 1246–1249 (1957).

Morton, G. A., "Infrared photoemission," *Proceedings of the IRE* **47,** 1467 (1959).

Moss, T. S., G. J. Burrell, and B. Ellis, *Semiconductor Opto-Electronics,* John Wiley & Sons, New York (1973).

Naval Ocean Systems Center, San Diego, CA, *Properties of Photoconductive Detectors,* NOLC Report 564 (a continuing series begun June 30, 1952).

North American Rockwell, Anaheim, CA, Detector 632-1, Data Sheet 976 (Mar. 9, 1972).

Optical Industry and Systems Directory, 23rd ed., Optical Publishing Co., The Pittsfield, MA (1977).

Petritz, R. L., "Fundamentals of infrared detection, *Proceedings of the IRE* **47,** 1459–1467 (Sep. 1959); G. R. Pruett and R. L. Petritz, "Detectivity and preamplifier considerations for indium antimonide photovoltaic detectors," *Proceedings of the IRE* **47,** 1524–1529 (Sep. 1959).

Pines, M., and R. Baron, "Characteristics of indium doped silicon infrared detectors," *Proceedings of the 1974 National Electron Devices Meeting,* Washington, DC, December 1974, Institute of Electrical and Electronics Engineers, New York, p. 446.

Pines, M., D. Murphy, D. Alexander, R. Baron, and M. Young, "Characteristics of gallium doped silicon infrared detectors," *Technical Digest of the IEEE,* Institute of Electrical and Electronics Engineers, New York, p. 502.

Potter, R. F., J. M. Pernett, and A. B. Naugle, "The measurement and interpretation of photodetector parameters," *Proceedings of the IRE* **47,** 1503 (1959).

Putley, E. H., "Far infrared photoconductivity," *Physics Status Solidi* (B, Basic Research) Vol. 6, 1964, p. 571.

Putley, E. H., "The pyroelectric detector," *Semiconductors and Semimetals,* R. K. Willardson, A. C. Beer, Eds., Academic Press, New York, Vol. 5, pp. 259–285 (1970).

Scheer, J. J., and J. Van Loar, "GaAs-Cs: a new type of photoemitter," *Solid State Communications* **3,** 189 (1965).

Smith, R. A., F. E. James, and R. P. Chasmar, *The Detection and Measurement of Infrared Radiation,* Clarendon Press, Oxford, England (1968).

Sonnenberg, H., "Low-work-function surfaces for negative-electron-affinity photoemitters," *Applied Physics Letters* **14,** 298 (1969).

Soref, R. A., "Extrinsic IR photoconductivity of Si doped with B, Al, Ga, P, As or Sb," *Journal of Applied Physics* **38,** 520 (1967).

Stair, R., W. E. Schneider, W. R. Walters, and J. K. Jackson, "Some factors affecting the sensitivity and spectral response of thermoelectric (radiometric) detectors," *Applied Optics* **4,** 703 (1965).

Tolman, R. C., *The Principles of Statistical Mechanics,* Clarendon Press, Oxford, England, pp. 94, 95, 97, 145, 163 (1938).

Uebbing, J. J., and R. L. Bell, "Improved photoemitters using AsAsInGaAs," *Proceedings of the IEEE* **56,** 1624 (1968).

Van Vliet, K. M., "Noise in semiconductors and photoconductors," *Proceedings of the IRE* **46,** 1004 (1958).

van der Ziel, A., "Noise in junction transistors," *Proceedings of the IRE* **46,** 1019 (1958).

Williams, R. L., "Speed and sensitivity limitations for extrinsic photoconductors," *Infrared Physics* **9,** 37–40 (1969).

Wolfe, W. L., "Photon number D^* figure of merit," *Applied Optics* **12**(3), 619–621 (Mar. 1973).

Zworykin, V. K., and E. G. Ramberg, *Photoelectricity and Its Applications,* John Wiley & Sons, New York (1949).

CHAPTER 5

Readout Electronics for Infrared Sensors

John L. Vampola
Santa Barbara Research Center
Goleta, California

CONTENTS

5.1 INTRODUCTION

The readout integrated circuit (ROIC) is a highly integrated set of focal plane electronic functions combined into a single semiconductor chip. Its primary function is to provide infrared detector signal conversion and amplification, along with time multiplexing of data from many detectors to just a minimum number of outputs. ROICs can contain tens to hundreds of thousands of individual *unit cells,* each with critical detector amplifiers and multiplexer switches, as shown in Fig. 5.1. ROICs are normally processed using conventional silicon integrated circuit technology. They are most often implemented in complementary metal oxide semiconductor (CMOS) technology, allowing for higher resolution and greater sensitivity in today's sensors.

This chapter covers each of the ROIC functions shown in Fig. 5.1, including preamplifier, signal processor, multiplexer, and video amplifier sections. These functions are addressed in terms of major design drivers such as noise, dynamic range, and power. Since the signal-to-noise ratio (SNR) is the major driver in most sensor designs, each of the circuits is detailed in this context. The simplest circuits are introduced first to provide the basis for more advanced circuits. The resistor transimpedance amplifier (RTIA), which is most common in discrete configurations but also utilized in ROIC configurations, will be addressed initially to introduce many of the basic signal and noise concepts required for analysis of other preamplifier circuits in subsequent sections.

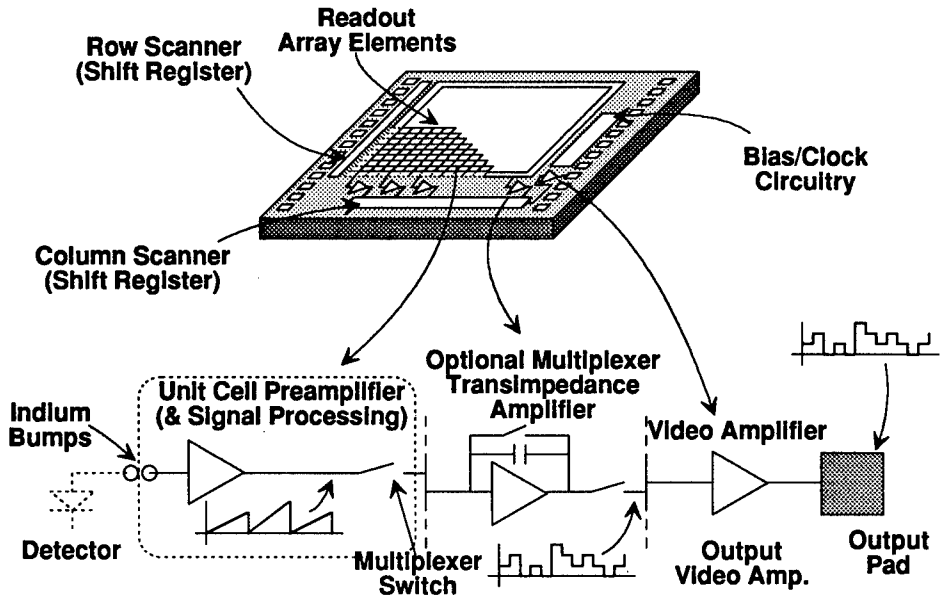


Fig. 5.1 The readout integrated circuit combines detector signal amplification, data multiplexing, and video output buffering of the infrared signal into a single chip. Shown is an example of a staring two-dimensional array.

High-impedance detectors, such as photovoltaics, extrinsic silicon, platinum silicide, and many photoconductors, are extremely sensitive to electromagnetic interference (EMI). A robust sensor design reduces EMI by locating the preamplifier as close as possible to the detector. Through the 1970s, preamplifiers were built up from discrete resistors, capacitors, and transistors into hybrids, which could be placed within inches of the detector. In the most sensitive applications, the front-end of the preamplifier, usually comprised of field effect transistors (FETs) and feedback resistors, was physically placed next to the detectors on the cold focal plane. Since a great amount of real estate would be required for such circuits, discrete amplifier designs put severe limitations on the number of detector channels that could be implemented in a given optical field of view.

The first integrated approaches to addressing both high-density scanning arrays and two-dimensional focal planes were designed in the 1970s. Visible sensors, the ancestors of today's camcorder focal planes, were demonstrated utilizing a single silicon chip composed of charge-coupled devices (CCDs) that served as both sensor and multiplexer. The advent of indium bump interconnect technology, whereby matching sets of small indium bumps are formed on the detector and the CCD, provided the mechanism for connecting a large array to its readout, forming a sensor chip assembly (SCA) as shown in Fig. 5.2. Because of this history, sensor users often refer to all ROICs as multiplexers or CCDs even though CCD devices may not be employed.

CCDs commonly utilize direct injection (DI) or gate modulation preamplifiers to buffer and accumulate the photon-induced current over a frame of scene data. These simple preamplifier types are covered in detail later in this chapter. Through the 1980s, as detector and integrated circuit densities increased, more elegant preamplifiers, as well as multiplexers and video drivers, were incorporated into the ROIC. These newer circuits expanded the use of SCAs to a broader set of applications by providing higher SNR, greater bandwidth, and better linearity. In addition to optimizing basic SCA functions,

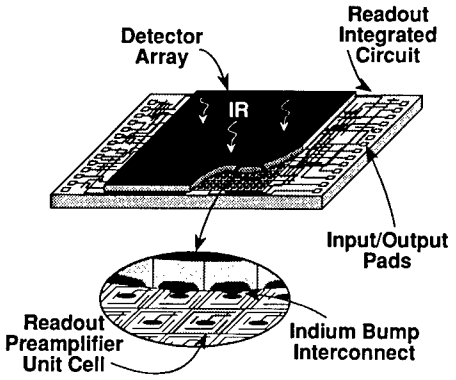


Fig. 5.2 The readout integrated circuit mates with the detector array via indium bumps to form the sensor chip assembly capable of integrating hundreds to millions of elements onto a single assembly.

many other signal processing functions have also been integrated into the SCA, such as detector frame store and circuits that eliminate amplifier drift.

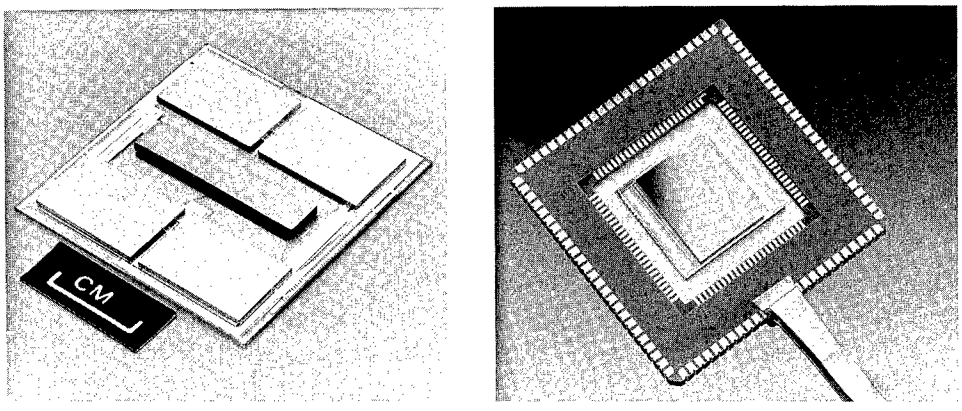
The ROIC, together with the detector, can be assembled to form various SCA configurations. The most common infrared configurations are the direct and the indirect hybrid SCAs of Fig. 5.3. A third type, the monolithic SCA configuration, is common in visible applications.

The direct hybrid SCA approach is common in high-density staring and scanning applications, which provide sufficient unit cell area under the detector element to accommodate a readout preamplifier and its associated circuitry. The direct hybrid SCA has fewer parts than the indirect and is therefore generally more producible.

Indirect hybrid SCAs can include one or more detector arrays fanned out to one or more ROICs via a single fanout board. Larger, more elaborate preamplifiers and signal processing electronics can be fabricated in the larger unit cells of indirect readouts, since the circuit area is no longer constrained to the real estate available under the detector element. The indirect hybrid can also reduce the stress caused by thermal mismatch between the detector and ROIC materials, thereby increasing the thermal cycle life of large SCAs.

A variation of the direct hybrid SCA design is the siderider SCA. This type of SCA includes additional signal processing circuitry, such as time delay integration (TDI), in the area of the ROIC adjacent to the detector array. The preamplifier signal, originating under each of the detector elements, is enabled to this siderider area of the ROIC for additional signal processing prior to being transmitted off the focal plane.

The monolithic device has both the detector and readout circuitry fabricated into a single semiconductor material. An example of a silicon monolithic device is the commercial camcorder SCA, which has the readout fabricated in the silicon adjacent to each of the detector elements. In this case the detector optical area is reduced to accommodate the readout circuitry, resulting in a low de-



(a) (b)

Fig. 5.3 (a) The 960 × 4 element detector with four readouts mounted on a motherboard is an example of an indirect scanning SCA. (Courtesy of DARPA) (b) The 640 × 480 element detector array coupled directly to the readout is an example of a staring direct hybrid SCA. (Courtesy of Santa Barbara Research Center)

tector (or optical) fill factor. Applications of the monolithic approach are also found in shorter wavelength infrared detector material, which utilizes CCD and FET devices.[1]

Common readout integrated circuit symbols, nomenclature, and units are given in Table 5.1.

5.2 MOSFET PRIMER

Although discrete preamplifiers can be designed with bipolar junction transistor (BJT) or junction field effect transistor (JFET) technology, integrated circuit forms of the preamplifier are most commonly fabricated in silicon CMOS technology because of the operating temperature range, power, and noise characteristics of the metal oxide semiconductor FET (MOSFET). Silicon CMOS devices can be designed[2] to operate from room temperature to below 10 K. This MOSFET primer is meant as an introduction only. Detailed explanations of semiconductor physics and MOSFET action are available through many sources.[3,4] Liu and Nagel[5] treat the modeling of MOSFETs.

The *N*-channel MOSFET, shown in Fig. 5.4, is composed of *n*-implanted drain and source regions isolated from each other by the *p*-doped silicon substrate, or *P*-well. A gate, usually composed of polysilicon, lies above a thin dielectric layer (usually SiO_2) on the semiconductor surface between the two diffusions. In the simplest transistor action, a positive gate-to-source voltage V_{gs} induces a field in the surface region of the semiconductor. If the gate voltage is above a specific threshold V_t, the resulting field repels majority mobile carriers (holes) and attracts electrons, forming a very thin inversion region, or *n*-channel, at the surface of the semiconductor. The *n*-channel then provides a current path between the source and drain. The *P*-channel MOSFET is the same as the *N*-channel device but it utilizes opposite doping and voltages.

A sample plot of the drain current versus drain voltage at several gate voltages is shown in Fig. 5.4. Higher gate voltages increase the density of electrons at the surface and thus increase the conductance of the channel between source and drain diffusions. This action is useful in ROICs, because MOSFETs can be utilized as switches for multiplexing signals and resetting integration capacitors. MOSFETs are dimensioned according to the width W and length L of the channel. A short channel length will result in a desirable lower "on" resistance. Lower "on" resistance also can be achieved by either connecting two switches in parallel or by simply increasing the channel width of a single device. When the switch is on, the voltage between the source and drain is very small, and the transistor acts as a voltage-dependent resistor:

$$I_d \text{ (linear)} \approx \frac{W}{L}\mu_n C_0(V_{gs} - V_t)V_{ds} \quad [\text{A}] \ , \tag{5.1}$$

where μ_n is the minority carrier mobility (electron in this case), C_0 is the gate capacitance per unit area, V_{gs} is the gate-to-source voltage, V_t is the threshold voltage, and V_{ds} is the drain-to-source voltage. The "on" resistance of the MOSFET in the linear region of operation is approximately

Table 5.1 Common Readout Symbols, Nomenclature, and Units

Symbol	Description	Units
A_{opt}	Optical area of the detector	cm^2
A_{sf}	Source follower buffer gain	V/V
A_v	Amplifier or buffer voltage gain	V/V
C_{clamp}	Clamp storage capacitor	F
C_{fb}	Feedback capacitance	F
C_0	MOSFET gate capacitance per unit area	F/cm^2
C_{sh}	Sample and hold storage capacitor	F
C_{st}, C_{stray}	Stray node capacitance	F
D^*	Normalized detector signal to noise	Jones
Δf	Power bandwidth	Hz
e_{in}	Input referred noise voltage	volts rms
e_n	Input noise voltage	volts rms
f	Frequency	F
g_m	MOSFET drain current to gate-source voltage (transconductance)	mohs
i_{det}	Detector noise current	amps rms
i_{in}	Input referred noise current	amps rms
i_{int}	Integrated photon-induced current	A
i_n	Input noise current	amps rms
I_d	MOSFET drain current	A
I_{dark}	Detector dark current	A
k	Boltzmann constant	J/K
L	MOSFET channel length	cm
m	MOSFET subthreshold factor	—
NEC	Noise equivalent charge	e^-
NEI	noise equivalent irradiance	$ph/cm^2\ s^{-1}$
q	Electron charge	C
q_c	Noise charge on a capacitor	e^- rms
r_{det}	Detector resistance	Ω
r_{in}	Amplifier input resistance	Ω
r_s	Source (detector resistance)	Ω
R	Resistor	Ω
R_{fb}	Feedback resistance	Ω
R_{on}	MOSFET "on" resistance	Ω
R_0	Zero bias detector resistance	Ω
t_{frame}	Period from frame to frame	s
t_{int}	Signal current integration time	s
T	temperature	K
T_{det}	Detector temperature	K
T_{Rfb}	Feedback resistor temperature	K
v_c	Noise voltage on a capacitor	volts rms
v_{out}	Output referred noise voltage	volts rms
V_{ds}	MOSFET drain-to-source voltage	V
V_{gs}	MOSFET gate-to-source voltage	V
V_t	MOSFET turn on threshold voltage	V
W	MOSFET channel width	cm
Z	Charge to voltage gain	V/C
Greek:		
η	Photon to electron conversion efficiency	e^-/ph
η_{ie}	Photon current injection efficiency	A/A
μ_n	N-channel MOSFET minority carrier mobility	$cm^2/V\ s^{-1}$
$1/f$ noise	Noise of $1/f$ power spectral density characteristics (drift)	

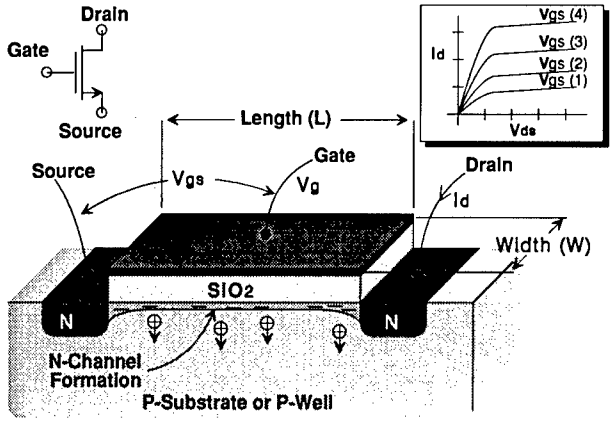


Fig. 5.4 Most ROICs utilize the MOSFET to perform as both amplifier and switch. Drain current is controlled by the gate-to-source voltage.

$$R_{\text{on}}\ (\text{linear}) \approx \frac{L}{W\mu^n C^0(V^{gs} - V^t)} \quad [\Omega] \ . \tag{5.2}$$

This simple explanation of transistor action is adequate, in the case of the switch, where the voltage between source and drain is driven to zero when the switch is on. When the drain-to-source voltage is increased, however, the drain current eventually saturates and no longer increases with the drain voltage. The transistor is then said to be in *saturation*, which is the case with most analog MOSFET amplifiers. The drain saturation current can be approximated as

$$I_d\ (\text{sat}) = \frac{W}{2L}\mu_n C_0(V_{gs} - V_t)^2 \quad [\text{A}] \ . \tag{5.3}$$

Analogous to the "on" resistance of a switch, the most important parameter describing the action of the MOSFET amplifier is its transconductance g_m, which is defined as the change in drain current for a given change in gate-to-source voltage

$$g_m\ (\text{sat}) \approx \frac{W}{L}\mu_n C_0(V_{gs} - V_t) \sim \left(2\frac{W}{L}\mu_n C_0 I_d\right)^{1/2} \quad [\text{mhos}] \ . \tag{5.4}$$

Equations (5.3) and (5.4) approximate the action of MOSFETs that are in strong inversion ($V_{gs} > V_t$). MOSFETs in weak inversion are discussed later in this chapter for the specific case of a direct injection preamplifier.

5.3 TRANSISTOR NOISE

Although most modern ROICs are comprised of MOSFETs and other components formed in CMOS integrated circuit technology, MOSFETs are not always

the best choice for low noise amplification of detector signals. There is a strong relationship between detector impedance and optimum readout technology. Although the silicon MOSFET covers most infrared applications, it is not well suited for all detectors. Specifically, discrete readout preamplifier implementations, popular in systems with few detector elements, can benefit from BJT, JFET, or MOSFET technology.

Detectors can be divided into two impedance categories: low-impedance sensors (lower than 10 kΩ), such as long-wave IR photoconductive HgCdTe detectors, and high-impedance sensors (greater than 10 kΩ), such as photovoltaic, extrinsic silicon, and platinum-silicide detectors. Noise in detectors is generated by the incident photon flux as well as inherent noise sources in the detecting element itself. The input active device of the ROIC preamplifier, a transistor, is usually the dominant noise contributor of the readout. The preamplifier should provide enough gain to balance, if not render negligible, downstream noise source contributions. If this is the case, the input transistor becomes the principal factor in readout noise performance.

The noise contribution of the input transistor is a function of the source impedance presented by the detector. Thus, it is important to match the detector with an appropriate readout input transistor. A schematic representation of a detector and the input transistor of an amplifier are shown in Fig. 5.5. The functions e_n and i_n are the equivalent input noise voltage and current of the transistor and are usually expressed in terms of noise power spectral density versus frequency. It can be generalized that a high transistor noise current, coupled with a high-impedance detector, can result in high-noise gain on the input node. Conversely, a low-impedance detector can tolerate high-noise current but will suffer from high-noise voltages. The equivalent input noise voltage e_{in} for a transistor or preamplifier in a voltage mode amplifier configuration is given by

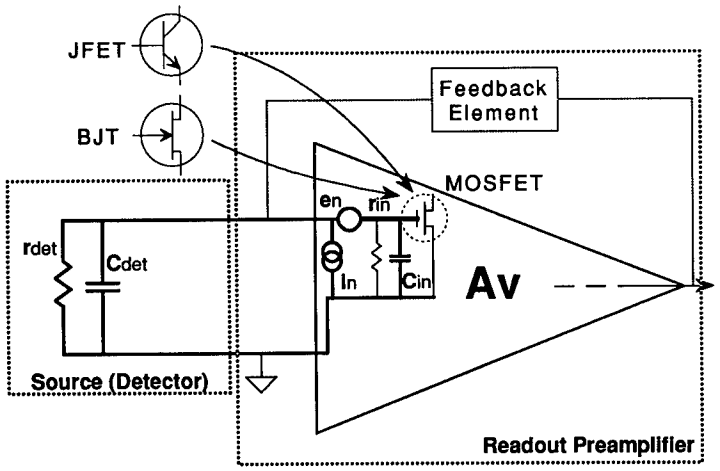


Fig. 5.5 Input referred noise of the preamplifier is usually driven by the input transistor. MOSFETs and JFETs have low i_n, while BJTs have low e_n.

$$e_{\text{in}} \approx \left\{\left[\frac{e_n r_{\text{in}}}{(r_{\text{in}} + r_s)}\right]^2 + (i_n r_{\text{in}} \| r_s)^2\right\}^{1/2} \approx [e_n^2 + (i_n r_s)^2]^{1/2} \quad [\text{V}/\sqrt{\text{Hz}}] \qquad (5.5)$$

for the nominal case where the voltage mode amplifier input impedance r_{in} is very high. The impedance $r_{\text{in}} \| r_s$ is the parallel combination formed by the input impedance of the amplifier (or transistor) and the source resistance. The output referred noise is the product of e_{in} and the voltage gain of the amplifier A_v. In current-to-voltage, or transimpedance, amplifiers, it is useful to express the noise in terms of an equivalent input noise current, i_{in}, instead of a voltage:

$$i_{\text{in}} \approx \left\{\left[\frac{i_n r_s}{(r_{\text{in}} + r_s)}\right]^2 + \left(\frac{e_n}{r_{\text{in}} \| r_s}\right)^2\right\}^{1/2} \approx \left[i_n^2 + \left(\frac{e_n}{r_s}\right)^2\right]^{1/2} \quad [\text{A}/\sqrt{\text{Hz}}] \qquad (5.6)$$

for the nominal case of a transimpedance amplifier where the input resistance is very low. The output referred amplifier noise is the product of i_{in} and the amplifier transimpedance Z_t.

Noise current and voltage for typical low-noise transistor technologies in the common source (or common emitter) configuration are plotted in Fig. 5.6. The noise is typical for low-noise devices such as discrete JFETS and BJTs, as well as for MOSFETs that might be found in a CMOS integrated circuit. Noise data for discrete BJT and JFETS are available from manufacturer's data books, whereas integrated circuit MOSFET noise is typically measured and modeled statistically for a given integrated circuit process and transistor geometry.

Within the system bandwidth of interest, the characteristics of e_n and i_n may not be flat with frequency, that is, white noise. These characteristics may

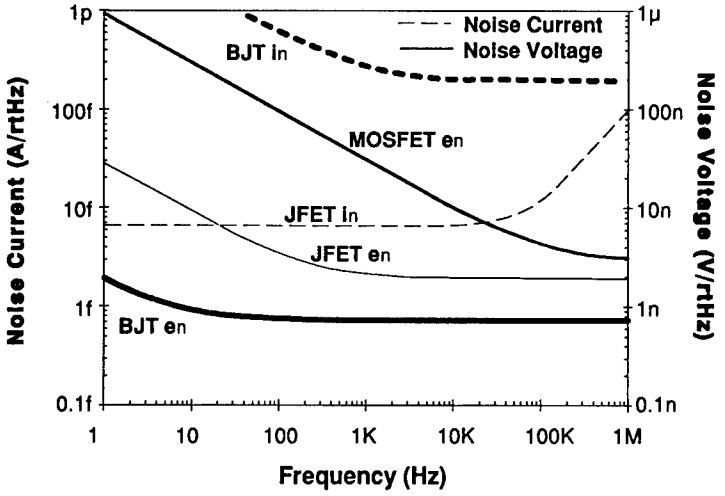


Fig. 5.6 Noise voltage and currents for typical low-noise BJTs, JFETs, and MOSFETS. MOSFET noise current is negligible.

be boosted at low frequencies, as is the case for $1/f$ noise, or boosted at high frequencies. Therefore, the total output rms noise must be calculated by integrating the power spectral densities, plotted in Fig. 5.6, for example, over the equivalent noise power bandwidth required, and then multiplying this by the gain of the amplifier:

$$v_{\text{out}}^2(e_n) = \int_{f_{\text{low}}}^{f_{\text{high}}} A_v^2(f) e_{\text{in}}^2(f)\, df \quad [\text{V}^2] \ , \tag{5.7}$$

or, for a transimpedance amplifier,

$$v_{\text{out}}^2(i_n) = \int_{f_{\text{low}}}^{f_{\text{high}}} Z_t^2(f) i_{\text{in}}^2(f)\, df \quad [\text{A}^2] \ , \tag{5.8}$$

where A_v and Z_t are the voltage gain and current transimpedance, respectively. The total rms output noise voltage from all noise components is summed in the power domain, for example,

$$v_{\text{out}}\ (\text{total}) = [v_{\text{out}}(i_n)^2 + v_{\text{out}}(i_{\text{det}})^2 + v_{\text{out}}(i_{\text{ph}})^2 + \ldots]^{1/2} \quad [\text{volts rms}] \ . \tag{5.9}$$

Figure 5.7 shows the ratio of amplifier (or input transistor) noise to detector thermal noise, as a function of detector source resistance, for the BJT, JFET, and MOSFET examples in the common emitter or source configuration. The calculations for detector (and resistor) thermal noise are covered later in this chapter. Notice that for photovoltaic detectors, which typically have impedances above 1 MΩ, MOSFET noise under the conditions given is lower than the detector thermal noise. MOSFETs, however, are not the optimum choice for detectors with impedance below 100 kΩ, such as some low-resistance photoconductors; bipolar transistors are the best choice in these cases. JFETS,

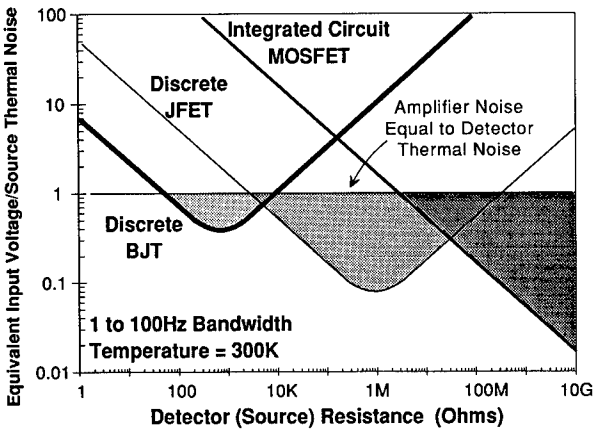


Fig. 5.7 Bipolar transistors offer low noise for low-impedance detector applications, while JFETs and MOSFETs offer low noise for high-impedance detectors.

common in discrete amplifier circuit configurations, offer greater performance than MOSFETs out to 100 MΩ, even at room temperature; and, although not shown, cooled JFETS have continued low-noise performance well beyond 100 MΩ, because shot-noise-inducing input bias current is significantly reduced with temperature.

There are two primary sources of noise in MOSFETs. The white (flat with frequency) channel thermal noise, usually input referred as a voltage from gate to source, is given as

$$e_n = \left(\frac{\frac{8}{3}k\,T\Delta f}{g_m}\right)^{1/2} \quad [\mathrm{V}] \tag{5.10}$$

for both MOSFETs and JFETs.[6] Unpublished data indicate that the actual channel thermal noise on some MOSFETs may be higher than that predicted here.

MOSFET $1/f$ noise comes primarily from current fluctuations caused by surface states at the interface of the MOSFET channel and the gate oxide.[7] These surface states capture an electron (or hole) and release it at a later time.[8] The $1/f$ noise is a strong function of the CMOS process and can vary significantly from one process lot to another. In general, $1/f$ for a given process can be reduced through circuit design by increasing the gate area, WL, or the gate capacitance, C_{0x}, of the MOSFET

$$e_{n(1/f)} \propto \left(\frac{1}{WLC_0 f^{\alpha}}\right)^{1/2} \quad [\mathrm{V}/\sqrt{\mathrm{Hz}}]\ , \tag{5.11}$$

where α is normally modeled as -1.0, but can vary between process lots.

5.4 ROIC PERFORMANCE DRIVERS

Sensor electronic designs, whether discrete or ROIC, are guided by requirements traceable to system performance parameters or input/output interface requirements. Key ROIC requirements are matched with respect to key system or interface requirements in Table 5.2.

The SNR is the prime design driver in most sensor systems. To achieve SNR objectives, trade-offs must often be made between detector temperature, circuit area, and power. Other important drivers include dynamic range, linearity, and operability. All requirements are interrelated and can usually be met given great enough real estate (detector size), power, and a low detector temperature. These conveniences are rarely allowed, resulting in designs requiring many trade-offs and compromises between parameters. The designer should develop a dialog with the sensor user so that the evolving design accurately reflects the users needs.

5.5 ROIC PREAMPLIFIER OVERVIEW

Most ROICs utilize preamplifiers that accumulate detector photon-induced current over a fixed integration time. The detector current, accumulated in an

Table 5.2 Relationships between Key Readout Circuit Requirements and System Performance and Interface Issues

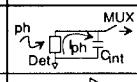
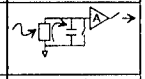
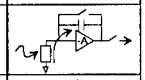
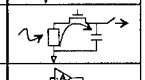
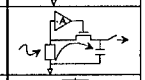
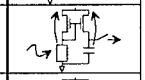
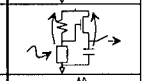
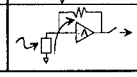
Major Readout Performance Parameters	Related System Parameter or Interface Impact	Comments
NEC (noise equivalent charge)	Sensitivity	Minimized to enhance SNR
Power dissipation	Cooldown time Life Weight	Limited cryogen/cooler life Cryogen weight/cooler size
Dynamic range	Maximum saturation signal	Loss of signal
Crosstalk	System MTF (resolution) Blooming of saturated elements	Element to element
Frequency response	System MTF (resolution) Latent images	Often related to crosstalk
Input impedance	Signal linearity Noise	Detector bias changes with signal Loss of optimum detector bias
Linearity reliability	Calibration Instrument life	Proper identification Confidence of success
Gain	Sensitivity	Signal amplified above system noise floor
Output video driver impedance	Sensitivity MTF	EMI from environment crosstalk between multiplexed elements

integration capacitor or CCD "bucket," results in a signal that is periodically sampled and multiplexed out of the preamplifier. The integration capacitor is reset and the process starts over again. A less common form of ROIC preamplifier provides continuous output voltage or current, which is proportional to the detector current, in lieu of an accumulated signal at the end of the frame time. The most common families of preamplifiers and their typical performance capabilities are given in Table 5.3. It is worth noting that the listed performance capabilities represent typical preamplifier applications and that it is possible, through careful design, to broaden the application space of the various configurations. Each of these configurations is briefly described here and then treated in detail in later sections. Many of these circuits and related topics are also addressed by Nelson, Johnson, and Lomheim.[9]

The self-integrator (SI) has the fewest unit cell components among all readout circuit configurations. Photon-induced charge is integrated over a given frame time directly onto the integration capacitance formed on the detector node, and is periodically transferred out of the unit cell via the multiplexer switch (MUX). The output of the SI is in the charge domain as opposed to the voltage domain. The action of the MUX, while enabled, resets the integration capacitor at the end of each frame.

A buffer amplifier can be added to the SI to provide voltage domain readout to the MUX. This implementation, normally a MOSFET source follower per detector (SFD), requires a reset switch on the detector node to reset the integration capacitor, because the MUX cannot provide reset through the buffer.

Table 5.3 Common ROIC Preamplifiers Selected According to the Limitations and Performance Requirements of the Specific Application

Circuit	Block Diagram	Circuit Description	Minimum Multiplexer Required (MUX)	Typical Noise (e^-) at Frame Rate**	Dynamic Range V_{sat}/V_{ns}	Frequency Response (KHz)	Power Per Cell (μW)	Non-Linearity	Detector Bias Uniformity	Minimum Unit Cell Area (μm x μm)	Irradiance Range at 1KHz F.R. ($ph/cm^2/s$)
SI - Self Integrator		Reset Integrator	CCD or Active Amp. Prior to Video Amp.	400 at 1KHz	≥200 High Noise Low Signal	>200	<0.5μW	Tracks Detector I-V	~1mV to 10mV	<10 x <10	$<1x10^9$ to $\sim3x10^{14}$
SFD - Source Follower/ Detector		Buffered Reset Integrator	Video Amplifier	50 - 100 at 10Hz	~2,000 Low Signal	>200	~0.5μW	Tracks Detector I-V	1mV	<20 x <20	$<1x10^9$ to $\sim3x10^{14}$
CTIA - Capacitor Feedback TIA		Miller Reset Integrator	Video Amplifier	40 - 150 at 10KHz	>10,000 - Low Noise High Signal	>200	3μW - 40μW	Reduced By Amp. Feedback	<0.5mV* (to 40mV)	35 x 35	$<1x10^9$ to $\sim3x10^{15}$
DI - Direct Injection		Injection Reset Integrator	Same as FEDI	80 at 10KHz	~2,500 Medium Noise	>20 at Higher Flux	<0.5μW	Tracks MOSFET & Det. I-V	10mV to 40mV	<20 x <20	$1x10^{12}$ to $\sim4x10^{15}$
FEDI - Feedback Enhanced DI		Injection Reset Integrator	CCD or Active Amp. Prior to Video Amp.	80 at 10KHz	5,000	>50	6μW - 30μW	Reduced By Open Loop Gain	<0.5mV* (to 40mV)	40 x 40	$1x10^{11}$ to $\sim3x10^{15}$
CM - Current Mirror		Gate Modulation Reset Integrator	CCD or Active Amp. Prior to Video Amp.	1,000 at 10KHz	~2,000 High Noise High Signal	>20	<0.5μW	Tracks MOSFET I-V	10 to 40mV	30 x 30	$1x10^{13}$ to $>5x10^{16}$
RL - Resistor Load		Gate Modulation Reset Integrator	CCD or Active Amp. Prior to Video Amp.	10,000 at 10KHz	~2,000 High Noise High Signal	>20	<0.5μW	Tracks Resistor I-V	10 to 40mV	30 x 30	$1x10^{13}$ to $>5x10^{16}$
RTIA - Resistor Feedback TIA		Current to Voltage Amplifier	Video Amplifier	60 - 150 at 1KHz	>10,000 Low Noise High Signal	>50	12μW - 500μW	Reduced By Amp. Feedback	<0.5mV* (to 40mV)	100 x 100	$<1x10^9$ to $>5x10^{16}$

* Lower Detector Bias Offset Achievable Utilizing Auto-Zeroing Type Amplifier
** Detector Capacitance in 0.3pf to 1pF Range

The detector bias of both the SI and SFD, although initially set at the beginning of a frame, changes as detector current integrates onto the input node. This detector debias results in signal output nonlinearity, since the signal characteristics of many detectors change as a function of bias. Neither the SI nor the SFD provides gain within the unit cell.

The capacitor feedback transimpedance amplifier (CTIA) solves the debias and gain limitations of the SFD and SI by incorporating the integration capacitor into the feedback loop of an inverting, high-gain differential amplifier, and thereby forcing charge to integrate into the feedback capacitor instead of the detector node capacitance. This results in a stable detector bias with a more linear signal transfer function. Because its gain is set by the feedback capacitor instead of the detector node stray capacitance, the CTIA can provide a high degree of signal amplification prior to multiplexing. The trade-off is that the CTIA requires significantly more area to implement than either the SI or the SFD.

Direct injection (DI) circuits provide a low-impedance detector interface, via the source of a MOSFET, that helps to keep the detector bias constant. Photon-induced charge integrates onto the capacitor at the drain of the MOSFET. The gain of the DI is set by the integration capacitor and, like the CTIA, gain can be quite high. The DI is limited to medium to high photon flux ranges. This is because the input impedance increases dramatically at low detector current levels, resulting in unstable detector bias, lower photon current integration, and reduced frequency response on the detector node. Also, since the input node of the DI is not reset at the end of each integration period, charge induced

in one frame of data can be integrated during the subsequent frame, resulting in reduced frequency response.

Feedback-enhanced DI (FEDI) incorporates an inverting amplifier between the detector node and the input MOSFET gate. This further reduces input impedance for applications at lower backgrounds. Like the SI, both the DI and FEDI provide charge domain output to the MUX; voltage mode output can be accomplished by adding a SFD to the output of the DI or FEDI.

Due to real estate constraints, it is not possible to accumulate all of the photon-induced charge at very high irradiance levels on a capacitor within the unit cell. Hence, the photon current must be scaled down before integration onto a reasonable size integration capacitor. For such applications, the current mirror (CM) preamplifier utilizes a load MOSFET, biased by the detector photon current, to induce or mirror a smaller current in the drain of a scaled down input MOSFET. The resistor load (RL) circuit provides a similar function by utilizing a resistor load in the detector to induce a lower current in the input MOSFET for accumulation onto the integration capacitor. Both CM and RL circuits suffer the same frequency response and detector bias stability issues as DI circuits.

The RTIA is similar to the CTIA but with the feedback capacitor and reset switch replaced by a resistor. The RTIA does not integrate detector current; rather, it provides a continuous output voltage that is proportional to the detector current. Since it is not reset after data are sampled, the RTIA has limitations in frequency response. It also requires high resistance feedback to provide gain comparable to the CTIA; large resistors require considerable unit cell area while tending to be high in $1/f$ noise or drift.

5.5.1 ROIC Preamplifier Signal-to-Noise Ratio

Sensitivity is normally the most important and challenging user-imposed requirement of a sensor design. Sensitivity can be expressed in terms of SNR or some other related parameter. The benchmark SCA performance parameters for SNR are typically functions of the sensor application, as given in Table 5.4. A common detector sensitivity parameter D^* is not a reasonable sensor flowdown, because high D^* can be achieved without fully addressing the system

Table 5.4 Useful Sensitivity Parameters Related to Sensor Applications

Parameter	Definition	Units	Common Application
NEdT	Noise equivalent temperature	K	Thermal imagers
NEI	Noise equivalent irradiance	ph/cm^2 s^{-1}	Radiometers and photon counters
NEP	Noise equivalent power	W	Imagers and radiometers
NEC	Noise equivalent charge (in a data frame)	e^-	Measure of ROIC, preamplifier, and CCD sensitivity Astronomy Visible imagers
D^*	Signal-to-noise normalized to photon energy, optical area, and system power bandwidth	A $\sqrt{\text{Hz}}$/W (Jones)	Comparison of detector technologies

SNR requirements. However, D^* is a valued tool for comparing one detector technology to another, since it is normalized to signal power, signal power bandwidth, and detector optical area.

Readout and SCA sensitivity are most often expressed in terms of any number of input referred noise quantities. Noise equivalent sensitivity parameters can be interpreted as the signal level that would result in a SNR of unity. The most universal measurement of readout sensitivity is noise equivalent charge (NEC); that is, the equivalent input referred noise electrons accumulated over a single frame or sample of data. NEC is specified with a given signal bandwidth or frame rate of interest. For a readout circuit that accumulates photon charge on an integration capacitor, NEC is the equivalent noise charge on the integration capacitor. The most commonly specified SCA level sensitivity parameter is noise equivalent irradiance (NEI), or the equivalent input irradiance for a SNR of unity; NEI holds the designer to a specific SNR at the sensor output. NEC can be referred to the input of the sensor, in terms of NEI through

$$\text{NEI} = \frac{\text{NEC}}{\eta A_{\text{opt}} t_{\text{int}}} \quad [\text{ph/cm}^2\ \text{s}^{-1}] \ , \tag{5.12}$$

where η is the quantum efficiency of the detector, A_{opt} is the detector optical area, and t_{int} is the integration time over which photon charge accumulates.

5.5.2 System Sensitivity Flowdown

In general, the readout circuit is an integral part of the sensor design and should not be separated from the SCA or higher level assemblies. Therefore, sensor requirements must be partitioned into detector, readout, and other subcomponent parameters before design work can begin. Although subcomponents such as off-focal-plane amplifiers and digitizers require an allocation of the specification space, these parts are well understood and will be assumed to have negligible system impact in the following analyses.

An example flowdown of sensor sensitivity to the detector and readout begins by dividing the SCA level NEI into two equal noise contributors. This initial allocation is made after contributions from the photon flux Q_B are removed from the system NEI:

$$\text{NEI}_{\text{SCA}} = (\text{NEI}_{\text{sys}}^2 - \text{NEI}_{ph}^2)^{1/2} = \left(\text{NEI}_{\text{sys}}^2 - \frac{Q_B}{\eta A_{\text{opt}} t_{\text{int}}}\right)^{1/2} \text{[photons]} \tag{5.13}$$

for photovoltaic detectors. In this case, the sum of the readout contribution, NEI_{ro}, and the detector contribution, NEI_{det}, yields the SCA total (NEI_{SCA}). The initial detector and readout allocations are

$$\text{NEI}_{ro} = \text{NEI}_{\text{det}} = \frac{\text{NEI}_{\text{SCA}}}{\sqrt{2}} \quad \text{[photons]} \ , \tag{5.14}$$

which is an approximation, since it does not allocate NEI to other second-order noise contributors such as the multiplexer, video driver, off-focal-plane electronics, and the digitizer.

Detector NEI. The value of $\mathrm{NEI_{det}}$ is determined, to first order, from the noise contribution of a photovoltaic detector in reverse bias[6]:

$$i_{\mathrm{det}} = (2qI_{\mathrm{dark}}\Delta f)^{1/2} \quad [\text{amps rms}] \ , \tag{5.15}$$

where the detector noise current is assumed to be white thermal noise, q is the electron charge, Δf is the measurement bandwidth, and I_{dark} is the backbias dark current of the detector. The total (rms) detector noise is determined by integrating the noise and noise transfer function over all frequencies. For a photon integrating sensor, in which detector charge is accumulated and periodically read out before reset, the transfer function of the accumulator drops the noise power bandwidth and noise rolls off at a frequency of approximately

$$f \approx \frac{1}{2T_{\mathrm{frame}}} \quad [\mathrm{Hz}] \ , \tag{5.16}$$

where T_{frame}, the period between frames, is approximately equal to the integration time. (The shape of various transfer functions in a sampled system will be covered in a subsequent section.) Since the shot noise in Eq. (5.15) is white, i.e., flat over all frequencies, the noise power bandwidth of the accumulator can be substituted for Δf. The total number of noise electrons, or NEC, accumulated in a given integration time is

$$\mathrm{NEC_{det}} = \frac{i_{\mathrm{det}}t_{\mathrm{int}}}{q} \quad [\text{electrons}] \ . \tag{5.17}$$

To fit within the allocated NEI,

$$\mathrm{NEI_{det}} > \frac{\mathrm{NEC_{det}}}{\eta A_{\mathrm{opt}}t_{\mathrm{int}}} = \left(\frac{I_{\mathrm{dark}}/t_{\mathrm{int}}q\eta^2}{A_{\mathrm{opt}}}\right)^{1/2} \quad [\mathrm{ph/cm^2\ s^{-1}}] \ , \tag{5.18}$$

resulting in a maximum dark current of

$$I_{\mathrm{dark}} < \mathrm{NEI_{det}}^2\eta^2 qA_{\mathrm{opt}} \quad [\text{amps rms}] \ . \tag{5.19}$$

As a note, the NEI contribution from the photon-induced noise is calculated by replacing the dark current I_{dark} with photon current I_{ph} in the preceding equations.

For detector operation near zero bias, where dark current is negligible, the detector thermal noise can be expressed in terms of the small-signal detector resistance at zero bias, R_0, and the temperature of operation, T_{det}:

$$i_{\mathrm{det}} = \left(\frac{4kT_{\mathrm{det}}\Delta f}{R_0}\right)^{1/2} \quad [\mathrm{A}] \ , \tag{5.20}$$

where k is the Boltzmann constant. This yields a maximum detector noise contribution of

$$\mathrm{NEI_{det}} > \frac{(2kT/t_{\mathrm{int}}R_0)^{1/2}}{q\eta A_{\mathrm{opt}}} \quad [\mathrm{ph/cm^2\ s^{-1}}] \ , \tag{5.21}$$

and a minimum zero-bias resistance,

$$R_0 > \frac{2kT}{t_{\mathrm{int}}\mathrm{NEI}_{\mathrm{det}}^2 q^2 \eta^2 A_{\mathrm{opt}}^2} \quad [\Omega] \ . \tag{5.22}$$

Both of these expressions for detector thermal noise, Eqs. (5.19) and (5.22), are approximations that fit well in specific regions of detector operation. Equation (5.21) is most commonly used in infrared photovoltaics, because detector bias is usually minimized to reduce dark current and, therefore, detector $1/f$ noise (drift). Detector performance is often expressed in terms of R_0A_{opt}, or I_{dark}, and η due to their dependence on noise as expressed in Eqs. (5.18) and (5.21). If the selected detector technology shows significant margin in these parameters, more of the total noise budget can be allocated to the readout or other downstream noise sources.

Readout NEI. Readout noise, which is a function of the circuit configuration and layout details, is more difficult to analyze than detector noise. All readout circuits can be conceptualized in a format similar to that shown in Fig. 5.8, in which the readout noise sources are combined into an equivalent input referred noise voltage e_n and noise current i_n. Both parameters are usually dominated by the noise of the preamplifier input transistors and their sampled transfer function to the output. In determining the values of e_n and i_n, the folding, or aliasing, of high-frequency noise into the system bandwidth of interest must be considered.

Once the appropriate input noise contributions are determined, the readout amplifier noise current is handled in the same way as detector noise current in Eq. (5.17); however, the input referred noise voltage must be converted to an equivalent current prior to applying it to Eq. (5.17):

$$i_{en} = \frac{e_n}{r_{\mathrm{det}} \| r_{\mathrm{in}}} \quad [\mathrm{amps\ rms}] \ , \tag{5.23}$$

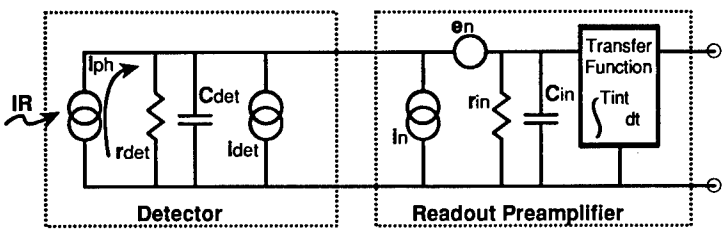


Fig. 5.8 Simple model of the sensor includes detector noise current and amplifier input referred noise voltage and current. Most preamplifiers can be conceptualized in this manner.

where $r_{\text{det}}\|r_{\text{in}}$ is the parallel combination of the small-signal detector and the amplifier input resistances. The NEC of the readout, with a fixed integration time of t_{int}, is

$$\text{NEC} = \left[i_n^2 + \left(\frac{e_n}{r_{\text{det}}\|r_{\text{in}}} \right)^2 \right]^{1/2} \frac{t_{\text{int}}}{q} \quad \text{[electrons]} \tag{5.24}$$

and is input referred to the detector as

$$\text{NEI}_{ro} > \frac{\text{NEC}}{\eta A_{\text{opt}} t_{\text{int}}} \quad \text{[photons]} \ . \tag{5.25}$$

Equations (5.24) and (5.25) can be utilized to determine the maximum noise contribution of the readout through Eq. (5.13). The results will guide the readout circuit type and design. Allocation of sensitivity should be made to detector drift, as well as to other less critical components not covered here, such as off-focal-plane electronics.

5.6 READOUT PREAMPLIFIERS

This section covers many of the most common readout preamplifier circuit families. Although the focus is on integrated circuit forms of the preamplifier, the RTIA is covered because it is utilized in sensors with few detector elements, it can be fabricated in ROIC format, and it provides an introduction to noise concepts necessary for subsequent preamplifier circuits.

5.6.1 Analysis of the Resistor Transimpedance Amplifier

The analysis tools discussed can be demonstrated and expanded by example for a common detector preamplifier circuit, the resistor transimpedance amplifier. The RTIA, usually employed in discrete rather than integrated circuit configurations, is useful in systems requiring only a few pixels (detector elements). The RTIA is also readily adaptable to ROIC format and may be combined with a sample/hold stage and multiplexers for application in sensors with larger numbers of elements. As with most photovoltaic readout amplifiers, the RTIA is a transimpedance or current-to-voltage amplifier. Unlike most preamplifiers, it does not accumulate, or integrate, signal charge over a given frame time; instead it presents a continuous output signal proportional to the input (photo) current. A schematic of a generic RTIA in a focal plane application, along with the predominant noise sources, is presented in Fig. 5.9.

The output response of the RTIA to a photon-generated current, or to any input current including noise, is

$$Z_t(f) = \frac{V_{\text{out}}}{I_{ph}} = \frac{R_{fb}}{[1 + (2\pi f R_{fb} C_{st})^2]^{1/2}} \quad \text{[V/A]} \ . \tag{5.26}$$

The in-band low-frequency transimpedance Z_t is set by R_{fb}, which is selected to provide sufficient gain to the system and to minimize its thermal noise contribution.

Noise contributed by the feedback resistor can be represented as a noise current across either the resistor or the input node by[6]:

$$i_{Rfb} \approx \left(\frac{4kT\Delta f}{R_{fb}}\right)^{1/2} \quad [\mathrm{A}] \ . \tag{5.27}$$

It is normally desirable to keep the feedback resistor noise below the detector thermal noise of Eq. (5.20); this results in a preliminary selection of R_{fb} as

$$R_{fb} > R_0 \frac{T_{Rfb}}{T_{\mathrm{det}}} \quad [\Omega] \ . \tag{5.28}$$

If the feedback resistor is on the focal plane and its temperature, T_{Rf}, is the same as that of the detector, T_{det}, then R_{fb} must be greater than the zero-bias resistance of the detector, R_0. If R_{fb} is too large, the very high amplifier transimpedance will result in saturation of the output signal.

If photon-induced noise is greater than detector thermal noise, then the feedback resistor is optimized to have lower noise than the photon-induced noise current i_{ph}:

$$i_{ph} = (2qI_{ph}\Delta f)^{1/2} \quad [\mathrm{A}] \ , \tag{5.29}$$

then

$$R_{fb} > \frac{2kT_{Rfb}}{I_{ph}} \quad [\Omega] \ . \tag{5.30}$$

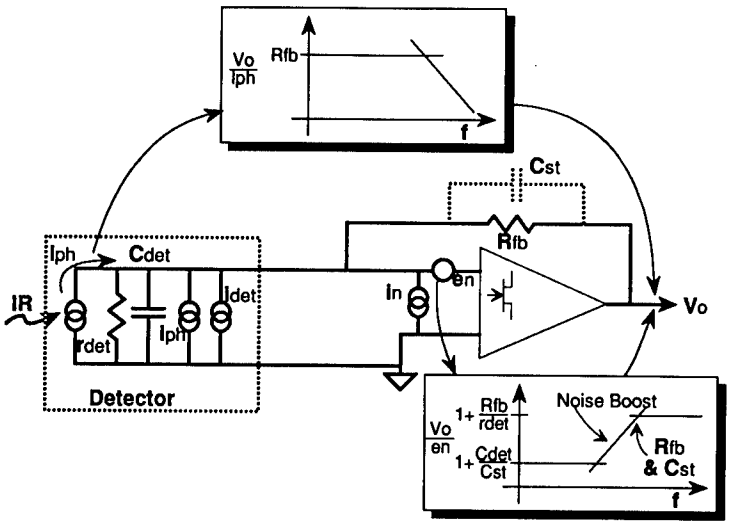


Fig. 5.9 The limits of the RTIA amplifier function are dominated by stray capacitance elements across the feedback resistor and on the input node.

Equations (5.29) and (5.30) can be used as guidelines for determining a reasonable transimpedance for the preamplifier. It may be desirable to have a higher R_{fb} than expressed above in order to further reduce its noise contribution and thus further limit its impact on the total noise, or to provide higher gain upfront and thus reduce the impact of downstream noise sources.

The feedback resistor must be selected with careful consideration to drift, or $1/f$ noise, which is common in high-ohmage resistors. In general, large surface area resistors exhibit low $1/f$ but, because of size, these are not easily integrated into cold focal plane configurations. Before committing to a final design, sample resistors should be measured both for $1/f$ and temperature stability, because any change in transimpedance with temperature can manifest as noise.

After selecting an appropriate transimpedance for the amplifier, the design should focus on selection of a low-noise-differential amplifier for RTIA implementation. Shot noise current of the amplifier, which is dependent on the input bias current of the input transistor, I_{in}, is governed by the same equation that predicts noise from photon current, and can be estimated at the input node as

$$i_n = (2qI_{in}\Delta f)^{1/2} \quad [\text{A}] \ . \tag{5.31}$$

Input bias current should be much less than the photon current, and in such cases shot noise is insignificant. For JFETs, this approximation can lose accuracy at very low frequencies where transistor $1/f$, or drift components, can dominate. Also, since input bias current is a strong function of temperature for JFETs, i_{in} can be negligible at cryogenic temperatures. Shot noise in MOSFETs is insignificant compared to the thermal noise of most detectors, provided no leaking p-n junctions or other stray current paths exist on the input node.

Transistor input referred noise voltage, e_n, does not have the same transfer function to the output as photon signal current in the RTIA. In the case of MOSFETs and JFETs, this noise is predominantly the channel thermal and $1/f$ noise described earlier in Eq. (5.10). The most straightforward way to analyze the impact of the input referred noise voltage is to refer to it as a noise current, by first determining its characteristics at the output node (through the amplifier) and then dividing by the circuit transimpedance. This allows all noise sources on the input node to be directly compared and summed as a single noise current. In Fig. 5.9, the transfer function of e_n to the output node is flat at low frequencies. At higher frequencies the stray capacitance on the input node "boosts" the voltage gain of the transfer function and noise gain increases. Eventually, the response goes flat again as parasitic or other capacitors across the feedback resistor become the dominant transimpedance element:

$$\frac{v_{\text{out}}}{e_n} = \left\{1 + \frac{R_{fb}}{R_{\text{det}}}\left[\frac{1 + (2\pi f C_{\text{det}} R_{\text{det}})^2}{1 + (2\pi f R_{fb} C_{fb})^2}\right]\right\}^{1/2} \quad [\text{V/V}] \ . \tag{5.32}$$

The input referred equivalent noise current caused by e_n of the RTIA is the ratio of Eq. (5.32) to Eq. (5.26):

$$i_{en}(f) = e_n\left(\frac{1}{R_{fb}} + \frac{1}{R_{\text{det}}}\right)\left[1 + (pfC_{\text{det}}R_{\text{det}})^2\right]^{1/2} \quad [\text{A}/\sqrt{\text{Hz}}] \ . \tag{5.33}$$

The total input referred noise current of the dominant RTIA noise sources is the sum of the photon, detector, feedback resistor, and transistor noises in the power domain:

$$i_{\text{total}} = (i_{ph}^2 + i_{\text{det}}^2 + i_n^2 + i_{en}^2)^{1/2} \quad [\text{A}/\sqrt{\text{Hz}}] \ . \tag{5.34}$$

The total noise should be compared to the system noise allocation discussed previously.

A detector and amplifier combination, such as the RTIA with a photovoltaic detector, are said to be "BLIP" (background limited performance) if the total noise is equal to the photon noise. The percent BLIP is calculated by dividing the photon noise by the total sensor noise, in this case Eq. (5.34),

$$\%\ \text{BLIP} = \frac{i_{ph}}{I_{\text{total}}} \times 100 \ . \tag{5.35}$$

The same BLIP can also be calculated via other sensitivity parameters such as D^*, SNR, and NEI.

5.6.2 Reset Integrators and Sampled Readout Circuits

The RTIA is commonly applied in a discrete component configuration in systems with few detector elements. However, the nature of most SCA applications is such that hundreds to thousands of detector elements must be amplified and multiplexed onto a few output lines in an integrated circuit configuration. In many cases, simple integrated circuit capacitors and switches act as the gain setting element in place of the resistor utilized in the RTIA. To design for and analyze the sensor sensitivity requires additional steps beyond those discussed in the previous section. A common problem encountered in switched circuit analysis is the aliasing of high-frequency signals or noise into the bandpass of the sampled signal, thereby producing a false signal or increasing noise beyond normal expectations.

An example of this aliasing effect can be seen in the spoked cartwheels of Western movies. The wheels appear to be moving slowly, stopped, or even reversed although the wagon is obviously moving forward at a high rate of speed. The cartwheel spokes are sampled by the film at a rate of 30 Hz. If the wheel passes a reference point in the camera's field of view at a rate of 29 spokes per second, the wheel will appear to be moving in reverse at about 1 spoke per second; this is because each subsequent spoke is imaged in a position slightly behind that of the preceding spoke in the previous frame. If the spokes are moving at exactly 30 or 60 per second, the camera will image subsequent spokes in the same position on each frame of film, giving the appearance of no spoke, or wheel, movement at all. These same effects are seen in sampled sensor systems, where undesirable high-frequency noise, signal, or interference appears at frequencies below half the scene sample rate, the so-called Nyquist rate or limit.

Sensors can be divided into two general classes of input preamplifier types: (1) those, such as the previously described RTIAs, that amplify or create a continuous output voltage representing the photon generated signal, and (2) those

that additionally accumulate and average (or integrate) the photon signal over a specified frame time. This second type of sensor is commonly referred to as a *reset integrator*. The resulting signal from either type of circuit can be multiplexed with other elements. In the analysis that follows, the reset integrators are assumed to integrate during the entire frame time except for a very short reset period.

5.6.3 Self-Integrating Readout

Integrating preamplifiers are most commonly used in ROIC designs where the charge accumulation element is a capacitor or a CCD. The capacitor accumulates photon-induced current over an integration period, resulting in a signal that can be multiplexed to the sensor output. Figure 5.10 shows one of the most basic integrating preamplifier unit cell designs, the self-integrator. The SI has advantages over all other designs in that it is composed of a single unit cell readout component, a small MOSFET switch, and thus requires minimum unit cell area. The drawback is that the SI does not provide signal gain in the unit cell and is thus subject to multiplexer and column amplifier noise.

Photon charge in the SI integrates onto the stray capacitance in the unit cell, which is formed primarily by the detector but also includes strays from the interconnect and the MOSFET switch. If necessary, additional capacitance can be added to the readout to increase storage capacity. After charge is integrated for an entire frame, the multiplexer is cycled and the stored charge is transferred onto the low-input impedance transimpedance column amplifier of gain Z. A reset switch on the multiplexer bus normally restores the detector to its pre-integration voltage bias after the amplifier drives the signal out of the multiplexer. However, the action of the transimpedance amplifier may be sufficient to perform the detector reset without the addition of a separate reset switch on the bus. The transfer function of the circuit from photon current to output of the column transimpedance amplifier, over the integration period t_{int}, is

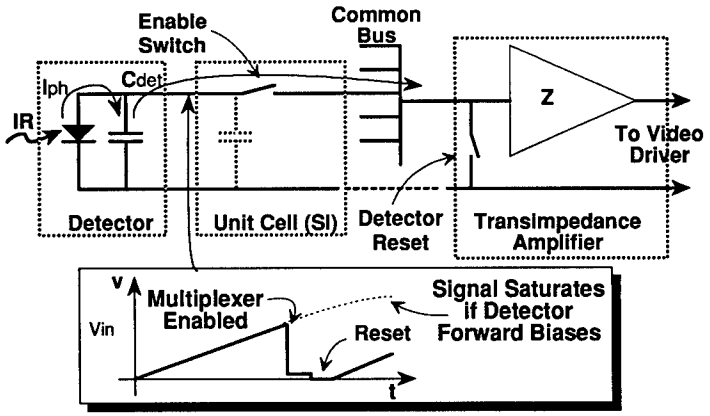


Fig. 5.10 The self-integrator accumulates photon charge on stray capacitance on the input node but provides no gain in the unit cell.

$$V_{\text{out}} = Z\int_0^{t_{\text{int}}} I_{ph}(t)\,dt + ZQ_r \quad [\text{V}] \ , \tag{5.36}$$

where Q_r is the initial charge stored on the input node and Z is the charge-to-voltage gain.

As the SI integrates photon and dark current onto the detector node capacitance, the voltage bias of the detector forms a ramp, as illustrated in Fig. 5.10. If allowed to integrate too long, however, the detector will become forward biased, resulting in response nonlinearity and additional detector shot noise, as photon-induced charge is shunted across the detector. The ramp on the detector node will follow the current to voltage (I-V) characteristics of the detector. For a photovoltaic detector, the voltage on the input node will follow the characteristics of a diode, which exhibits high nonlinearity as its voltage transitions from reverse to forward bias. The dynamic range of the SI is thus limited to the flat back bias signal excursion of the detector.

***kTC* Noise.** Thermal noise (commonly referred to as kTC noise) is stored on the detector node capacitance at the moment the switch or multiplexer of the SI is disabled (opened). This results in small variations in detector voltage or input node charge (ΔQ_r) from one reset time to the next, which in turn results in variation or noise at the end of integration according to Eq. (5.36). This variation in reset voltage is true for any switched capacitor. The thermal noise characteristics of the circuit, when the switch is closed, are the same as in any resistor-capacitor circuit with a resistance R_{on}:

$$e_R = (4kTR_{\text{on}}\Delta f)^{1/2} \quad [\text{V}] \ . \tag{5.37}$$

The rms noise on the capacitor can be calculated by integrating the thermal noise of the resistor over all frequencies through the transfer function of the RC circuit:

$$v_c^2 = 4kTR_{\text{on}}\int_0^{\alpha} \frac{1}{1 + (2\pi f C R_{\text{on}})^2}\,df \quad [\text{V}^2] \ , \tag{5.38}$$

which can be rewritten as

$$v_c = \left[4kTR_{\text{on}}\frac{\pi}{2}f(-3\text{ dB})\right]^{1/2} = \left(\frac{KT}{C}\right)^{1/2} \quad [\text{volts rms}] \ . \tag{5.39}$$

Notice that the noise produced by resetting a capacitor through a switch is not dependent on the resistance of the switch. In Eq. (5.39), the noise equivalent bandwidth [Δf, of Eq. (5.37)] is the product of $\pi/2$ and the -3-dB bandwidth of the circuit, which is true for all white noise sources acted on by a single pole filter. The noise voltage in Eq. (5.39) can also be expressed in terms of noise charge stored on the capacitor by multiplying by the node capacitance:

$$q_c = (kTC)^{1/2} \quad [\text{C}] \ . \tag{5.40}$$

Since kTC noise charge goes up with increased capacitance, it is normal practice to minimize these strays on the input node of the SI. In the case where photon noise dominates, Eqs. (5.29) and (5.40) can be combined to determine the maximum permissible stray capacitance on the input node

$$C\ (\max) \ll \frac{i_{ph} t_{\text{int}}^2}{kT} \quad [\text{F}] \ . \tag{5.41}$$

Unless special care is taken in the selection of capacitor size, kTC noise can be a dominant noise source in sampled integrated circuit applications. Minimizing this capacitance places special requirements on the detector capacitance; since the detector is normally the dominant contributor to input node capacitance, kTC noise must be considered in all switched capacitor circuits.

Integrated Noise Transfer Function. Individual contributions of all noise sources can be determined by applying the appropriate switched noise transfer function to each source. Noise current on the input node of any integrating amplifier, including the SI, is the simplest to analyze because it has the same transfer function as the signal current I_{ph}. For this example, the integrated current of the SI forms a voltage on the input node and is reset (transferred) at the end of the integration time, thus possessing the characteristics of the reset integrator. The change in voltage at the end of integration represents, for the SI, the total charge shifted out at the end of integration.

A reset integrator can be modeled as illustrated in Fig. 5.11. The voltage on the input node, shown as a continuous ramp, is the integral of the detector current $x(t)$. The resulting output voltage $y(t)$ can be expressed in terms of the convolution of the input current and the impulse response $h(t)$:

$$y(t) = x(t) * h(t) \quad [\text{V}] \ . \tag{5.42}$$

The impulse response of the reset integrator $h(t)$ over the integration period t_{int} is

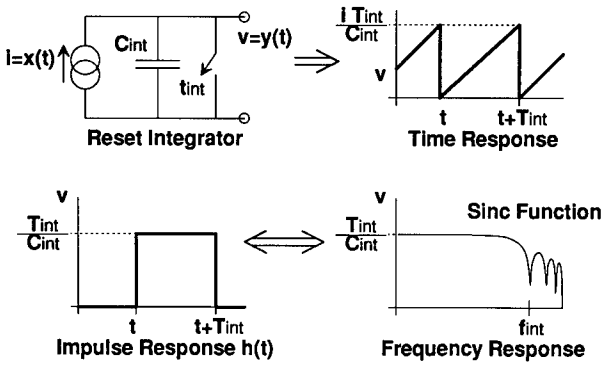


Fig. 5.11 The reset integrator action can be modeled as a continuous integrator that is sampled and subtracted as discrete time increments.

$$h(t) = 1/C_{\text{int}} \int_{t-}^{t-+t_{\text{int}}} \delta(x)\, dx \quad [\text{V}] \ , \tag{5.43}$$

which is simply a step to an amplitude of $1/C_{\text{int}}$ at time equal zero, and returns to zero (is reset) after t_{int}. The integral of the delta function is often referred to as a *window function,* because it applies only to a single integration time, and is zero before integration starts $(t-)$ and after the end of integration.

The frequency response of the reset integrator can be determined by evaluating the Laplace transform:

$$Y(f) = X(f)H(f) \quad [\text{V}] \ , \tag{5.44}$$

where the Laplace transform of $h(t)$ of Eq. (5.43) is

$$H(f) = \frac{t_{\text{int}}}{C_{\text{int}}} \frac{\sin(\omega t_{\text{int}}/2)}{\omega t_{\text{int}}/2} \quad [\text{V/A}] \ , \tag{5.45}$$

where

$$\omega = 2\pi f \quad [\text{rad/s}] \tag{5.46}$$

$$H(f) = \frac{t_{\text{int}}}{C_{\text{int}}} \frac{\sin(\pi f t_{\text{int}})}{\pi f t_{\text{int}}} = \frac{t_{\text{int}}}{C_{\text{int}}} \frac{\sin(x)}{x} \quad [\text{V/A}] \ . \tag{5.47}$$

The sine expression of the form $\sin(x)/x$ in Eq. (5.47) is commonly referred to as the *sinc function.* This derivation of Eq. (5.47) is strictly valid for an ideal reset integrator that is reset instantaneously in time. In practice, however, some dead time exists between the reset and the onset of integration (due to switching time). Equation (5.47) provides a good approximation in cases where the dead time is a small fraction of the integration time.

Signals that are integrated over time and then reset have the sinc transfer function shown in Fig. 5.12, and are in fact attenuated at frequencies higher than the Nyquist limit. Figure 5.13 shows the result of integrating a slowly changing (dc) photon signal with a high-frequency sine wave riding on it; the high-frequency ac component could be either signal, noise, or interference picked up on the input node. A large net accumulation of signal charge results in this case, through the integration of the dc signal over the entire integration time. The ac noise is also accumulated but tends to cancel out its own excursions above and below the dc signal, thus resulting in a significant reduction in amplitude. Since the integration time is assumed to be nearly equal to the time between frames, the only unaliased frequencies that can pass through the reset integrator are those less than half the frame rate. All other residual energy above the Nyquist frequency is attenuated and "folded" down between Nyquist and zero, thus resulting in higher noise than would otherwise be expected in the passband.

In addition to *kTC* noise, other noise sources of the SI include detector thermal noise and photon noise, as discussed previously. Also, the column amplifier is usually the major active noise source in most SI designs, although

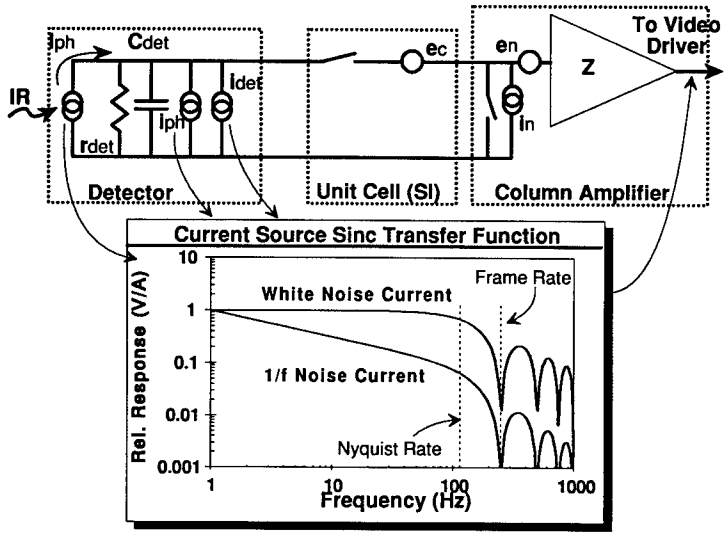


Fig. 5.12 Simple integrator of the SI provides basis of a sample noise model. Integrated current signal and noise sources have a sinc transfer function.

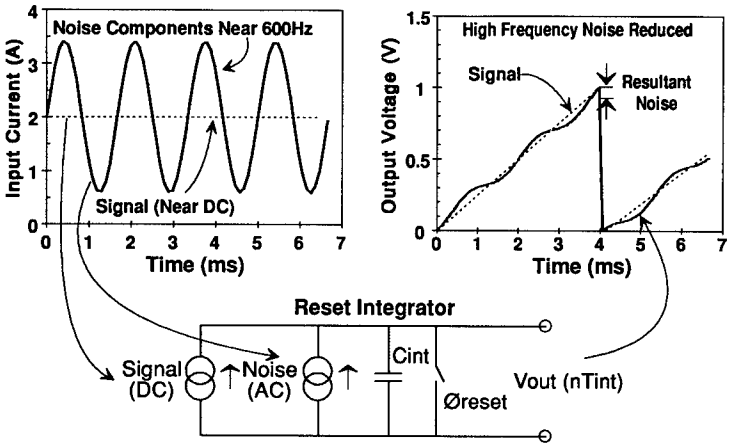


Fig. 5.13 Noise above the Nyquist frequency is attenuated in integrating amplifiers, reducing any aliasing effects in the passband.

it is outside the unit cell. The noise contribution of the column amplifier is dependent on its type, which can be any one of several unit cell amplifiers discussed later in this section.

5.6.4 Source Follower per Detector Readout

The source follower per detector readout preamplifier (SFD), shown in Fig. 5.14, is similar to the SI but with the addition of a MOSFET for voltage mode output and a reset switch. Only three components, other than the detector,

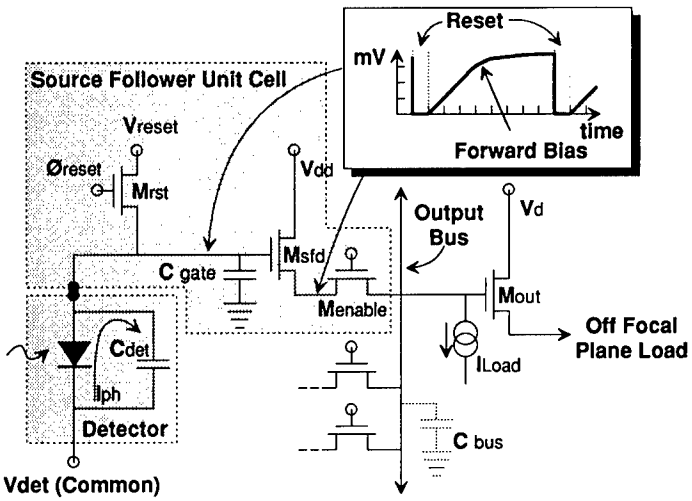


Fig. 5.14 SFD buffers the voltage resulting from photon charge integration to the video output driver via a source-follower MOSFET configuration. The SFD requires a reset switch on the input node.

are contained in the unit cell; this makes the SFD desirable for applications that require small unit cells.

Photon current is integrated onto the input node stray input capacitance formed by the gate of the source-follower MOSFET, the interconnect, and the detector capacitance. Unlike the SI, the ramping input voltage of the SFD is buffered by the source-follower MOSFET and then multiplexed, via an enable switch, to a common bus prior to the video output buffer. The SFD, which has a voltage mode output, requires no bus amplifier. After the multiplexer read cycle, the input node is reset and the integration cycle begins again. The dynamic range of the SFD is limited in the same way as that of the SI, by the current voltage characteristics of the detector. Charge on the input node creates a voltage at the end of integration according to

$$V_{\text{in}} = \frac{I_{\text{det}} t_{\text{int}}}{C} \quad [\text{V}] \ , \tag{5.48}$$

where I_{det} is the sum of photon and dark currents and C is the stray capacitance on the readout input node. The total excursion of input voltage should not exceed the detector forward bias, or severe signal nonlinearities and additional shot noise can occur. In photovoltaic sensors, this limit can be as low as tens of millivolts or as high as several volts. Additional charge storage can be accommodated by adding an integrated circuit capacitor to the input node.

The buffer MOSFET of the SFD is designed both for low noise and near unity gain. This gain, between the detector node and multiplexer, approaches unity as the MOSFET g_m and load resistance R_l are increased:

$$A_{sf} = \frac{1}{1 + (1/g_m R_l)} \quad [\text{V/V}] \ , \tag{5.49}$$

where R_l is the combined load resistance presented by the multiplexer, current source load, and video driver input circuits.

MOSFET drain current in the source follower is not driven by noise; rather, it is driven by g_m requirements and the signal bandwidth of the multiplexer. The current source load for the SFD is located opposite the multiplexer, outside the unit cell, so that power only dissipates in the unit cell MOSFET when it is enabled through the multiplexer onto the video buffer line. The source-follower load current must be large enough to drive, or slew, the capacitance of the multiplexer bus. This is covered in detail for the source-follower video driver in Sec. 5.9.

SFD Noise. The SFD is typically utilized in low background applications, such as astronomy,[10] where long integration times accumulate sufficient charge for readout. Gain is fixed by the near-unity gain of the MOSFET source follower and the integration capacitor, which is usually dominated by the detector. In high-sensitivity applications, the detector capacitance must be reduced to very low levels and thus provide high input node gain so that amplifier and multiplexer noise is overcome. Again the *kTC* noise caused through the reset of the input node must be considered. The major noise sources of the SFD are *kTC*, MOSFET 1/*f*, and MOSFET channel thermal noise.

It is important to address 1/*f* in the input MOSFET because this can often dominate the channel thermal noise to frequencies well above 10 kHz. MOSFET 1/*f* noise, as discussed in an earlier section, decreases as the product of the channel width and length (W and L) increases. The product of W and L also drive gate capacitance and, in Fig. 5.15, the relative total 1/*f* noise is shown as a function of the MOSFET gate-to-detector capacitance ratio. When gate capacitance is much lower than detector capacitance, the photon-to-voltage gain is high but the 1/*f* noise is also high. When gate capacitance is much larger than detector capacitance, the gain of the SFD is substantially reduced according to Eq. (5.48) and SNR is once again reduced. Optimum gate capacitance occurs at the point where it is equal to the detector capacitance. In practice, the detector capacitance is often sufficiently large such that the MOS-

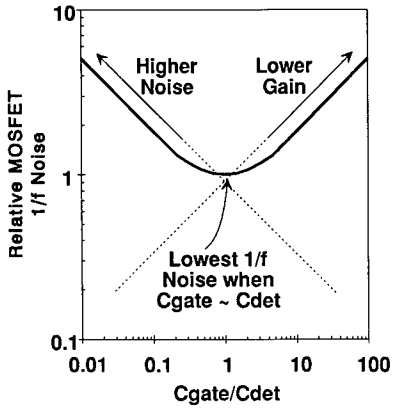


Fig. 5.15 Minimum MOSFET 1/*f* noise contribution occurs when gate capacitance is equal to detector capacitance.

FET size, and therefore gate capacitance, cannot be fully optimized due to area limitations in the unit cell.

MOSFET noise at the output of the SFD is the product of MOSFET input referred noise voltage (whether $1/f$ or channel thermal noise), source-follower gain (~ 1), and square root of the equivalent noise bandwidth of the noise, Δf. Periodic sampling through the multiplexer results in the folding of high-frequency noise below the Nyquist frequency of the circuit. In the case of channel thermal noise, the equivalent noise bandwidth, Δf of Eq. (5.10), is the only parameter to optimize for lowering noise.

The following discussion addresses the MOSFET channel thermal contribution to source-follower noise. MOSFET $1/f$ noise, to first order, is geometry and process dependent and not directly dependent on drain current or g_m. To reduce MOSFET channel thermal noise, the designer's degree of freedom is limited to reducing the noise bandwidth Δf since increasing g_m of the MOSFET does not have the obvious effect on noise that Eq. (5.10) implies. An increase in g_m, as a result of increased drain current, W/L ratio, or other means, results in a corresponding increase in the bandwidth Δf, thus canceling any low-noise benefit.

Noise Equivalent Bandwidth. Noise in the SFD can be controlled by reducing the noise equivalent bandwidth Δf of the circuit. The concept of noise equivalent bandwidth was mentioned previously in the analysis of capacitor kTC noise; here, noise equivalent bandwidth will be explained for the specific case of a single pole RC filter applied to white noise, as in Fig. 5.16. The output of the SFD can be conceptualized as such a circuit where the drive resistance R is $1/g_m$ of the MOSFET, and the load capacitance C is formed by strays on the multiplexer bus.

Noise equivalent bandwidth is a tool utilized to simplify circuit noise analysis, so the integral of noise sources and their transfer function need not be

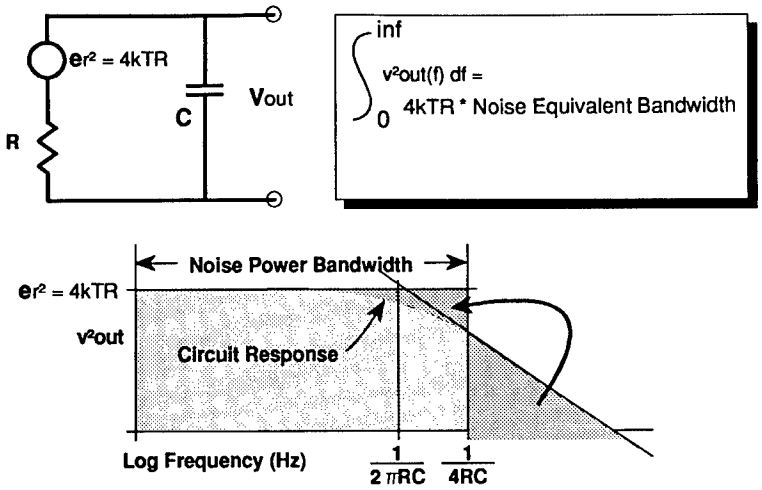


Fig. 5.16 Noise equivalent or power bandwidth is a brick wall approximation utilized in the estimation of total rms noise of a circuit.

evaluated to an infinite frequency. The white thermal noise voltage of the resistor circuit in Fig. 5.16 can be written

$$e_n^2 = 4kTR\int_0^{\alpha} A_v(f)\,df = 4kTR\int_{f_{\text{low}}}^{f_{\text{high}}} df \quad [\text{V}^2] \ , \tag{5.50}$$

where f_{high} and f_{low}, the equivalent noise bandwidth, are selected to provide the equivalence of the two integrals. In this expression, $A_v(f)$ is the transfer function of the single pole circuit comprised of a resistor and a capacitor. The integral of Eq. (5.50), over frequencies from zero to infinity through the filter, yields

$$e_n^2 = 4kTR\Delta f = \frac{4kTR}{4RC} = 4kTR\frac{\pi}{2}f(-3\text{ dB}) \quad [\text{V}^2] \ , \tag{5.51}$$

where $1/(4RC)$ is the equivalent noise or power bandwidth Δf. The term f_{high} of Eq. (5.50) is equal to the noise bandwidth if f_{low} is assumed to be at or near 0 Hz. The Δf of white noise acted on by a single pole filter is the product of $(\pi/2)$ and the half power response, or pole $f(-3\text{ dB})$:

$$e_n\ (\text{rms}) = e_n\ (\text{white})\left[\frac{\pi}{2}f(-3\text{ dB})\right]^{1/2} \quad [\text{V}^2] \ . \tag{5.52}$$

The equivalent noise bandwidth of a source can be dominated by a downstream single pole filter, such as the output impedance and load capacitance of the video amplifier, in which case, that pole would substitute into Eq. (5.52). If a system is ac coupled, and is consequently rejecting low-frequency noise and signal, Δf for a white noise source must be replaced by

$$\Delta f = f_{\text{high}} - f_{\text{low}} \quad [\text{Hz}] \ , \tag{5.53}$$

where f_{low} and f_{high} are the equivalent power bandwidth roll-on and roll-off frequencies, respectively. The result of Eq. (5.53) will be less accurate as f_{low} approaches f_{high}.

Also shown in Fig. 5.16 is the noise power spectral density as a function of frequency and the "brick wall" filter equivalent, or equivalent noise bandwidth, for the circuit. It should be noted again that Fig. 5.16 shows the derivation of power bandwidth for the common case of an RC circuit to a white-noise source. The equivalent noise bandwidth of noise sources other than white, such as $1/f$, applied to transfer functions different from the RC example can also be calculated.

In Eq. (5.53), f_{low} is a function of the data processing method and the implementation of ac coupling (if, in fact, ac coupling is implemented). For example, frame-to-frame differencing reduces or eliminates stationary or slowly changing elements of a scene. In data so processed, only high-frequency noise and scene changes are observed. Low-frequency noise and $1/f$ noise are sharply attenuated, and f_{low} in this case increases toward the Nyquist frequency. Systems that incorporate periodic scene recalibration will have a roll-on f_{low} on the order of half the recalibration frequency; all systems have a lower limit

on the bandwidth. Integration of $1/f$ noise from zero frequency has severe mathematical consequences, and it is important to determine the roll-on frequency of a focal plane system (or processing system). However, in most cases, the roll-on for white noise can be assumed to be zero with insignificant error in the analysis.

SFD Noise Equivalent Bandwidth. For the SFD, the bandwidth of the source follower during multiplexer read is

$$f(-3\text{ dB}) = \frac{g_m}{2\pi C_{\text{mux}}} \quad [\text{Hz}] \ , \tag{5.54}$$

where C_{mux} is the bus capacitance of the multiplexer plus other strays. The noise bandwidth of this single pole circuit is $\pi/2$ times the bandwidth of Eq. (5.54):

$$\Delta f = \frac{g_m}{4C_{\text{mux}}} \quad [\text{Hz}] \ . \tag{5.55}$$

Substituting Eq. (5.55) into (5.10) yields a total input referred channel noise contribution of

$$e_n = \left(\frac{2}{3}\frac{kT}{C_{\text{mux}}}\right)^{1/2} \quad [\text{V}] \ , \tag{5.56}$$

where g_m drops out, leaving C_{mux}, to first order, as the only noise-limiting parameter under the control of the designer.

Increasing the g_m of the source follower unfortunately does not change the thermal noise of the circuit, as expected in Eq. (5.56). If C_{mux} is too large, however, g_m must be increased to sustain a bandwidth large enough to pass the detector signal through the multiplexer. The SFD in this configuration has a MOSFET thermal noise contribution with a form similar to kTC noise in Eq. (5.39).

5.6.5 Capacitor Feedback Transimpedance Amplifier

This class of amplifier addresses a broad range of detector interface and performance requirements across many applications. The CTIA provides a highly stable detector bias, high photon current injection efficiency, high gain, and low noise, and has overall performance equal to and often better than most RTIA configurations. Further, the CTIA is readily integrated into silicon and has the high-frequency response and high modulation transfer function (MTF) advantages of most other reset integrators.

The CTIA, or reset Miller integrator, is shown in Fig. 5.17. Photon charge causes a slight change in voltage on the inverting input node of a differential amplifier. The amplifier, with open-loop gain in the hundreds to tens of thousands, responds with a sharp reduction in output voltage. This change in output is coupled back to the input node through the feedback capacitor, causing photon-induced charge to flow onto the feedback capacitor and oppose the initial

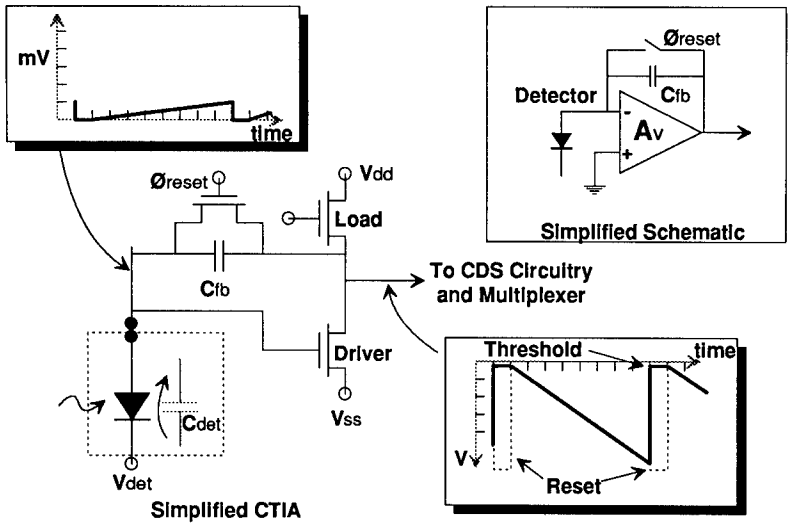


Fig. 5.17 The CTIA utilizes a high-gain inverting amplifier coupled via a feedback capacitor to achieve a high-gain linear dynamic range.

effects of charge on the input node. As detector current accumulates over the frame time, this results in a ramp at the output. At the end of integration, the output voltage is sampled and multiplexed to the output video drivers, and the switch across the feedback capacitor is cycled closed to achieve reset. Signal processing such as correlated double sampling (CDS) is often employed on the output of the CTIA, within the unit cell, primarily to reduce drift, but also to rereference the output signal to a more convenient voltage range. CDS is discussed in a later section of this chapter.

The CTIA gain, assuming a large open-loop amplifier gain, is

$$V_{\text{out}} = \frac{I_{ph} t_{\text{int}}}{C_{fb}} \quad [\text{V}] \ , \tag{5.57}$$

where C_{fb} is the feedback capacitor. This can be compared to gain of the RTIA, in which transimpedance is equal to the feedback resistor, while the CTIA transimpedance is often expressed over a single frame of data as

$$Z_t = \frac{V_{\text{out}}}{I_{ph}} = \frac{t_{\text{int}}}{C_{fb}} \quad [\text{V/A}] \ . \tag{5.58}$$

Given a feedback capacitor of 50 fF (5×10^{15} F) and a 100-Hz frame rate, this translates into an equivalent transimpedance of about 2×10^{11} Ω. This can be done in the CTIA without the careful resistor selection required by the RTIA, because transimpedance is achieved through the switched capacitor network in the amplifier feedback loop instead of a resistor.

The CTIA signal action can be modeled using the equivalent Miller capacitance on the input node, as shown in Fig. 5.18. The feedback capacitor, coupled with the high open-loop gain of the amplifier, results in an equivalent capac-

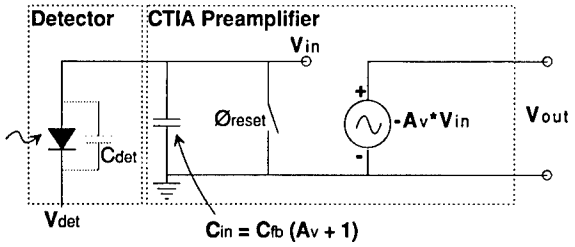


Fig. 5.18 The CTIA acts as a very large integration capacitor on the input node followed by a high-gain voltage amplifier.

itance on the input node that is orders of magnitude larger than the feedback capacitor itself. The excursion in detector bias, ΔV_{in}, can be determined from this model as

$$\Delta V_{\text{in}} = \frac{I_{\text{det}} T_{\text{int}}}{C_{fb}(1 + A_v)} \quad [\text{V}] \ , \tag{5.59}$$

thus providing a stable bias, and therefore a highly linear amplifier, for most reasonable open-loop gains A_v. The output voltage is

$$V_{\text{out}} = \frac{A_v I_{\text{det}} t_{\text{int}}}{C_{fb}(1 + A_v)} \quad [\text{V}] \ , \tag{5.60}$$

which reduces to Eq. (5.58) for the normal case of $A_v \gg 1$.

Like all reset integrators, the detector noise sources are integrated and follow the same sinc function described in Eq. (5.47) [after making appropriate substitution for Z_t, as described by Eq. (5.58)]. The CTIA does have major noise contributions from the input transistor(s) of the differential amplifier that would tend to cause undesirable drift and offset. Most CTIA implementations utilize on-chip CDS within the unit cell to reduce low-frequency noise ($1/f$) and drift, and eliminate the kTC noise caused by reset of the feedback capacitor. Also, a sample/hold circuit is often used to store the end of integration signal value from the CDS circuit. This stored value is then sequentially multiplexed to the video driver during integration of the next frame of data.

The output noise of the CTIA from the input MOSFETs is a function of the CDS circuit (if employed), the noise gain, and the noise bandwidth Δf. The voltage gain of the MOSFET white and $1/f$ noise sources is a function of the detector impedance, as shown in Fig. 5.19:

$$\frac{v_{\text{out}}}{e_n} \approx \frac{C_{fb} + C_{\text{det}}}{C_{fb}} \quad [\text{V/V}] \ , \tag{5.61}$$

which assumes detector node capacitance dominates detector impedance over the noise bandwidth of interest. The equivalent noise bandwidth Δf is a function of circuit implementation and generally determined through small-signal analysis of the specific implementation. Noise contributions from other MOSFETs

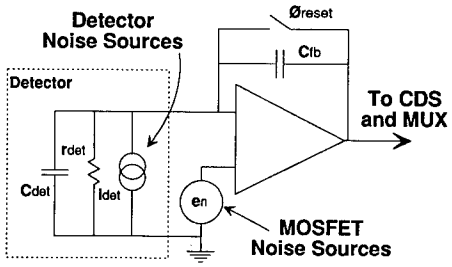


Fig. 5.19 The CTIA noise model must consider gain caused by detector capacitance and resistance. Detector capacitance must be minimized.

within the CTIA are normally negligible to the input MOSFET; however, they can be input referred as a noise voltage e_n.

5.6.6 Injection Circuits

A form of reset integrator, injection circuits perform integration through the channel of an active transistor or CCD channel. Usually the detector node is not directly reset; rather, charge accumulated on the integration capacitor (or CCD bucket), on the output of the injection transistor, is periodically reset or transferred. Since the detector is not directly reset, residual photon-induced charge from one frame time can be integrated into the next frame time causing frame-to-frame crosstalk. Figure 5.20 shows two of the more popular injection circuit configurations.

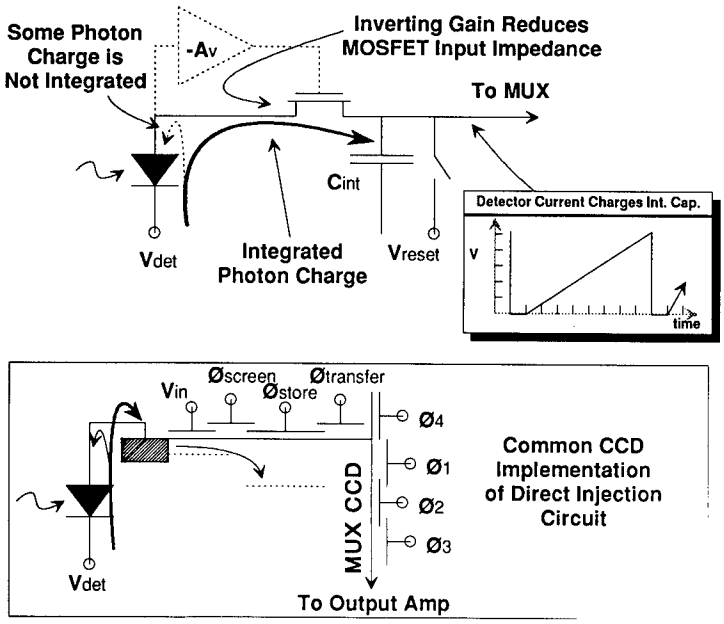


Fig. 5.20 The direct injection circuit is used in both CCD and conventional multiplexer designs. The FEDI is a DI with the addition of an inventor.

Direct Injection. The DI circuit of Fig. 5.20 has been used as an input to CCDs for many years. This readout configuration requires minimal area in the unit cell for implementation; it is perhaps second only to the SI in compactness. Its applications include medium to high proton irradiance conditions that provide sufficient photon current to maintain high MOSFET transimpedance g_m, therefore providing a low impedance and stable bias for the detector.

Photon current in DI circuits is injected, via the source of the input transistor, onto an integration capacitor that has been reset prior to the beginning of the frame. As the photon current integrates, it charges the capacitor throughout the frame; a multiplexer then reads out the final value and the capacitor is reset. In lieu of a multiplexer, a CCD can be utilized to pump the charge out of the unit cell. Unlike the SFD and SI, the gain of the DI, and the CTIA discussed earlier, is set by the integration capacitor, which can be quite small and is not dependent on the detector capacitance. The integration capacitor can be buffered by a source follower to provide voltage mode output. In the case of CCD implementation, the accumulated photon charge is clocked down the CCD to a sense gate, or floating diffusion, whose capacitance sets the gain.

Drain, or channel currents, in the DI are typically very low because they result from photon interaction with the detector. In many injection applications, the MOSFET operates in weak inversion, also known as the subthreshold region, in which drain current is exponentially related to the gate-to-source voltage, as described earlier.

To reduce detector noise, it is important that a uniform, near-zero-voltage bias be maintained across all the detectors. This is especially true for long-wavelength detectors, which can have significant dark current in reverse bias. Detector bias is not directly set on the DI circuits; instead, it is set through the node on the common side of the detector, and includes the MOSFET gate voltage plus the individual threshold voltage of each MOSFET. MOSFET threshold variations across a focal plane can result in bias variations (nonuniformities) of 10 to 50 mV or more between detector elements. Because of this broad range of detector biases, the detector common must be adjusted such that all detectors are biased away from the higher noise forward-bias region of the detector; however, this will cause some elements to receive tens of millivolts reverse bias and can result in higher detector $1/f$ noise and fixed pattern noise in the form of nonuniform detector dark currents.

The input to the readout should also provide a low impedance to the detector. This provides a stable detector bias and a high photon current injection efficiency. If the input impedance of the DI is too high, a fraction of the photon current will be shunted across the detector and not injected into the MOSFET, resulting in a loss of SNR. This injection efficiency, or ratio of integrated photon current I_{int} to photon current, is

$$\text{Injection efficiency} = \frac{I_{\text{int}}}{I_{ph}} = \frac{r_{\text{det}} g_m}{1 + r_{\text{det}} g_m} \quad [\text{A/A}] \ , \tag{5.62}$$

where r_{det} is the resistance of the detector and g_m is the inverse of the input impedance looking into the source of the MOSFET.

The term g_m is a strong function of drain current for MOSFETs in the weak inversion (or subthreshold) region, where most DI circuits operate. The rela-

tionship between drain current and gate-to-source voltage for a MOSFET in weak inversion was given by Swansen and Meindl[11] and can be written

$$I_d = \frac{W}{L}\mu_n K_1 \exp\left(V_g - \frac{mkT}{q}\right) \quad [\text{A}] \tag{5.63}$$

for an n-channel MOSFET, where K_1 includes geometry and operational characteristics of the MOSFET. For a transistor in subthreshold, the g_m is given by the first derivative of Eq. (5.63) with respect to drain current and gate-to-source voltage:

$$g_m = \frac{\delta I_d}{\delta V_{gs}} = \frac{I_d q}{mkT} \quad [\text{mhos}] \ . \tag{5.64}$$

Note that g_m in subthreshold is independent of the device geometry parameters. Because high g_m is a direct function of drain current (i.e., photon current), there are limitations in the background flux capability of the DI; at very low photon currents, g_m becomes very poor and the injection efficiency falls.

Low g_m not only affects the injection efficiency of the DI but also reduces the bandwidth of the detector element, thereby impacting frame-to-frame crosstalk and MTF. If g_m is the dominant impedance on the detector node, a condition required for high injection efficiency, the frequency response of the detector node becomes

$$f(-3\text{ dB}) = \frac{g_m}{\pi(C_{\text{det}} + C_{gs})} \quad [\text{Hz}] \ , \tag{5.65}$$

where C_{gs} is the MOSFET gate-to-source capacitance. Since g_m is directly proportional to the photon current, DI use is limited in high-speed low-photon-flux applications.

Noise in the DI includes the same detector, MOSFET, and kTC noise sources discussed earlier for other preamplifiers. MOSFET noise e_n can be output referred as a current i_o to the integration capacitor. Since i_o is integrated, it is acted on by the sinc function of the reset integrator discussed earlier. The transfer function of the input MOSFET noise sources to output current is

$$i_o \approx \frac{e_n g_m}{1 + r_{\text{det}} g_m} = \text{injection efficiency} \frac{e_n}{r_{\text{det}}} \quad [\text{A}/\sqrt{\text{Hz}}] \ . \tag{5.66}$$

Note that MOSFET noise is negligible for cases where r_{det} is high. The transfer function for all detector noise sources (thermal, $1/f$, and photon-induced) is the sinc function discussed for the SI; however, the DI detector noise is acted on by the charge injection efficiency and the input node bandwidth discussed above.

Feedback-Enhanced Direct Injection. Feedback-enhanced direct injection circuits (FEDI) address some shortfalls of the DI by introducing a feed-forward inverting amplifier off the detector node, as shown in Fig. 5.20. The inverting

amplifier, with gain of A_v, reduces the input impedance of the DI and therefore increases the injection efficiency and bandwidth. The effect is to reduce the minimum operating photon flux range of the FEDI by approximately an order of magnitude below that of the DI. All previously given equations for the DI can be utilized to estimate performance for the FEDI, with the substitutions of g'_m and C'_{gs} for g_m and C_{gs}, where

$$g'_m = g_m(1 + A_v) \quad [\text{A/V}] \ , \tag{5.67}$$

$$C'_{gs} = C_{gs}(1 + A_v) \quad [\text{F}] \ . \tag{5.68}$$

The detector bias uniformity across the array is dependent on the threshold variation of the inverting amplifier for the FEDI. Threshold nonuniformities from the inverting amplifier can be addressed through the use of auto-zeroing circuits, which utilize feedback during the zeroing period to reset the amplifier input. These zeroing circuits also reduce drift components of the amplifier.

The output referred MOSFET preamplifier noise for the FEDI is

$$i_o \sim \frac{e'_n g'_m A_v}{(1 + A_v)(1 + r_{\text{det}} g'_m)} \approx \text{injection efficiency} \frac{e'_n}{r_{\text{det}}} \quad [\text{A}/\sqrt{\text{Hz}}] \ , \tag{5.69}$$

where the dominant preamplifier noise is the input referred noise of the inverting amplifier e'_n. The noise contribution of the injection transistor of the DI, e_n, is negligible, because it is reduced by approximately A_v. Detector noise sources are treated in the same manner as they are for the DI.

5.6.7 Gate Modulation Circuits

Gate modulation circuits utilize photon current to modulate the gate voltage and thereby induce an output current in the MOSFET. The transistor output drain current accumulates onto an integration capacitor or a CCD bucket. Because of this gate modulation action, the detector bias is required to change, or debias, as scene-induced current changes. Thus, detectors utilized in gate modulation circuits must be operated in reverse bias where $1/f$ noise and dark current nonuniformity can be a problem. A reset switch may be incorporated onto the detector node to reduce crosstalk from one frame to the next, resulting in a higher MTF—in this special case of gate modulation circuits—than in their injection circuit counterparts. The gate modulation circuits discussed in this section are most commonly employed in high to very high background applications.

Resistor Gate Modulation. The resistor load gate modulation circuit (RL) is shown in Fig. 5.21. Detector bias is a function of detector (photon) current, load resistance R_l, and resistor bias V_d:

$$V_{\text{in}} = (I_{ph} + I_{\text{dark}})R_l + V_d \quad [\text{V}] \ . \tag{5.70}$$

The load resistor is designed for low $1/f$ noise,[12] high temperature stability, and for uniformity from cell to cell. Polysilicon resistors in integrated circuit form are commonly utilized as the load. The MOSFETs in gate modulation

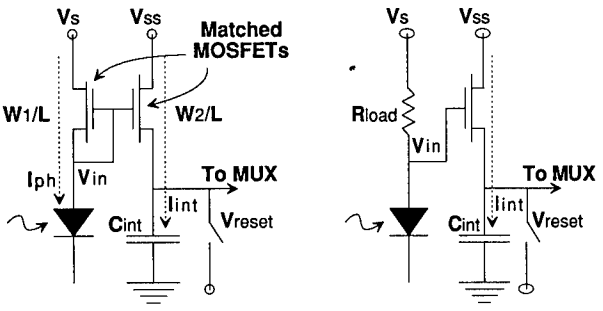


Fig. 5.21 The gate modulation circuit biases the detector through a high-impedance resistor or current mirror MOSFET: (a) current mirror input and (b) resistor load input.

circuits are normally run in weak inversion. Therefore, the resulting integrated drain current can be determined by applying Eq. (5.70) to the weak inversion (subthreshold) MOSFET of Eq. (5.63):

$$I_{\text{int}} = I_d = K_2 \exp(V_{\text{in}}) \quad [\text{A}] \ , \tag{5.71}$$

where K_2 is dependent on resistor bias, MOSFET threshold voltage, temperature, and other factors.

In high background applications, where the signal of interest is small compared to optics and scene background irradiance, the resistor gate modulation circuit provides a design that can reject much of these background components. When the background alone is present on the detector, the bias on the detector or the load resistor can be adjusted for negligible drain current or integration of charge. As signal is applied, the MOSFET drain current increases exponentially with photon current and thereby allows some level of background flux rejection.

The output uniformity from one element to another is dependent on variations in MOSFET threshold voltage, load resistance, and detector resistance. Because of these nonuniformities, a substantial amount of fixed pattern noise is characteristically presented at the output of the SCA.

Current Mirror Gate Modulation. In the current mirror gate modulation amplifier (CM), the resistor of the RL circuit is replaced with a MOSFET, as shown in Fig. 5.21. In CM operation, photon current flowing into the drain of the first of two closely matched transistors induces a common gate-to-source voltage change in both transistors; this results in a similar, or mirrored, current in the second transistor. If the source voltages, V_s and V_{ss}, of two matched MOSFETs are connected, both will have the same gate-to-source voltages, which will induce a current in the output MOSFET identical to the detector current flowing through the input MOSFET. In this specific case, the integration current is a linear function of the detector current (unlike the nonlinear case of the RL). This CM action can also be used to scale up or down the integration current in the second transistor by varying the transistor geometries according to

$$I_{\text{int}} = I_{\text{det}} \frac{W_2/L_2}{W_1/L_1} \quad [\text{A}] \ , \tag{5.72}$$

where W and L are the width and length dimensions of the two MOSFETs. The CM is frequently used in applications of very high background, where there is insufficient area in the unit cell to accumulate the detector current over an entire frame period; in this case, the width of the second transistor can be scaled down to a fraction of that of the load transistor, thereby scaling integrated current to a fraction of detector current. The current gain can also be set through the bias voltages V_s and V_{ss}:

$$I_{\text{int}} = I_{\text{det}} \frac{W_2/L_2}{W_1/L_1} \exp(V_s - V_{ss}) \quad [\text{A}] \ . \tag{5.73}$$

This analysis has assumed 100% injection efficiency of detector current into the input load MOSFET. Injection efficiency is a function of detector resistance and MOSFET g_m, which is dependent on detector current (photon current), and is calculated using the same equations as for the DI discussed earlier.

Nonuniformity from element to element can be quite large in CM designs. This is because the threshold voltages of the current mirror MOSFET pair are not identical and also because injection efficiency can vary due to detector resistance variations, factors that result in the high fixed pattern noise commonly observed in these designs. CM advantages over the RL include greater linearity and absence of a load resistor. The frequency response is calculated using the DI circuit equations discussed in the preceding subsection, unless the detector is reset each frame, in which case the frequency response follows the sinc function of reset integrators.

5.7 SIGNAL PROCESSING

Incorporation of signal processing on an ROIC is often desirable in order to reduce off-focal-plane electronics, reduce the data rate, or perform processing prior to sampling and multiplexing. The two areas where on-chip signal processing can occur are (1) within the unit cell itself and (2) in the multiplexer prior to the output video amplifier. The most common forms of ROIC signal processing are band-limiting (provided by all reset integrator preamplifiers), sample and hold, correlated double sampling, and time-delay integration. Other less common examples, not covered here, include gain and offset correction, signal digitization,[13] single event gamma circumvention (correction of corrupt signal), and frame-to-frame differencing to remove clutter and background offsets.

5.7.1 Sample and Hold

Simultaneous integration of all elements of the sensor is often required. The signal for simultaneous integration is accumulated over a given time period in a snapshot mode. To reset the detector and preamplifier to begin integration of the next frame, the signal from the previous frame must be sampled and stored temporally for sequential readout by the multiplexer. The most common form of this type of sample and hold is composed of a MOSFET sampling switch,

a hold capacitor, and a unity gain buffer amplifier. Figure 5.22 shows such an implementation on the output of a SFD preamplifier. A simple output MOSFET source follower and load serves as buffer amplifer prior to multiplexing. The sample and hold circuit resides in the unit cell, and therefore puts limitations on minimum cell size because ample area is required for the three transistors and capacitor.

The key design objectives of the sample and hold circuit are settling time and noise. Settling time is governed by the time required to charge or discharge the hold capacitor; it is also a function of the drive capability of the preamplifier or buffer circuit in front of the sampling switch. The output buffer must be able to drive or slew the multiplexer bus capacitance; the output current required for this is calculated in the same way as for the output source-follower video drivers, covered in a later section of this chapter.

Noise of the sample and hold circuit is governed by the kTC noise of the hold capacitor v_c, which must be significantly less than the output referred noise of the preamplifier, v_{out} (preamp). A kTC noise equal to $1/5v_{\text{out}}$ (preamp) will result in less than a 2% increase in noise:

$$v_c = \frac{kT}{C_{sh}} \leq \frac{v_{\text{out}}\ (\text{preamp})}{5} \quad [\text{volts rms}] \tag{5.74}$$

for a minimum sample hold capacitance C_{sh} of

$$C_{sh} \geq \frac{25kT}{v_{\text{out}}^2\ (\text{preamp})} \quad [\text{F}]\ . \tag{5.75}$$

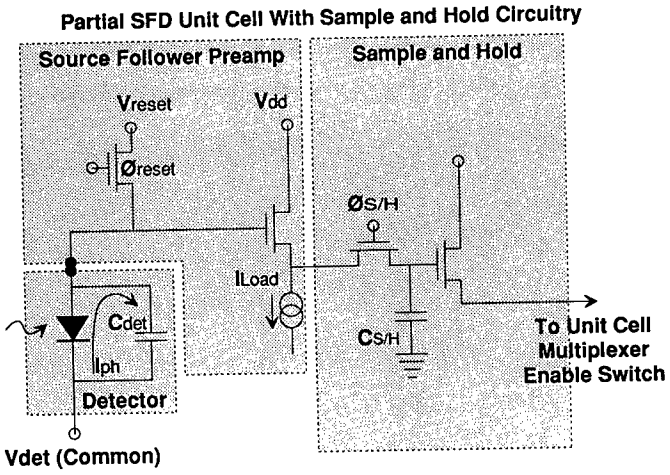


Fig. 5.22 The sample and hold function provides frame store in the unit cell. In the voltage domain configuration shown, the SFD is followed by a MOSFET switch and hold capacitor. The output is buffered with a MOSFET source follower.

5.7.2 Correlated Double Sampling

Drift and $1/f$ noise are often dominant noise contributors in readout preamplifiers. Therefore, it is often desirable to recalibrate, or rezero, the amplifier chain periodically in order to achieve lower noise and greater absolute accuracy. This is normally accomplished in reset integrators by rezeroing the output of the preamplifier at the beginning of integration.

Figure 5.23 shows a preamplifier, in this example the SFD, with a clamp circuit on the output. The output signal is initially sampled across the clamp capacitor during the onset of photon integration (after the detector is reset). The action of the clamp switch and capacitor subtracts any initial voltage from the output waveform. Because the initial sample is made before significant photon charge integrates onto the capacitor, the final integrated photon signal swing is unaltered; however, any offset voltage or drift present at the beginning of integration is removed, or subtracted, from the final value. This process of sampling each pixel twice, once at the beginning of the frame and again at the end, and providing the difference is called correlated double sampling (CDS). This process can be performed within the unit cell as shown in the figure, or it can be performed numerically off the focal plane in a digital processor.

CDS reduces or eliminates low-frequency noise but at the expense of increased noise at higher frequencies, as shown in Fig. 5.24. The value of the initial CDS sample represents dc offsets, low-frequency drift, and high-frequency noise; this initial value is subtracted from the final value, which also includes dc offset, low-frequency drift, and high-frequency noise. Since the two samples occur within a short period of time, the dc and drift components of each sample do not change significantly; hence, these terms cancel in the subtraction process. On the other hand, high-frequency waveforms (noise) change significantly between the two samples and bear little resemblance or correlation to each

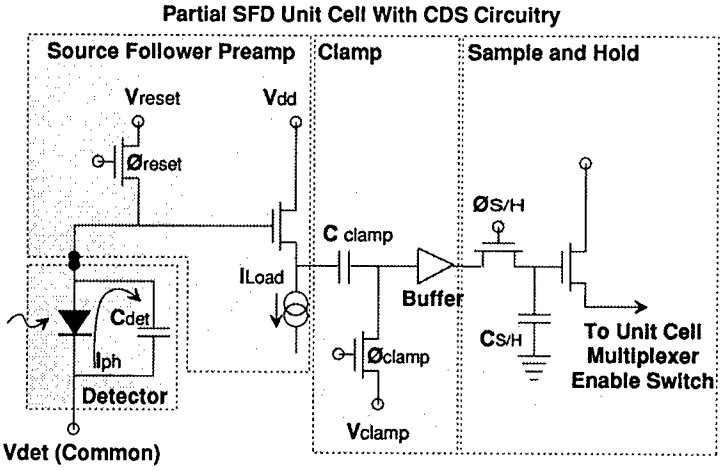


Fig. 5.23 Correlated double sample achieved in the unit cell with the addition of clamp and sample hold circuitry.

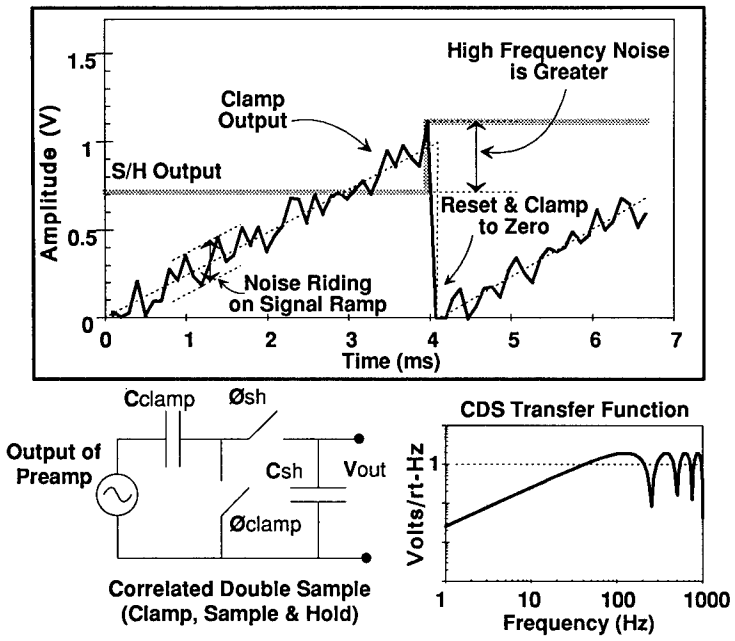


Fig. 5.24 CDS removes dc offsets and drift, but increases high-frequency noise.

other. The net effect is that these two uncorrelated high-frequency components sum to twice the noise power.

The sample transfer function of a noise source acted on by CDS processing can be calculated in the same manner as that for the reset integrator. The only difference is that the presampling input to the CDS, from the amplifier noise sources, is acted on by the band-limited gain of the amplifier, A_v, instead of an integrator,

$$V_{\text{out}}(t) = y(t) = A_v h(t) = A_v[\delta(t) - \delta(t - t_{int})] \text{ [V]} . \tag{5.76}$$

Then, using the Laplace transform,

$$V_{\text{out}}(\omega) = Y(\omega) = A_v H(\omega) = A_v[1 + \exp(-j\omega t_{\text{int}})] \text{ [V]} , \tag{5.77}$$

which results in

$$v_{\text{out}}(\text{f}) = A_v[2 - 2\cos(2\pi f t_{\text{int}})]^{1/2} \text{ [V}/\sqrt{\text{Hz}}\text{]} , \tag{5.78}$$

where it is assumed samples occur at the beginning and end of the frame. The sample function of Eq. (5.78) changes somewhat if significant time lapses between the end of one frame and the beginning of the next frame. The integral of Eq. (5.78) over all frequencies can be approximated for white noise by

$$v_{\text{out}} \text{ (CDS)} = e_n A_v (2\Delta f)^{1/2} \text{ [volts rms]} , \tag{5.79}$$

provided the equivalent noise bandwidth of the noise source, Δf, is many times greater than the frame rate.

Noise in the CDS circuitry is dominated by kTC and buffer noise, as it was for the sample and hold case of the previous section. To minimize clamp capacitor kTC noise, C_{clamp} is scaled in the same manner as the hold capacitor. Additionally, C_{clamp} must be large enough to overcome any strays on its output side in order to maintain near unity gain. Stray capacitance, which includes the MOSFET switch and the gate of the output buffer MOSFET, together with the clamp capacitor, reduces the signal gain of the clamp, A_{clamp}:

$$A_{\text{clamp}} = \frac{C_{\text{clamp}}}{C_{\text{stray}} + C_{\text{clamp}}} \quad [\text{V/V}] \ . \tag{5.80}$$

It is desirable to have a clamp gain of at least 0.95 resulting in

$$C_{\text{clamp}} \geq 20\, C_{\text{stray}} \quad [\text{F}] \ , \tag{5.81}$$

although higher capacitance may be required to reduce kTC noise.

5.7.3 Time-Delay Integration

A simple scanning SCA includes a single row of detector elements that scan a scene and multiplex the resulting signal to the output. To generate an entire scene, the array is scanned from one side of the field of view to the other. Sensitivity is limited in a sensor of this type by the dwell time, or equivalent time that an element is "looking" at a specific point in the scene. To reduce flicker and provide reasonable scene refresh rates, this scene is typically scanned 30 to 100 times per second. The dwell time of a given element will tend to reduce if a short scan time, high scene resolution, or large field of view is required. Placing a second row of detectors next to the first row produces a second image of the scene that is displaced in time. The two images can be added (integrated), after the first is scene delayed, to double the signal level with only a modest increase in noise of the two frames. By adding rows of time-delay elements and performing the time-delay integration (TDI), the SNR of a scanning system can be improved by

$$\text{SNR improvement} = \sqrt{n} \ , \tag{5.82}$$

assuming the system is detector limited, where n is the number of elements or rows in TDI.

The TDI function can be performed off focal plane (normally after digitization) or on the ROIC. On-chip TDI requires temporary frame storage for each TDI row, and a multiplexer to transport the signal for the summing process. TDI can be performed via a large array of storage capacitors, however, it is most commonly implemented utilizing a CCD. The CCD transports the signal charge from one element to the next for accumulation with the next detector in TDI, as in Fig. 5.25. The transport time between cells provides the delay necessary for synchronization to the signal from the next detector. The CCD accumulates all TDI elements before the signal is buffered off the ROIC

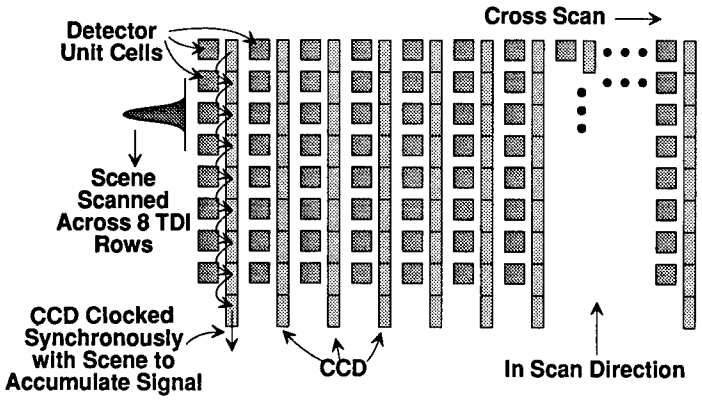


Fig. 5.25 TDI can be performed by utilizing a CCD, which provides delay between TDI detector before summing.

via the video amplifier. The CCD in this example provides both TDI and the multiplexer functions. More details on CCD operation are provided in the following section.

5.8 DATA MULTIPLEXERS

The multiplexer (MUX), in its simplest form, is a series of switches or transfer wells that sequentially transports sampled data from many pixel elements and encodes them onto a common bus. Signals from tens to hundreds of thousands of detector elements can be multiplexed through a video driver to a single output pad on the readout. Two common forms of multiplexers are available to the readout designer: (1) CCDs, which utilize a series of sequentially enabled potential wells, or metal insulator semiconductor (MIS) capacitors, to transfer charge to a floating gate or diffusion; or (2) a set of switches that is enabled sequentially to a common bus as shown in Fig. 5.26. Staring arrays utilize two multiplexers, one for the column and one for the row MUX. The column multiplexer shifts data at low speed from each unit cell to the end of the column; the data are then further multiplexed, at high speed, with other elements of the same row from subsequent columns. The output is thus formatted with pixel one of the first row through the last pixel in the first row followed by the data from pixel one in the second row through the last pixel in the second row, and so on. If multiple outputs are utilized, data can be formatted to address quadrants or interlaced columns of the ROIC unit cells.

5.8.1 CCD Multiplexers

The invention of the CCD in 1970 launched the optical industry, including infrared, into the area of high-density multielement solid-state focal planes. Evolution of most modern hybrid focal planes can be traced back to the introduction of CCD technology. Several detailed sources of information on the subject are available that provide a more in-depth review of CCD physics and

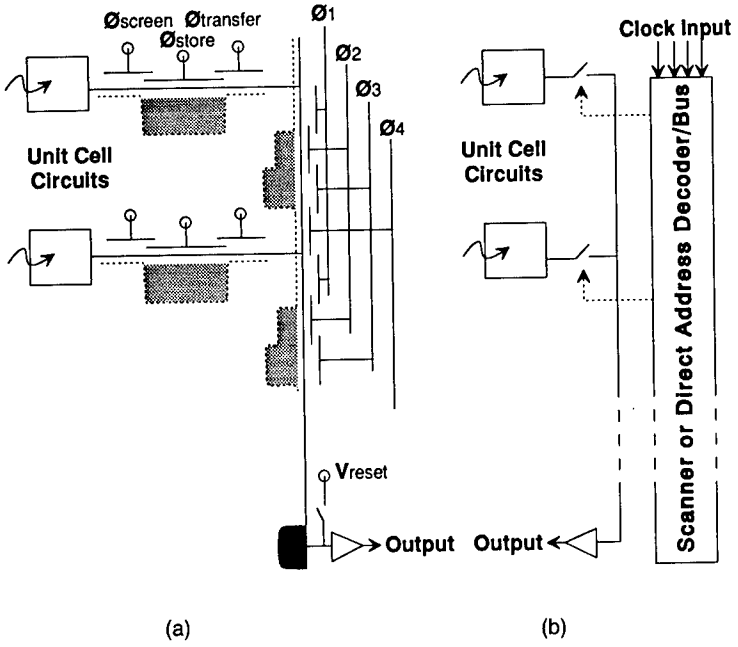


Fig. 5.26 Multiplexers include CCD and MOSFET enable switches: (a) CCD multiplexer and (b) direct readout multiplexer.

their wide range of applications and capabilities.[14–16] The following is a brief description of the CCD in ROIC applications.

The CCD basically provides charge storage and transfer along the surface of a semiconductor. Charge is manipulated through the attraction of carriers to a field set up between a gate and the substrate of the semiconductor. In MOSFETs, the application of a voltage, above a specific threshold, results in a channel formation under the gate between the source and drain diffusions, as discussed in the earlier MOSFET primer. In the absence of source/drain diffusion to attract carriers for channel formation, the area under the gate assumes a potential capable of storing localized minority carrier charge. The storage potential under the gate is referred to as a "bucket," and removal of the gate voltage results in the loss of charge storage capability in a manner analogous to a bucket being emptied. Charge is injected into a CCD through an adjacent diffusion or gate structure.

The action of the CCD is shown in the three-phase CCD multiplexer of Fig. 5.26(a). Charge, in this case minority carrier electrons in the *p*-type semiconductor, tends to flow toward the highest surface potential. In this case the surface potential is provided (and manipulated) by a positive gate bias on $\emptyset_1$. Charge is actually stored in a thin layer near the semiconductor surface, but is commonly depicted schematically as a bucket in the figure. The potential (voltage) on $\emptyset_2$ can be increased to attract charge from under the first gate. The voltage on $\emptyset_1$ is reduced to zero, thus forcing charge near the semiconductor surface to flow under $\emptyset_2$. This action can be repeated with $\emptyset_2$ and $\emptyset_3$ resulting in charge transfer to the right in the figure. Notice that additional charge "packets" in the CCD channel also flow in a similar manner. The CCD must

be capable of providing isolation between these signal charge packets as shown in Fig. 5.26. A two-phase CCD can be constructed as in Fig. 5.27(b). In this case, alternate gates are built on different thicknesses of gate oxide, thereby inducing a potential gradient across the surface under the two connected gates and providing charge transport from one gate to the next.

The CCD can perform both the detector and ROIC functions. Photon current can be accumulated directly into the CCD bucket or in an adjacent diode (detector). This monolithic approach is common in visible sensors, and has been demonstrated in the infrared.[1,17]

The charge is actually stored in a thin layer at the surface of the semiconductor in a MIS capacitor. The effective voltage applied to the MIS capacitor is

$$V_a = V_g - V_{FB} \quad [\text{V}] \ , \tag{5.83}$$

where V_{FB} is the voltage generated by the work function between the metal and semiconductor, and the fixed charge in the insulator.[4] The maximum storage capacity of a MIS capacitor is

$$N_{\max} = V_a \frac{C_{\text{gate}}}{q} = V_a \frac{qA_c\varepsilon_o\varepsilon_c}{T_{ox}q} \quad [\text{electrons}] \ , \tag{5.84}$$

where V_a is limited by the gate/oxide breakdown voltage.

As charge is transferred down a CCD the packet can gain charge through uncontrolled lateral dark currents, can lose charge, or can gain a random charge element (noise) through fluctuations resulting from surface state interactions or dark current noise. Charge lost through incomplete transfer can result in low signal at the output, or can result in crosstalk because the lost charge can appear in subsequent charge packets. Charge transfer efficiency

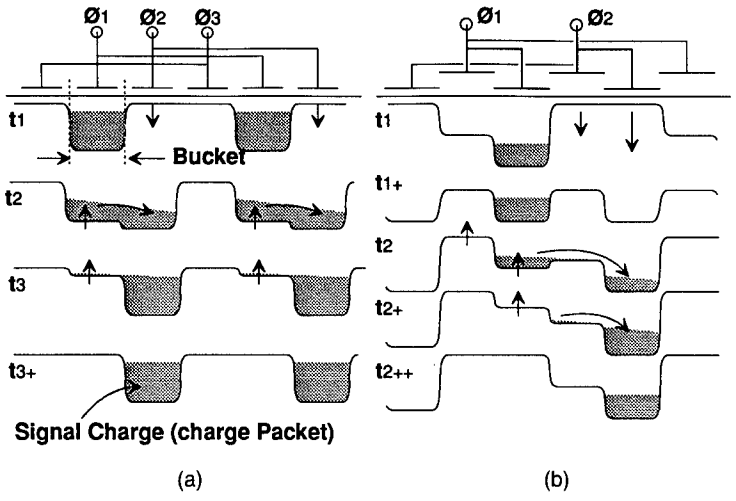


Fig. 5.27 Charge is transferred along a CCD as the surface potential is sequentially changed: (a) three-phase CCD and (b) two-phase CCD.

η_x, the ratio of charge transferred into the next CCD cell to the original charge prior to transfer, is related through

$$\eta_x + \varepsilon_x = 1 \ , \tag{5.85}$$

where ε_x is the transfer loss factor or charge transfer inefficiency; ε_x is related to the transfer time, because slower CCD clock rates allow for greater time for charge to transfer from one bucket to the next. Charge transfer efficiency for a good CCD is greater than 0.9999 per transfer. At low frequencies, <1 MHz, the major contributor to transfer inefficiency is interface-state trapping, whereby charge is trapped and subsequently released. This effect can be minimized through the injection of a constant current (dc) input to the CCD, or a "fat-zero" charge that fills most of these surface states. At high frequencies the charge transfer efficiency is limited by self-induced drift or repulsion of the carriers, by thermal diffusion, and by drift as a result of fringe fields such as those presented by adjacent gates and diffusions in the CCD.

Noise in the CCD manifests itself as fluctuations in the number of carriers in a charge packet. The major noise mechanisms in the CCD are related to those in MOSFETs and detectors:

1. as with $1/f$ noise in MOSFETs, fluctuations in surface state interaction with charge
2. analogous to channel thermal noise in the MOSFET, variation caused by thermally generated carriers in the semiconductor
3. noise generated through the absorption of photons (photon noise).

The input circuit to a CCD includes the charge domain preamplifier circuits discussed earlier: SI, DI, FEDI, RL, or CM. Voltage mode preamplifiers can be converted to charge mode by inserting a capacitor in series with the output. The output charge from the CCD can be transferred to a source-follower gate through a diffusion at the end of the CCD. The source-follower gate must be reset before the next signal is transferred. Additional input and output circuits are utilized for the CCD and are covered in Refs. 1 and 17.

5.8.2 Direct Address and Scanning Multiplexers

Direct address and scanning multiplexers became popular when CMOS technology was introduced into readout designs. Prior to this, NMOS or PMOS integrated circuits were commonly used and MOSFET multiplexer switches and their drivers, although not uncommon, were more difficult to implement. The MOSFET switch-type multiplexers are also simpler to interface to voltage mode preamplifier circuits, such as the CTIA and SFD, than CCDs, which operate in the charge domain.

The direct address and scanning multiplexers are shown in Fig. 5.28. In the direct address multiplexer, a digital code provides a unique address to each unit cell that is to be enabled onto the bus. This type of multiplexer requires several input address lines, but allows for random access to any element on the focal plane.

The scanner multiplexer utilizes a flip-flop-type scanner to pass sequentially an enable signal down a row or column of unit cell multiplexer switches. A simplified implementation, shown in Fig. 5.28, requires one line to reset, $\emptyset_r$,

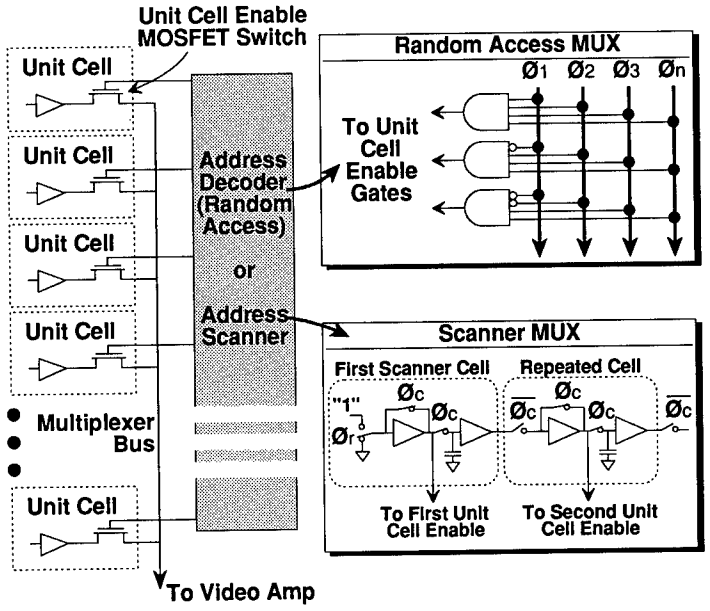


Fig. 5.28 Direct address and scanning multiplexers use MOSFET switches to enable detector-derived signals onto a common bus.

and another, $\emptyset_c$, to clock the scanner; $\emptyset_r$ resets the first scanner cell to a logical "1" at the beginning of the multiplexing sequence, $\emptyset_c$ and its inverse subsequently clock the logical "1" down the scanner, thus providing sequential unit cell enable.

The scanner multiplexer is easily integrated into readouts and can be utilized to generate other timing waveforms such as unit cell reset and sample/hold clocks. The on-focal-plane scanner has reduced the required number of input/output lines on large focal planes to fewer than 10.

The multiplexer itself does not produce noise, because the enable transistors are hard "on" or completely "off"; however, the kTC of the bus and noise of any buffer amplifiers must be considered in the design. Scanning multiplexers can achieve speeds upward of 10 million signals per second.

5.9 OUTPUT VIDEO AMPLIFIERS

The output video driver buffers the encoded signal string from the ROIC multiplexer off the focal plane, through the cryogenic and ambient interface cables, and finally to a set of warm electronics where any necessary signal conditioning is provided prior to digitization or display. The primary concern of the video driver is the cryogenic power dissipation required to provide adequate frequency response and dynamic range.

The limits of video driver power can be determined by evaluating the extremes in driver circuit configurations and output load capacitance. At the lower limit, power can be calculated from the energy required to charge and discharge the load capacitor presented by the cable of Fig. 5.29:

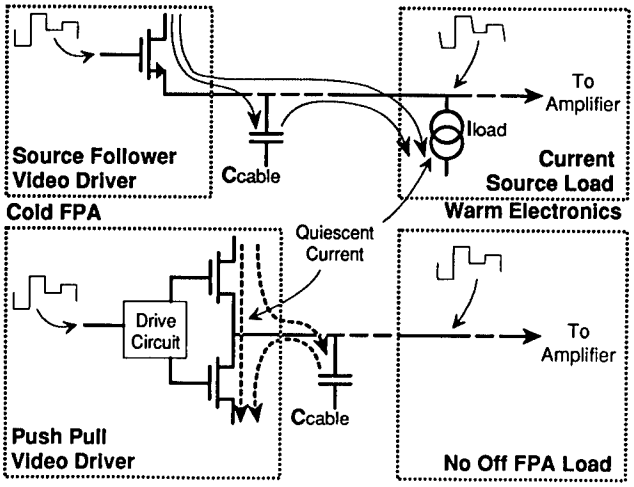


Fig. 5.29 Key drivers in video driver power are data rate, drive capacitance, and driver design. The lower power push-pull has greater ROIC complexity.

$$E(C_{\text{load}}) - \frac{1}{2}C_{\text{load}}\Delta V_{\text{max}}^2 \quad [\text{J}] \ . \tag{5.86}$$

The total power is the product of the energy and the repetition rate, plus the energy required to provide the idle current, I_{idle}, in the push-pull circuit:

$$P_{\text{out}}\ (\text{min}) > \frac{0.5\ C_{\text{load}}\Delta V_{\text{max}}^2 N}{T_{\text{frame}}} + (\Delta V_{\text{max}} + 1.5)I_{\text{idle}} \quad [\text{W}] \ , \tag{5.87}$$

where C_{load} is the total output capacitance, ΔV_{max} is the maximum output voltage swing, N is the number of elements, and T_{frame} is the time between two frames of data. The idle current required for the push-pull is in the 10- to 100-μA range.

A video driver circuit configuration that can approach this minimum power is the push-pull concept in Fig. 5.29. Push-pull circuits require several components to properly bias and drive the output transistors, as well as a quiescent current to keep them both active and linear. Although these push-pull circuits are somewhat complex, the near unity gain push-pull of Fig. 5.29 has advantages in speed and power over most other common video driver configurations.

The most common type of video driver for cooled focal planes is the source follower (in MOSFET implementations) of Fig. 5.29. The primary advantages of this circuit are simplicity, relative low power, and near unit gain; also, this configuration implements a current source load off the focal plane to minimize focal plane cryogenic power. The source follower responds to a positive change in input voltage by charging the stray load capacitor through the low-impedance source of the active N-channel MOSFET. Since the MOSFET in this configuration can only source current, or charge the load capacitance, an off-focal-plane current load is needed to discharge the capacitor for negative transi-

tioning signals. The current load required to discharge the capacitor is dependent on the time allocated for slewing the output T_{slew}, through the maximum output voltage swing $\Delta V_{\max}$:

$$I_{\text{load}} = \frac{\Delta V_{\max} C_{\text{load}}}{T_{\text{slew}}} \quad [\text{A}] \ . \tag{5.88}$$

To provide a reasonable time for downstream electronics to settle, T_{slew} is typically no more than one-third the total data valid time. The data valid time T_{data} for a given output is approximately

$$T_{\text{data}} < \frac{T_{\text{frame}}}{N} \quad [\text{s}] \ . \tag{5.89}$$

This gives

$$I_{\text{load}} > \Delta V_{\max} C_{\text{load}} \frac{3N}{T_{\text{frame}}} \quad [\text{A}] \ . \tag{5.90}$$

To keep the MOSFET operating in a linear fashion, the maximum voltage across the output transistor must be approximately 1 V greater than the maximum output voltage swing. The worst case video driver power occurs when the maximum voltage is across the video driver transistor:

$$P_{\text{driver}} (\max) = (\Delta V_{\max} + 1) \Delta V_{\max} C_{\text{load}} \frac{3N}{T_{\text{frame}}} \quad [\text{W}] \tag{5.91}$$

and for large voltage swings

$$P_{\text{driver}} (\max) \approx 3 C_{\text{load}} \Delta V_{\max}^2 \frac{N}{T_{\text{frame}}} \quad [\text{W}] \tag{5.92}$$

or up to six times the minimum power of the push-pull. Knowledge of the typical infrared signal can result in a lower average power of the source-follower driver; the circuit should be designed so the MOSFET has lower voltage across it in the nominal signal output case, and has higher voltage across it only when there are infrequent large differences between the nominal output voltage and the maximum deviation from the nominal output. In Fig. 5.29, the nominal might be 1.5 V below the power supply V_{dd}, while the maximum deviation may occur at $V_{dd} - 6$ V.

5.10 POWER DISSIPATION

Power dissipation on any cryogenic focal plane, although very small when compared to conventional room temperature analog electronics, is a major system driver because it often translates into greater system weight, increased power in active coolers, larger cooler volume, and limitations in mission life. For example, the mission life of a space sensor can be driven by the consumption

rate of a cryogenic liquid or solid; or, where active cooling is utilized, cryogenic power can force the use of larger coolers, which in turn consume more electrical power and add greater weight. Passive cooler (radiative cooler) size also increases considerably as focal plane increases, and it is ultimately limited by space viewing issues as well as size and weight constraints. In view of these factors, it is not only important to reduce readout power, but to design sensors for the highest operating temperature that the selected detector technology can support.

Cold focal plane power often includes electrical power dissipation (I^2R) in the readout, detector, and cable; radiative power from the environment to the focal plane and interconnect cable; and power conducted through the interconnect cable to the focal plane. The readout design influences both I^2R losses in the readout and the thermal load of the cable, because this load is a function of the number of cable traces required by the readout. Of secondary importance are on-chip bias and clock generating circuits, such as the scanner multiplexer, which can be designed with minimal thermal impact for high-density sensor arrays. High-impedance detectors, such as photovoltaics, have negligible impact on focal plane power given that detector current is in the fempto-amp to nano-amp range.

Power dissipation occurs in two primary areas of the readout: the preamplifier section, with associated signal processing, and the video output driver. In some cases, where the preamplifier output is in the charge rather than the voltage domain, the multiplexer transimpedance amplifier must be considered. The drivers of power dissipation are (1) the data rate, which is set by the system specification, and (2) the linearity and sensitivity requirements of the readout, which drive the circuit design. The data rate, a combination of the detector element frame rate and the total number of elements, drives the slew rate and frequency response of both preamplifier and output video driver.

Total readout power, including the preamplifier, basic signal processing, and video driver, is a strong function of the detector unit cell preamplifier design. Unit cell power can be estimated, to first order, based on the number of active transistors within the unit cell. Estimates for typical power per unit cell were summarized earlier in Table 5.3.

Preamplifiers, such as the self-integrator, direct injection, current mirror, and resistor load types, have negligible current in the unit cell, that is, on the order of the detector current. Power for these circuits is in the nanowatt range and is also negligible. These circuits are normally current mode output and, as such, require a transimpedance amplifier on the output of the multiplexer or a CCD for charge transfer. In the case of CCD, power is very small and dominated entirely by the output video amplifier and on-chip clock and bias circuits. Transimpedance amplifiers, on the other hand, require power in the milliwatt range. The transimpedance amplifier is shared between the unit cells on the multiplexer bus and dissipates 1 to 10 nW per unit cell.

Preamplifiers such as the feedback-enhanced direct injection, capacitor feedback transimpedance amplifier, and source follower per detector types have amplifiers within the unit cell that require substantial bias current (on the order of 0.5 to 6 μA per amplifier). This results in a power of 3 to 40 μW per unit cell. In the case of the SFD, power can be cycled off except during the multiplexer read time.

Signal processing in the unit cell can also consume significant power if buffer amplifiers, such as the source follower, are utilized. These buffers consume about 1 to 3 μA per buffer. A typical bias voltage is 6 V, resulting in 5 to 20 μW of power per buffer.

Output video amplifier power is covered in Sec. 5.9.

5.11 DYNAMIC RANGE

Dynamic range (DR) is usually defined as maximum unsaturated photon flux, Q_{sat}, divided by NEI, as measured at the minimum specified background irradiance:

$$DR = \frac{Q_{sat}}{NEI} \tag{5.93}$$

or

$$DR = 20 \log\left(\frac{Q_{sat}}{NEI}\right) \quad [dB] \ . \tag{5.94}$$

Equation (5.93) defines instantaneous dynamic range, in which no user or external control is utilized to lower the gain for high-amplitude signals. Automatic gain switching, reduced integration time, and other methods can be used to achieve higher effective system dynamic ranges. Extremely high instantaneous dynamic ranges can be achieved through the use of nonlinear gain amplifiers in the readout circuit. The typical dynamic range of a focal plane can be calculated by dividing the maximum output voltage swing (on the order of 5 V) by the output noise floor of the sensor (typically on the order of 1 mV or less), and is usually in the range of 60 to 80 dB.

The maximum output voltage swing for any given ROIC is usually limited by either the integrated circuit process itself or the maximum acceptable focal plane power restrictions. The challenge of the readout designer is to achieve a large signal swing with the lowest possible noise floor. The minimum noise floor is a function of noise contributions of the preamplifier, multiplexer, and output driver, as well as external EMI. These noise sources can be filtered to some degree by on-focal-plane or off-focal-plane electronics. For example, the output of low data rate sensors can be filtered to reduce noise levels in astronomy applications.

For white noise, the rms noise contribution of a given source is proportional to the square root of the bandwidth. This relationship is useful in determining the dynamic range capability of a sensor. High data rate multiplexers and video drivers require high bandwidths. Therefore, one technique of increasing dynamic range is to decrease the corresponding bandwidth (data rate), and hence the noise of each output, by decreasing the number of elements multiplexed to a single output and then increasing the total number of outputs on the readout.

The general trend of minimum output noise versus multiplexer data rate is plotted in Fig. 5.30. The right axis shows dynamic range, in terms of digitizer

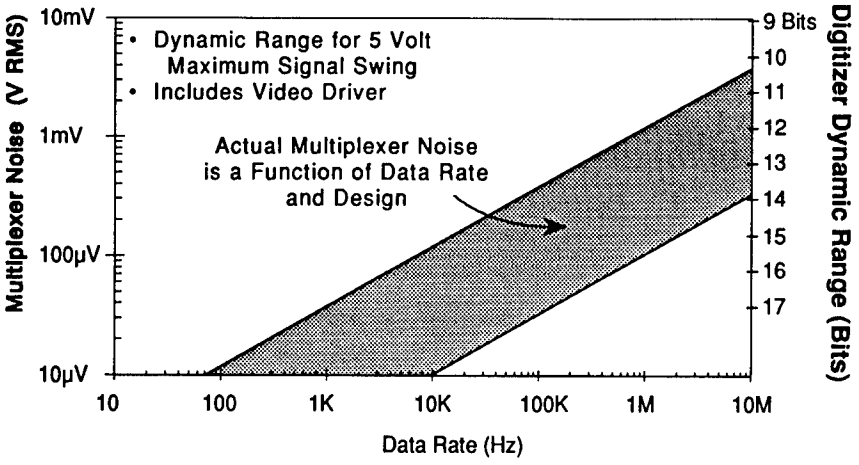


Fig. 5.30 Multiplexer noise floor and dynamic range are a function of data rate of a given output. A 12-bit dynamic range can be achieved in the 1- to 5-MHz region.

bits, for a maximum voltage swing of 5 V; the LSB (least significant bit) corresponds to the minimum sensor noise at the given data rate. The resolution of the focal plane falls to 12 to 13 bits in the 1-MHz data rate range, and is reduced to the 10 to 11 range at about the 5-MHz data rate.

5.12 CROSSTALK AND FREQUENCY RESPONSE

Two basic forms of signal crosstalk occur in SCAs. The first type is crosstalk from one physical detector element to another, which can be either electrical or optical in nature; the second is crosstalk on a detector element between one frame of data and its next frame of data (in time), which can also be expressed as frequency response in scanning sensors. Because a fraction of the photon energy striking a detector element is sensed in either an adjacent element or in a subsequent frame, both forms of crosstalk result in decreased image sharpness. Crosstalk can also manifest as a ghost or a latent image that fades with time. Poor frequency response in staring sensors can result in slow response to changes in the scene. Because frequency response of the detector channel is often a function of the specific readout preamplifier design, the use of preamplifier circuits that completely reset all storage elements (capacitors), prior to acquisition of a new frame of data, should be considered in critical applications where high-frequency response is required.

The output video amplifier is also a source of crosstalk problems in focal plane designs. The video amplifier, and all subsequent off-focal-plane circuits, must settle to a sufficient level prior to signal digitization. This settling level can be a source of confusion, because it is often specified to be within the system digitizers' least significant bit, which is in fact rarely required. Figure 5.31 shows a simplified representation of the output of a focal plane, where r_{out} is the video amplifier output impedance and C_{load} is the total output load capacitance including cables and other strays. Assuming the frequency response of the output data stream is dominated by a single pole of the circuit, the frequency response is

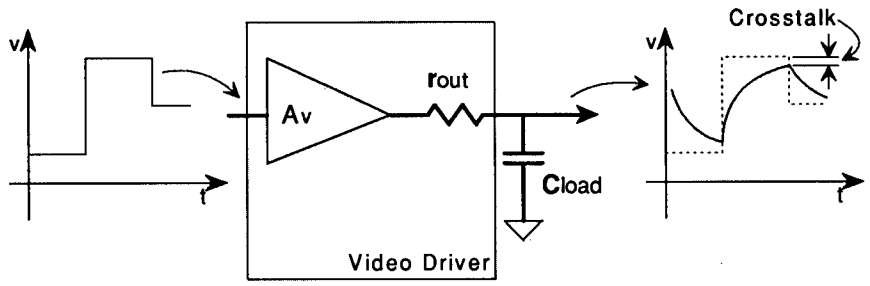


Fig. 5.31 The output configuration of a ROIC is designed to minimize crosstalk and on-focal-plane power.

$$f = \frac{1}{2\pi C_{\text{load}} r_{\text{out}}} \quad [\text{Hz}] \tag{5.95}$$

and the crosstalk between adjacent data stream elements is approximately

$$\text{crosstalk} = \exp\left(\frac{-t}{r_{\text{out}} C_{\text{load}}}\right), \tag{5.96}$$

where t is the elapsed time between the onset of new data being driven onto the output and when the digitizer acquires, or samples, the analog data. The crosstalk required for a typical sensor (including the detector) is normally on the order of 0.5 to 10%. If the video output circuit were designed to settle within half the least significant bit, on a 12-bit digitizer, the crosstalk of the video stage would be forced to less than 0.01%. Crosstalk this low would require a very high video circuit bandwidth and, since noise is related to bandwidth, this could result in increased noise. Settling time, or crosstalk, should be allocated in a reasonable fashion between the detector, the preamplifier, the video driver, and subsequent circuitry. The dominant output pole need not be the elements shown in Fig. 5.31, but could be characteristics of other downstream amplifier stages in the signal processing chain prior to digitization. The same analysis can be made for these other elements by appropriate substitution in Eqs. (5.95) and (5.96).

In most focal plane applications, significant crosstalk between elements that are not located optically adjacent to each other is unacceptable. An example is that in which subsequent columns of detector data are multiplexed to a single data stream; crosstalk occurs between the last element in one row of detector elements and the first element in the next row. In this case, a faster video settling time may be required at the expense of reduced noise bandwidth.

5.13 DESIGN METHODOLOGY

The full readout design includes many components not covered in the limited space of this chapter. This information, together with the references and bibliography, provides many of the tools required for successful circuit trades and designs. The circuit design approach is only a part of the overall procedure.

As mentioned, the design requires a well-understood set of requirements allocated to all sensor subcomponents. Circuit modeling and design is an iterative approach that leads to successful integrated circuit layout and processing. Performance of the resultant readout device is then fed back into the model to provide information for redesign or for future ROIC applications.

Acknowledgments

The author would like to recognize the following colleagues for the many discussions covering this subject: Will Frye, Mostyn Gale, Mark Goodnough, Mary Hewitt, Alan Hoffman, Joe Norworth, Terrence Lomheim, and Geof Orias. We also recognize support of the technology by the U.S. government, Santa Barbara Research Center, and Hughes Aircraft Company.

References

1. T. L. Koch, J. H. de Loo, M. H. Kalisher, and J. D. Phillips, "Monolithic n-channel HgCdTe linear imaging arrays," *IEEE Transactions on Electron Devices* **ED-32**(8) 1592–1598 (Aug. 1985).
2. M. E. McKelvey, C. R. McCreight, J. H. Goebel, N. N. Moss, and M. L. Savage, "Characterization of direct readout Si:Sb and Si:Ga infrared detector arrays for space-based astronomy," *Proceedings of the SPIE* **868,** 73–80 (1987).
3. S. M. Sze, *Physics of Semiconductor Devices,* 2nd ed., John Wiley & Sons, New York (1981).
4. A. S. Grove, *Physics and Technology of Semiconductor Devices,* John Wiley & Sons, New York (1967).
5. S. Lui and L. W. Nagel, "Small-signal MOSFET models for analog circuit design," *IEEE Journal of Solid-State Circuits* **SC-17**(6), 983–998 (Dec. 1982).
6. C. D. Motchenbacher and F. C. Fitchen, *Low-Noise Electronic Design,* John Wiley & Sons, New York (1973).
7. I. M. Hafez, G. Ghibauso, and F. Balestra, "A study of flicker noise in MOS transistors operated at room and liquid helium temperatures," *Solid State Electronics* **33**(12), 1525–1529 (1990).
8. W. V. Backensto and C. R. Viswananathan, "Bias-dependent 1/f noise model of an MOS transistor," *IEEE Proceedings* **127**(2), 87–92 (Apr. 1980).
9. M. D. Nelson, J. F. Johnson, and T. S. Lomheim, "General noise processes in hybrid infrared focal plane arrays," *Optical Engineering* **30**(11), 1682–1700 (Nov. 1991).
10. A. W. Hoffman, "Operation and calibration of self-integrating multiplexed arrays," *Proceedings of Workshop on Ground Based Astronomical Observations with Infrared Detectors,* C. G. Williams, B. E. Ecklin, Eds., March 1987, Hilo, Hawaii, p. 29.
11. R. Swanson and J. D. Meindl, "Complementary MOS transistors in low voltage circuits," *IEEE Journal of Solid-State Circuits* **SC-7**(2), 146–153 (Apr. 1972).
12. S.-L. Jang, "A model of 1/f noise in polysilicon resistors," *Solid State Electronics* **33**(9), 1155–1162 (1990).
13. D. E. Ludwig, N. D. Woodall, and M. M. Spanish, "On-focal plane analog-to-digital conversion with detector gain and offset compensation," *Proceedings of the SPIE* **1097,** 73–84 (1989).
14. G. S. Hobson, *Charge-Transfer Devices,* John Wiley & Sons, New York (1978).
15. M. J. Howes and D. V. Morgan, *Charge-Coupled Devices and Systems,* John Wiley & Sons, New York (1979).
16. C. H. Sequin and M. F. Tompsett, *Charge Transfer Devices,* Academic Press, New York (1975).
17. R. D. Thom, T. L. Koch, J. D. Langan, and W. J. Parrish, "A fully monolithic InSb infrared CCD array," *IEEE Transactions on Electron Devices* **ED-27**(1), 160–170 (Jan. 1980).

Bibliography

Bailey, R. B., L. J. Kozlowski, J. Chen, D. Q. Bui, K. Vural, D. D. Edwall, R. V. Gil, A. B. Vanderwyck, E. R. Gertner, and M. B. Gubala, "256 × 256 hybrid HgCdTe infrared focal plane arrays," *IEEE Transactions on Electron Devices* **38**(5), 1104–1109 (May 1991).

Capone, B., L. Skolnik, R. Taylor, F. Shepherd, S. Roosild, W. Ewing, W. Kosonocky, and E. Kohn, "Evaluation of a Schottky infrared charge-coupled device (IRCCD) staring mosaic focal plane," *Optical Engineering* **18**(5), 535–541 (Sep./Oct. 1979).

Carson, J. C., and S. N. Shanken, "High volume producibility and manufacturing of Z-plane technology," *Proceedings of the SPIE* **1097,** 138–149 (1989).

Celik-Butler, Z., and T. Y. Hsiang, "Spectral dependence of 1/f noise on GATE bias in N-mosfets," *Solid State Electronics* **30**(4), 419–423 (1987).

Cohen, J., *Introduction to Noise in Solid State Devices,* NBS TN 1169, National Bureau of Standards, Washington, DC (Dec. 1982).

Fang, Z.-H., A. Chovet, Q.-P. Zhu, and J.-N. Zhao, "Theory and applications of 1/f trapping noise in MOSFETs for the whole biasing ranges," *Solid State Electronics* **34**(4), 327–333 (1991).

Fleetwood, D. M., and J. H. Scofield, "Evidence that similar point defects cause 1/f noise and radiation-induced-hole trapping in metal-oxide-semiconductor transistors," *Physical Review Letters* **64**(5), 579–582 (Jan. 1990).

Forrest, W. J., A. Moneti, C. E. Woodward, J. L. Pipher, and A. Hoffman, "The new near-infrared array camera at the University of Rochester," *Astronomical Society of the Pacific* **97,** 183–198 (Feb. 1985).

Fowler, A. M., I. Gatley, F. Stuart, R. R. Joyce, and R. G. Probst, "The NOAO 1–5 micron imaging camera: a new national resource," *Proceedings of the SPIE* **972,** 107–121 (1988).

Fronen, R. J., and F. N. Hooge, "1/f noise in a p-i-n diode and in a diode laser below threshold," *Solid State Electronics* **34**(9), 977–982 (1991).

Grabowski, F., "Influence of dynamical interactions between density and mobility of carriers in the channel on 1/f noise of MOS transistors below saturation—I. mechanisms," *Solid State Electronics* **32**(10), 909–913 (1989).

Grabowski, F., "Influence of dynamical interactions between density and mobility of carriers in the channel on 1/f noise of MOS transistors below saturation—II. implications," *Solid State Electronics* **32**(10), 915–918 (1989).

Gustafsson, S., R. Sundblad, and C. Svensson, "Noise behavior of a static induction transistor between 77K and 300 K," *Solid State Electronics* **30**(4), 439–443 (1987).

Janesick, J. R., T. Elliott, S. Collins, M. M. Blouke, and J. Freeman, "Scientific charge-coupled devices," *Optical Engineering* **26**(8), 692–714 (Aug. 1987).

Johnson, J. F., and T. S. Lomheim, "Hybrid infrared focal plane signal and noise modeling," *Proceedings of the SPIE* **1541,** 110–126 (1991).

Lee, T. H., W.-C. Chang, W. A. Miller, G. R. Torok, K. Y. Wong, B. C. Burkey, and R. P. Khosla, "A four million pixel CCD image sensor," *Proceedings of the SPIE* **1242,** 10–16 (1991).

Lomheim, T. S., R. M. Shima, J. R. Angione, W. F. Woodward, D. J. Asman, R. A. Keller, and L. W. Schumann, "Imaging charge-coupled device (CCD) transient response to 17 and 50 MeV proton and heavy-ion irradiation," *IEEE Transactions on Nuclear Science* **37**(6), 1875–1885 (Dec. 1990).

McCreight, C. R., J. A. Estrada, J. H. Goebel, M. E. McDelvey, D. D. McKibbin, R. E. McMurray, Jr., T. T. Weber, J. Farhoomand, N. N. Moss, and M. L. Savage, "Low-background detector arrays for infrared astronomy," *Proceedings of the SPIE* **973,** 250–255 (1988).

McLean, I. S., *Electronic and Computer-Aided Astronomy,* John Wiley & Sons, New York (1989).

Mikoshiba, H., "1/f noise in n-channel silicon-gate MOS transistors," *IEEE Transactions on Electron Devices* **ED-29**(6), 965–970 (June 1982).

Mullens, R. S., and T. I. Kamins, *Device Electronics for Integrated Circuits,* John Wiley & Sons, New York (1977).

Norton, P. R., "Infrared image sensors," *Optical Engineering* **30**(11), 1649–1663 (Nov. 1991).

Oh, S.-Y., and R. W. Dutton, "A simplified two-dimensional numerical analysis of MOS devices—DC case," *IEEE Transactions on Electron Devices* **ED-27**(11), 2101–2110 (Nov. 1980).

Reimbold, G., "Modified 1/f trapping noise theory and experiments in MOS transistors biased from weak to strong inversion—influence of interface states," *IEEE Transactions on Electron Devices* **ED-31**(9), 1190–1198 (Sep. 1984).

Schiebel, R. A., "1/f noise in HgCdTe MISFETs," *Solid State Electronics* **32**(11), 1003–1007 (1989).

Skotnick, T., and W. Marciniak, "A new approach to threshold voltage modeling of short-channel MOSFETs," *Solid State Electronics* **29**(11), 1115–1127 (1986).

Swanson, R., and J. D. Meindl, "Ion-implanted complementary MOS transistors in low-voltage

circuits," *IEEE Journal of Solid-State Circuits* **SC-7**(2), 146–153 (Apr. 1972).

Van Der Ziel, A., "Integral expression for 1/f noise in MOSFETS at arbitrary drain bias," *Solid State Electronics* **29**(1), 29–30 (1986).

Van Vliet, C. M., "A survey of results and future prospects on quantum 1/f noise and 1/f noise in general," *Solid State Electronics* **34**(1), 1–21 (1991).

Wu, E. N., and A. Van Der Ziel, "On the influence of substrate doping on the input conductance and the induced GATE noise in MOSFETS," *Solid State Electronics* **27**(11), 945–946 (1984).

Zhu, X. C., A. Van Der Ziel, and K. H. Duh, "Low-frequency noise spectra in MOSFETS made by the DMOS process," *Solid State Electronics* **28**(4), 325–328 (1985).

CHAPTER 6

Thermal and Mechanical Design of Cryogenic Cooling Systems

P. Thomas Blotter
J. Clair Batty
Utah State University
Space Dynamics Laboratory
Logan, Utah

CONTENTS

6.1 INTRODUCTION

This chapter deals with cryostats designed to provide desired temperatures in electro-optical instruments. The approach is a practical "how to" one with essential basic theory and data accompanied by specific application examples. The examples are at the "hand-held calculator" level. References are made to computer codes that may be useful to the serious professional involved in cryostat design. The topic of cryostats, often called "coolers," is so extensive that only a survey of fundamental concepts can be included here. References and an extensive bibliography are included to provide the reader with access to the vast body of literature that exists on the topic.

6.1.1 Fundamental Tasks

The fundamental tasks in the design of a cryogenic cooling system are to provide:

1. *An appropriate low-temperature heat sink:* The heat sink could be (a) a depletable liquid or solid cryogen within the system, (b) a refrigerator and an appropriate power or heat source, or (c) low-temperature radiators. These three types of heat sinks are discussed in this chapter.
2. *Adequate thermal isolation from a warm environment:* Temperature differences of several hundred degrees often exist between the low-temperature heat sink and the surroundings, providing a large potential for overwhelming heat fluxes.
3. *Proper thermal linkage between cooled instrument components and the heat sink:* Focal plane arrays and other optical components must usually be maintained at steady temperatures within rather narrow limits. This often provides an immense challenge in thermal design.
4. *Electrical reliability:* From an electrical point of view, wires running between the warm surroundings and cooled components inside the instrument should have maximum conductivity, maximum diameter, and minimum length and be well shielded to minimize voltage drops and signal distortion. From a thermal point of view, wires are heat paths into the cold world of the cryostat and should have minimum conductivity, minimum diameter, and maximum length. Thus, there is a direct conflict in objectives and compromise is always necessary.
5. *Mechanical integrity of the entire system:* This mechanical task is often in direct conflict with the thermal tasks listed above. Cooled components usually must be precisely located and rigidly supported, often under harsh dynamic conditions such as a rocket ride into space. From a mechanical point of view, the support structure should be massive and strong to provide the necessary rigidity. From a thermal point of view, the support structure should be light with small cross-sectional areas to minimize parasitic heat transfer. Again, compromise is necessary to arrive at a satisfactory design.

6.2 BASIC PRINCIPLES OF THERMAL DESIGN

6.2.1 Conduction

Conduction occurs when more energetic (higher temperature) particles of a solid, liquid, or gaseous medium strike their less energetic (lower temperature) neighbors more often and with greater vigor than the lower temperature neighbors strike back. This is described mathematically by the Fourier one-dimensional heat conduction law as

$$Q = -kA\, dT/dx \ , \tag{6.1}$$

where

Q = heat rate (watts) in the x direction
dT/dx = change of temperature with respect to change in position (K/m) (The negative sign is necessary because the temperature must decrease as x increases for heat to flow in the positive x direction.)
k = thermal conductivity (W/m K)
A = heat transfer area (m^2) normal to the direction of heat flow.

The greatest challenge in conduction heat transfer is to determine precisely the thermal conductivity of the material proposed for use in a thermally sensitive system. Transport of thermal energy in a solid is due to two major effects: Vibrational waves of the atomic lattice and migration of the free electrons. Thermal conductivity k is thus regarded as the sum of a lattice component k_l and an electronic component k_e:

$$k = k_l + k_e \ . \tag{6.2}$$

In pure metals k_e is generally much larger than k_l. In metal alloys, the lattice contribution increases in relative importance and the dependence of thermal conductivity on temperature differs dramatically from that of pure metals.

Nonmetallic solids usually have a lower thermal conductivity than metals; however, certain highly ordered crystalline solids such as diamond, sapphire, or beryllium oxide can have a value of k_l that exceeds k for many good metallic conductors.

The designer of conductive thermal linkages in cryogenic systems must recognize that the thermal conductivity of the link may be dramatically influenced by temperature, oxygen content, and purity as well as individual piece history involving annealing, welding, soldering, work hardening, and surface oxidation.

Thermal insulators, used to isolate thermally components at different temperatures from each other in cryogenic systems, utilize materials of low thermal conductivity such as Teflon, Kevlar, or composite materials. Again, it is important to realize that the thermal conductivity of these materials can vary significantly, not only with temperature and composition, but also configuration and orientation.

Conduction Example 1: Low Thermal Impedance Link. We want to provide a conductive thermal link, 0.66 m in length, between a heat sink at 9.7 K and

a focal plane mount to be maintained at $T \leq 9.9$ K. The total heat load to be conducted through the link is 30 mW. Select an appropriate material and determine the required cross-sectional area of the link.

Solution and Comment. The very small temperature drop allowed, together with the relatively long length of the link, suggests that a material with a high thermal conductivity should be selected. One reasonable choice would be well-annealed 0.99999 pure copper having an expected thermal conductivity near 1000 W/m at 10 K. The link is so nearly constant in temperature that variations in thermal conductivity with temperature are negligible. From the Fourier conduction law

$$\begin{aligned} Q\,dx &= -kA\,dT \\ Q\int_0^L dx &= -kA\int_{T_1}^{T_2} dT \\ QL &= kA(T_1 - T_2) \\ Q &= \frac{(T_1 - T_2)}{L/kA} = \frac{\Delta T}{R}\ , \end{aligned} \tag{6.3}$$

where L is the length of the link in meters and $L/kA = R$ is the thermal resistance of the link in kelvins per watt. This thermal link must have a thermal resistance less than

$$\frac{0.2\text{ K}}{0.030\text{ W}} = 6.7\text{ K/W} \tag{6.4}$$

or

$$A = \frac{QL}{k\Delta T} = \frac{0.030(0.66)}{1000(0.2)} = 9.9 \times 10^{-5}\text{ m}^2\ . \tag{6.5}$$

This area would be provided by a cylindrical rod having a diameter of 0.011 m = 0.442 in. A 0.5-in. nominal diameter off-the-shelf rod should be adequate.

Machining, bending, or heating of the rod should be minimized because such procedures have been found to reduce the thermal conductivity. It is also assumed that the thermal link is located in a low-temperature cavity such that radiation heat loads are negligible.

Conduction Example 2: Thermal Isolator. We need to isolate thermally a telescope forebaffle at an average temperature of 60 K from a base plate at an average temperature of 10 K. A glass epoxy (G-10) cylinder 15 in. (0.381 m) in outer diameter having a wall thickness of 0.045 in. (1.143 mm) is proposed. What length is required to provide a thermal resistance of 800 K/W for the isolator?

Solution and Comment. Since the temperature difference is large, the variation of thermal conductivity of G-10 with temperature must be taken into account. Hence, we write

$$Q = \frac{-kA\ dT}{dx}$$

$$= \frac{-A}{L}\int_{T1}^{T2} k\ dT$$

$$= \frac{A}{L}\left(\int_{T\text{ref}}^{T1} k\ dT - \int_{T\text{ref}}^{T2} k\ dT\right) , \tag{6.6}$$

where k is a function of temperature. Values of k and $\int k\ dT$ for several materials are shown in Tables 6.1 through 6.10 and Figs. 6.1 through 6.10. Selecting values from Table 6.8 we write

$$Q = \frac{\pi(0.380\text{ m})(0.001143\text{ m})}{L\text{ (m)}}(0.099 - 0.006)\frac{\text{W}}{\text{cm}}\left(\frac{100\text{ cm}}{\text{m}}\right)$$

$$= \frac{0.0127}{L}\text{ W} . \tag{6.7}$$

Also

$$Q = \frac{\Delta T}{R} , \tag{6.8}$$

where ΔT is the temperature difference between the warm and cold ends and R is the thermal resistance. Hence,

$$\frac{0.0127}{L} = \frac{\Delta T}{R} = \frac{60 - 10}{800}\frac{\text{K}}{\text{(K/W)}}$$

$$L = (0.0127)\left(\frac{800}{50}\right) = 0.203\text{ m} = 8.0\text{ in.} . \tag{6.9}$$

Conduction Example 3: Wire Heat Load. An electrolytic tough pitch copper wire, 0.020 in. (5.08×10^{-4} m) in diameter and 24 in. (0.61 m) long, runs from the 300 K vacuum shell of a cryostat to an internal component at 10 K. Estimate the rate of heat conducted through the wire neglecting I^2R and radiation loads.

Solution and Comments. We will take into account the variation in thermal conductivity due to the large temperature difference over the length of the wire:

$$Q = \frac{-kA\ dT}{dx}$$

$$= \frac{A}{L}\int_{T2}^{T1} k\ dT$$

$$= \frac{A}{L}\left[\int_{T\text{ref}}^{T1} kdT - \int_{T\text{ref}}^{T2} kdT\right] \tag{6.10}$$

Table 6.1 Thermal Conductivity and Integral of Aluminum Alloy—6061

T (K)	k (W/cm K)	$\int$ kdT (W/cm)
350	1.830	445.300
325	1.785	402.925
300	1.740	360.550
275	1.695	318.175
250	1.650	275.800
225	1.570	236.550
200	1.490	197.300
190	1.445	182.850
180	1.400	168.400
170	1.370	154.700
160	1.340	141.000
150	1.295	128.050
140	1.250	115.100
130	1.205	103.050
120	1.160	91.000
110	1.120	79.800
100	1.080	68.600
90	1.020	58.100
80	0.980	48.100
77	0.960	45.190
70	0.920	38.600
65	0.900	34.105
60	0.880	29.610
55	0.840	25.410
50	0.800	21.210
45	0.740	17.510
40	0.680	13.810
35	0.600	10.610
30	0.520	7.810
25	0.450	5.385
20	0.348	3.390
15	0.252	1.890
10	0.164	0.935
8	0.130	0.553
6	0.105	0.318

Table 6.2 Thermal Conductivity and Integral of Copper (from Ref. 1)

T (K)	k (W/cm K)	$\int$ kdT (W/cm)
350	3.920	5402.201
325	3.950	5301.951
300	3.980	5201.701
275	4.010	5101.451
250	4.040	5001.201
225	4.085	4899.076
200	4.130	4796.951
190	4.160	4754.901
180	4.190	4712.850
170	4.220	4670.800
160	4.250	4628.750
150	4.336	4584.176
140	4.423	4539.601
130	4.566	4492.501
120	4.710	4445.401
110	4.853	4398.301
100	4.996	4351.201
90	5.139	4304.101
80	6.260	4235.900
77	6.596	4215.440
70	7.380	4167.700
65	7.940	4133.600
60	8.500	4099.500
55	11.500	4027.000
50	14.500	3954.500
45	17.500	3882.000
40	20.500	3809.500
35	31.750	3650.750
30	43.000	3492.000
25	68.000	3214.500
20	105.000	2782.000
15	156.000	2133.000
10	196.000	1237.500
8	186.000	856.000
6	159.000	502.000

Note: Well-annealed 99.999% pure copper.

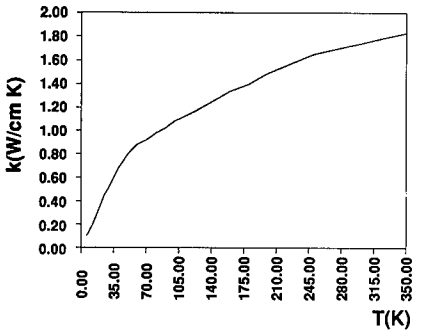


Fig. 6.1 Thermal conductivity of aluminum-6061.

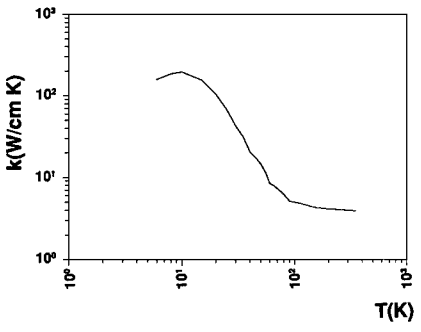


Fig. 6.2 Thermal conductivity of copper.

Table 6.3 Thermal Conductivity and Integral of OFHC—Copper (from Ref. 1)

T (K)	k (W/cm K)	∫ kdT (W/cm)
350	4.020	1697.744
325	4.025	1596.869
300	4.030	1495.994
275	4.034	1395.119
250	4.039	1294.244
225	4.049	1193.025
200	4.059	1091.806
190	4.068	1051.156
180	4.076	1010.506
170	4.097	969.584
160	4.118	928.663
150	4.142	887.216
140	4.167	845.769
130	4.217	803.640
120	4.267	761.512
110	4.377	717.917
100	4.487	674.322
90	4.772	627.432
80	5.153	577.570
77	5.289	561.925
70	5.687	523.942
65	6.149	493.516
60	6.612	463.090
55	7.301	425.851
50	7.990	388.613
45	8.967	342.800
40	9.944	296.988
35	11.014	248.388
30	11.365	192.750
25	10.430	137.645
20	8.731	92.100
15	6.200	51.100
10	4.800	23.600
8	3.600	15.200
6	2.800	8.800

Table 6.4 Thermal Conductivity and Integral of Copper-Beryllium (from Ref. 1)

T (K)	k (W/cm K)	∫ kdT (W/cm)
110	0.464	26.898
100	0.433	23.343
90	0.402	19.788
80	0.371	16.233
77	0.361	15.167
70	0.340	12.678
65	0.322	11.068
60	0.304	9.458
55	0.283	8.043
50	0.262	6.628
45	0.238	5.436
40	0.215	4.243
35	0.188	3.301
30	0.162	2.358
25	0.135	1.616
20	0.107	1.011
15	0.078	0.548
10	0.049	0.231
8	0.039	0.143
6	0.029	0.075

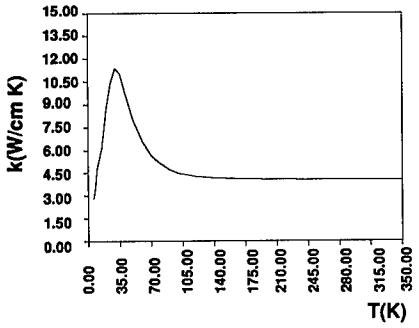


Fig. 6.3 Thermal conductivity of OFHC-copper.

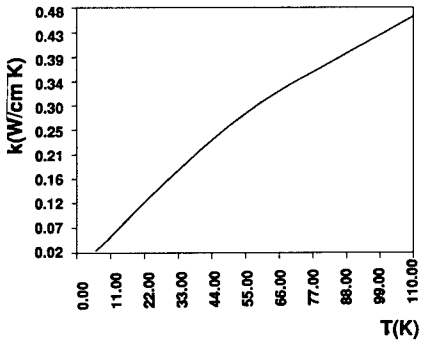


Fig. 6.4 Thermal conductivity of copper-beryllium.

Table 6.5 Thermal Conductivity and Integral of Copper Electrolytic Tough Pitch (from Ref. 1)

T (K)	k (W/cm K)	∫ kdT (W/cm)
350	3.866	1827.650
325	3.883	1728.900
300	3.900	1630.150
293	3.904	1602.500
275	3.916	1531.400
250	3.933	1432.650
225	3.950	1333.900
200	3.966	1235.150
190	3.973	1195.650
180	3.980	1156.150
170	3.986	1116.650
160	3.993	1077.150
150	4.000	1037.650
140	4.120	994.650
130	4.240	951.650
120	4.360	908.650
110	4.480	865.650
100	4.600	822.650
90	5.050	772.150
80	5.500	721.650
77	5.725	702.900
70	6.250	659.150
65	6.625	627.900
60	7.000	596.650
55	8.200	549.650
50	9.400	502.650
45	10.600	455.650
40	11.800	408.650
35	12.950	343.900
30	14.100	279.150
25	13.987	209.100
20	13.300	139.850
15	10.650	86.600
10	8.000	33.350
8	6.350	19.000
6	4.700	7.950

Table 6.6 Thermal Conductivity and Integral of Stainless Steel (from Ref. 1)

T (K)	k (W/cm K)	∫ kdT (W/cm)
350	0.162	38.615
325	0.157	34.879
300	0.153	31.142
275	0.148	27.406
250	0.143	23.669
225	0.137	20.261
200	0.130	16.853
190	0.127	15.588
180	0.124	14.323
170	0.122	13.078
160	0.119	11.834
150	0.115	10.677
140	0.112	9.520
130	0.108	8.436
120	0.105	7.353
110	0.099	6.356
100	0.094	5.360
90	0.088	4.439
80	0.083	3.586
77	0.081	3.338
70	0.078	2.759
65	0.074	2.397
60	0.070	2.035
55	0.065	1.713
50	0.059	1.392
45	0.053	1.110
40	0.046	0.829
35	0.039	0.641
30	0.032	0.453
25	0.026	0.287
20	0.019	0.165
15	0.013	0.095
10	0.006	0.030
8	0.005	0.021
6	0.003	0.012

Note: AISI 301, 303, 304.

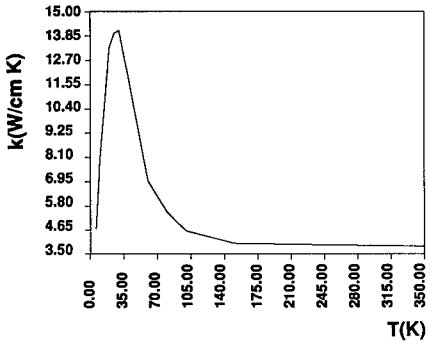


Fig. 6.5 Thermal conductivity of copper electrolytic tough pitch.

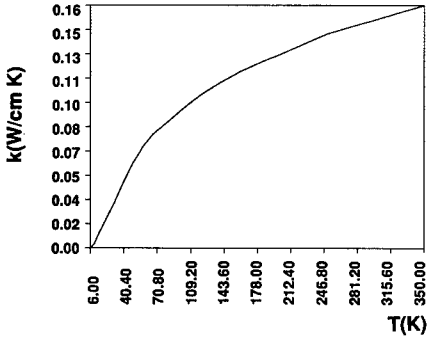


Fig. 6.6 Thermal conductivity of stainless steel.

Table 6.7 Thermal Conductivity and Integral of Glass-10-A

T (K)	k (mW/m K)	∫ kdT (W/cm)
350	1130.	2.024
325	1055.	1.816
300	980.	1.609
275	905.	1.401
250	830.	1.194
225	755.	0.986
200	680.	0.779
190	646.	0.719
180	612.	0.660
170	578.	0.600
160	544.	0.541
150	515.	0.489
140	486.	0.436
130	462.	0.391
120	438.	0.346
110	414.	0.301
100	390.	0.256
90	375.	0.219
80	360.	0.181
77	349.	0.170
70	325.	0.149
65	307.5	0.133
60	290.	0.116
55	275.	0.103
50	260.	0.090
45	245.	0.077
40	230.	0.064
35	215.	0.054
30	200.	0.043
25	187.5	0.033
20	175.	0.024
15	160.	0.016
10	145.	0.008
8	130.	0.005
6	101.	0.003

Note: G-10 tube cured under pressure.

Table 6.8 Thermal Conductivity and Integral of Glass-10-B

T (K)	k (mW/cm K)	∫ kdT (W/cm)
350	7.600	1.657
325	7.350	1.486
300	7.100	1.315
275	6.850	1.143
250	6.600	0.972
225	6.100	0.820
200	5.600	0.667
190	5.380	0.617
180	5.160	0.566
170	4.940	0.516
160	4.720	0.465
150	4.510	0.420
140	4.300	0.375
130	4.100	0.335
120	3.900	0.295
110	3.700	0.255
100	3.500	0.215
90	3.260	0.186
80	3.020	0.157
77	2.948	0.148
70	2.780	0.128
65	2.660	0.113
60	2.540	0.099
55	2.420	0.084
50	2.300	0.070
45	2.142	0.061
40	1.985	0.052
35	1.828	0.044
30	1.671	0.035
25	1.514	0.026
20	1.357	0.017
15	1.200	0.009
10	0.800	0.006
8	0.640	0.004
6	0.480	0.003

Note: G-10 support tube material (1543/E787 fiberglass).

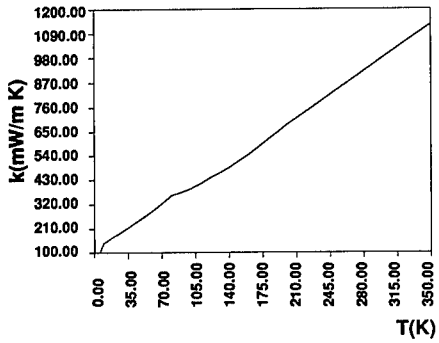


Fig. 6.7 Thermal conductivity of glass-10-A.

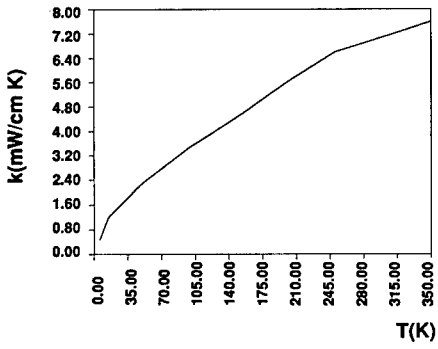


Fig. 6.8 Thermal conductivity of glass-10-B.

Table 6.9 Thermal Conductivity and Integral of Nylon (from Ref. 2)

T (K)	k (W/m K)	$\int$ kdT (W/m)
350	0.349	107.035
325	0.349	98.274
300	0.349	89.512
293	0.349	87.059
275	0.349	80.751
250	0.349	71.989
225	0.349	63.228
200	0.349	54.467
190	0.349	50.962
180	0.349	47.458
170	0.349	43.953
160	0.347	40.477
150	0.343	37.040
140	0.338	33.622
130	0.334	30.103
120	0.329	26.382
110	0.324	22.902
100	0.320	21.344
90	0.308	17.191
80	0.296	13.037
77	0.292	11.951
70	0.280	10.485
65	0.272	9.438
60	0.263	8.391
55	0.254	7.328
50	0.236	6.138
45	0.218	4.949
40	0.195	4.068
35	0.171	3.227
30	0.147	2.385
25	0.123	1.544
20	0.098	0.945
15	0.071	0.650
10	0.044	0.354
8	0.034	0.236
6	0.023	0.118

Table 6.10 Thermal Conductivity and Integral of Teflon (from Ref. 2)

T (K)	k (W/m K)	$\int$ kdT (W/m)
350	0.284	83.167
325	0.280	76.677
300	0.276	70.187
293	0.275	68.370
275	0.273	63.697
250	0.269	57.207
225	0.265	50.717
200	0.261	44.227
190	0.259	41.631
180	0.258	39.035
170	0.256	36.439
160	0.254	33.901
150	0.251	31.343
140	0.249	28.747
130	0.247	26.166
120	0.246	23.613
110	0.245	21.075
100	0.245	18.652
90	0.240	16.272
80	0.234	13.893
77	0.233	13.189
70	0.227	11.617
65	0.223	10.495
60	0.219	9.372
55	0.215	8.255
50	0.209	7.186
45	0.203	6.118
40	0.191	5.193
35	0.179	4.287
30	0.167	3.381
25	0.154	2.474
20	0.135	1.730
15	0.107	1.189
10	0.079	0.648
8	0.068	0.432
6	0.057	0.216

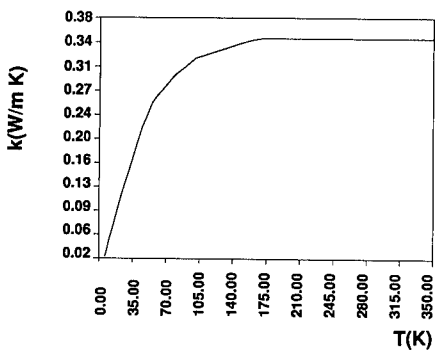


Fig. 6.9 Thermal conductivity of nylon.

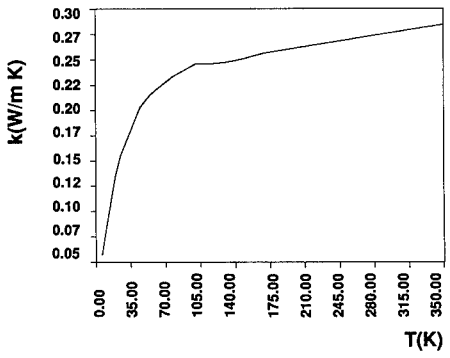


Fig. 6.10 Thermal conductivity of Teflon.

Using values for $\int k dT$ from Table 6.5 for electrolytic tough pitch copper

$$Q = \frac{\pi(5.08 \times 10^{-4}\ \text{m})^2}{4(0.61\ \text{m})}(1630.7 - 33)\frac{\text{W}}{\text{cm}}\frac{(100\ \text{cm})}{\text{m}}$$

$$= 53.1\ \text{mW} \ . \tag{6.11}$$

This example shows why copper wires impose large heat loads on cryostats. Other materials with lower thermal conductivities are generally preferred for low-amperage signal carrying wires. However, for current carrying power cables, material selection should be based on more than minimum thermal conductivity. If the selected wire diameter is too small, I^2R loads may be excessive. If the diameter is larger than necessary, the thermal conduction loads are increased.

Where electrical currents are small (less than ~0.1 A), the use of a low thermal conductivity material, such as manganin or constantan, is recommended to minimize total heat load. Where electrical currents are larger, a material such as copper or beryllium copper should be considered.

6.2.2 Heat Capacity

When concerned with cool down times or the amount of heat removal required to cool a particular component, the concept of heat capacity or thermal inertia becomes important. The heat capacity of a material is defined as

$$\text{heat capacity} = m \int_{T_1}^{T_2} C\, dT \ , \tag{6.12}$$

where

m = mass of the component (kg)
C = specific heat of the material (kJ/kg K)
T_1 = initial temperature
T_2 = final temperature.

For designers of cryogenic systems, accurate determination of the specific heat of the material as a function of temperature is a challenge. For many engineering materials, the specific heat becomes very small in the cryogenic temperature range.

Values of specific heat and $\int_{T_{\text{ref}}}^{T} C\, dT$ for several materials are shown in Tables 6.11 through 6.14 and Figs. 6.11 through 6.14. The use of the tables is illustrated by the following example.

Heat Capacity Example. Compare the heat inputs required to increase the temperature of 1 kg of 6061-T6 aluminum from 275 to 300 K with that required to change the temperature of the same material from 25 to 50 K.

Table 6.11 Specific Heat and Integral of Aluminum Alloy-6061

T (K)	C (J/g K)	$\int$ CdT (J/g)
350	0.931	200.790
325	0.905	179.439
300	0.879	158.088
275	0.853	136.738
250	0.827	115.387
225	0.780	95.893
200	0.733	76.400
190	0.714	69.256
180	0.695	62.113
170	0.672	55.386
160	0.648	48.660
150	0.616	42.438
140	0.584	36.216
130	0.550	30.943
120	0.516	25.670
110	0.468	20.670
100	0.420	15.670
90	0.346	11.880
80	0.300	8.650
77	0.290	7.700
70	0.250	5.900
65	0.216	4.824
60	0.182	3.748
55	0.154	2.969
50	0.126	2.190
45	0.105	1.660
40	0.085	1.130
35	0.061	0.738
30	0.045	0.502
25	0.033	0.372
20	0.021	0.241
15	0.009	0.110
10	0.000	0.000
8	0.000	0.000
6	0.000	0.000

Table 6.12 Specific Heat and Integral of Copper (from Ref. 3)

T (K)	C (J/g K)	$\int$ CdT (J/g)
350	0.399	98.467
325	0.392	88.930
300	0.386	79.393
275	0.379	69.855
250	0.373	60.318
225	0.356	51.208
200	0.340	42.098
190	0.342	38.627
180	0.345	35.157
170	0.338	31.768
160	0.331	28.380
150	0.321	25.159
140	0.312	21.938
130	0.300	18.929
120	0.288	15.920
110	0.270	13.215
100	0.251	10.510
90	0.230	8.096
80	0.202	5.929
77	0.194	5.340
70	0.171	4.055
65	0.153	3.287
60	0.135	2.519
55	0.116	1.936
50	0.096	1.353
45	0.077	0.964
40	0.059	0.575
35	0.041	0.323
30	0.026	0.150
25	0.015	0.047
20	0.007	0.000
15	0.000	0.000
10	0.000	0.000
8	0.000	0.000
6	0.000	0.000

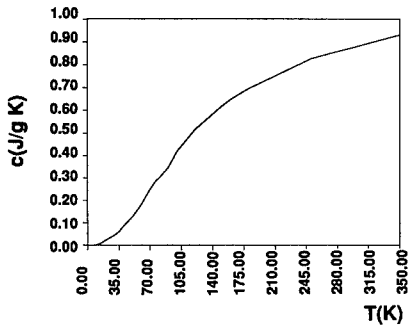


Fig. 6.11 Specific heat of aluminum-6061.

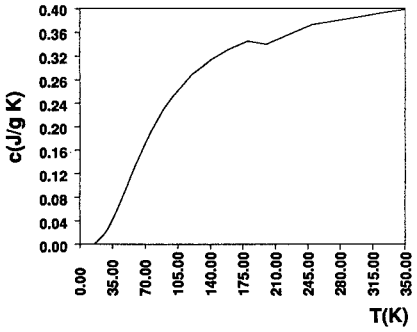


Fig. 6.12 Specific heat of copper.

Table 6.13 Specific Heat and Integral of OFHC-Copper (from Ref. 3)

T (K)	C (J/g K)	∫ CdT (J/g)
350	0.393	98.885
325	0.388	89.385
300	0.384	79.885
275	0.380	70.385
250	0.376	60.886
225	0.367	51.638
200	0.359	42.390
190	0.353	38.855
180	0.347	35.320
170	0.338	32.025
160	0.329	28.730
150	0.320	25.435
140	0.311	22.140
130	0.299	19.135
120	0.288	16.130
110	0.271	13.410
100	0.255	10.690
90	0.231	8.380
80	0.206	6.070
77	0.195	5.558
70	0.171	4.358
65	0.153	3.499
60	0.136	2.640
55	0.117	2.154
50	0.097	1.668
45	0.077	1.181
40	0.058	0.694
35	0.045	0.530
30	0.032	0.366
25	0.020	0.201
20	0.007	0.037
15	0.002	0.012
10	0.001	0.002
8	0.000	0.001
6	0.000	0.000

Table 6.14 Specific Heat and Integral of Copper Electrolytic Tough Pitch (from Ref. 3)

T (K)	C (J/g K)	∫ CdT (J/g)
1000	0.459	298.943
950	0.453	276.443
900	0.447	253.943
800	0.437	209.709
750	0.433	187.926
700	0.429	166.369
650	0.424	145.034
600	0.419	123.957
550	0.413	103.133
500	0.408	82.600
450	0.402	62.339
400	0.396	42.393
350	0.389	22.737
300	0.381	3.419

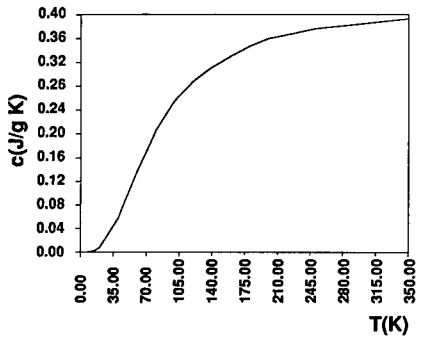


Fig. 6.13 Specific heat of OFHC-copper.

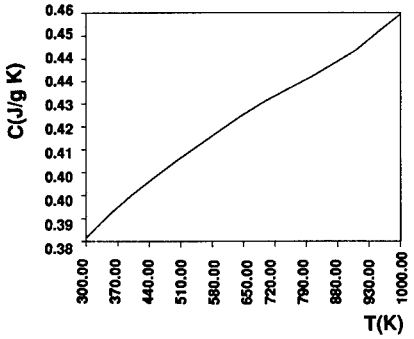


Fig. 6.14 Specific heat of copper electrolytic tough pitch.

Solution and Comments. From Table 6.11 the following values are obtained:

T	$\int C\,dT$
300	158.088
275	136.738
50	2.198
25	0.372

Heat required to increase the temperature of 1 kg of aluminum from 275 to 300 K is

$$1 \text{ kg } (158.1 - 137) \frac{\text{kJ}}{\text{kg}} = 21.4 \text{ kJ} . \tag{6.13}$$

Heat required to increase the temperature of 1 kg of aluminum from 25 to 50 K is

$$1 \text{ kg } (2.2 - 0.4) \frac{\text{kJ}}{\text{kg}} = 1.86 \text{ kJ} . \tag{6.14}$$

The heat capacity of 6061 aluminum in the 25 to 50 K range is less than 10% of the heat capacity in the 275 to 300 K range.

6.2.3 Thermal Expansion

Cryostats built at ambient temperature may experience temperature changes approaching 300 K as the cryogen is loaded. A change in lengths during this process depends on the material coefficient of thermal expansion, $\alpha(T)$(m/m K).

Since the coefficient of thermal expansion is often a strong function of temperature, the expression for $\alpha(T)$ should be integrated over the desired temperature range in order to determine the total change in material length. Much of the current literature presents thermal expansion data as integrals of the $\alpha(T)$ curve from a reference temperature to some specified final temperature. The reference temperature required for such integrals is usually taken to be 293 K:

$$\int_{293}^{T} \alpha(T)\,dT = \frac{L - L_0}{L_0} , \tag{6.15}$$

where

$\alpha(T)$ = coefficient of thermal expansion as a function of temperature
L = length of specimen at T K
L_0 = Length of specimen at 293 K.

Thermal expansion $\times$ (100.0) = percent thermal expansion(%):

$$\Delta L = L_0 \int_{293}^{T} \alpha(T)\,dt = L_0 \left(\frac{L - L_0}{L_0} \right) . \tag{6.16}$$

Examples of values of thermal expansion and $(L - L_0)/L_0$ are shown in Tables 6.15 through 6.18 and Figs. 6.15 through 6.18.

Table 6.15 Thermal Expansion of Aluminum Alloy-6061 (from Ref. 4)

T (K)	α x10^{-6} (1/K)	$(L-L_o)/L_o$ (%)
350	22.291	0.128
325	22.489	0.072
300	22.521	0.016
293	22.458	0.000
275	22.360	-0.039
250	21.964	-0.096
225	21.282	-0.149
200	20.247	-0.202
190	19.718	-0.221
180	19.114	-0.241
170	18.431	-0.259
160	17.662	-0.278
150	16.801	-0.295
140	16.368	-0.312
130	15.250	-0.327
120	14.455	-0.343
110	13.653	-0.356
100	12.623	-0.370
90	11.252	-0.381
80	9.534	-0.393
77	8.963	-0.395
70	7.572	-0.400
65	6.562	-0.404
60	5.579	-0.408
55	5.067	-0.410
50	4.594	-0.412
45	3.770	-0.414
40	2.635	-0.416
35	1.763	-0.416
30	1.295	-0.417
25	1.236	-0.418
20	1.158	-0.419
15	1.100	-0.419
10	1.066	-0.420
8	1.000	-0.420
6	0.837	-0.421

Table 6.16 Thermal Expansion of Copper (from Ref. 4)

T (K)	α x10^{-6} (1/K)	$(L-L_o)/L_o$ (%)
350	16.646	0.092
325	16.086	0.051
300	16.124	0.011
293	16.190	0.000
275	16.285	-0.029
250	16.256	-0.070
225	15.889	-0.110
200	15.199	-0.150
190	14.868	-0.164
180	14.530	-0.178
170	14.209	-0.193
160	13.932	-0.207
150	13.731	-0.221
140	13.202	-0.235
130	12.603	-0.247
120	11.961	-0.260
110	11.216	-0.271
100	10.341	-0.283
90	9.341	-0.292
80	8.586	-0.301
77	7.886	-0.304
70	7.022	-0.309
65	6.766	-0.312
60	6.355	-0.315
55	5.584	-0.318
50	4.429	-0.322
45	3.042	-0.323
40	2.542	-0.324
35	2.042	-0.325
30	1.542	-0.326
25	1.000	-0.327
20	0.500	-0.327
15	0.000	-0.327
10	0.000	-0.327
8	0.000	-0.327
6	0.000	-0.327

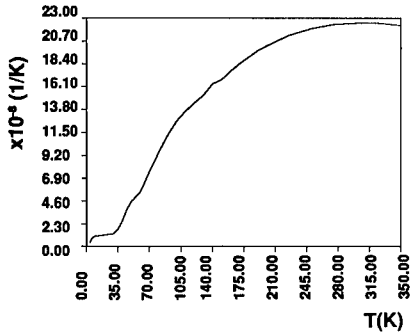


Fig. 6.15 Thermal expansion coefficient of aluminum-6061.

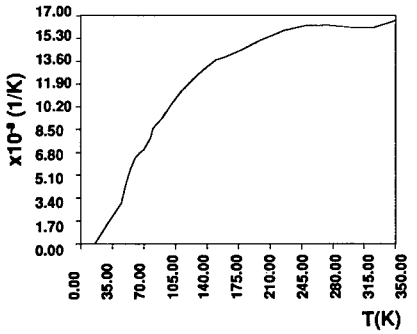


Fig. 6.16 Thermal expansion coefficient of copper

Table 6.17 Thermal Expansion of OFHC-Copper (from Ref. 4)

T (K)	α $\times 10^{-6}$ (1/K)	$(L-L_o)/L_o$ (%)
350	16.391	0.094
325	16.628	0.053
300	16.669	0.011
293	16.600	0.000
275	16.500	-0.029
250	16.137	-0.070
225	15.626	-0.109
200	15.041	-0.148
190	14.808	-0.163
180	14.587	-0.178
170	14.387	-0.192
160	14.217	-0.207
150	14.087	-0.220
140	13.130	-0.234
130	12.456	-0.247
120	11.993	-0.260
110	11.471	-0.271
100	10.714	-0.282
90	9.636	-0.293
80	8.248	-0.301
77	7.784	-0.304
70	6.654	-0.309
65	5.837	-0.312
60	5.049	-0.315
55	4.331	-0.317
50	3.429	-0.320
45	2.928	-0.321
40	2.447	-0.322
35	1.735	-0.323
30	1.158	-0.324
25	1.059	-0.325
20	0.557	-0.325
15	0.000	-0.326
10	0.000	-0.326
8	0.000	-0.326
6	0.000	-0.326

Table 6.18 Thermal Expansion of Stainless Steel AISI-302 (from Ref. 4)

T (K)	α $\times 10^{-6}$ (1/K)	$(L-L_o)/L_o$ (%)
350	17.526	0.136
325	17.419	0.091
300	17.682	0.047
293	17.652	0.000
275	17.626	-0.002
250	16.919	-0.042
225	15.587	-0.081
200	14.011	-0.120
190	13.466	-0.132
180	13.131	-0.145
170	13.096	-0.158
160	13.043	-0.171
150	13.035	-0.184
140	12.817	-0.197
130	12.406	-0.210
120	11.999	-0.223
110	11.598	-0.234
100	10.803	-0.246
90	9.448	-0.255
80	7.643	-0.264
77	7.065	-0.267
70	5.781	-0.271
65	5.028	-0.273
60	4.537	-0.276
55	4.039	-0.278
50	2.856	-0.281
45	1.897	-0.281
40	1.284	-0.282
35	1.010	-0.282
30	0.834	-0.282
25	0.634	-0.284
20	0.157	-0.284
15	0.000	-0.284
10	0.000	-0.284
8	0.000	-0.284
6	0.000	-0.284

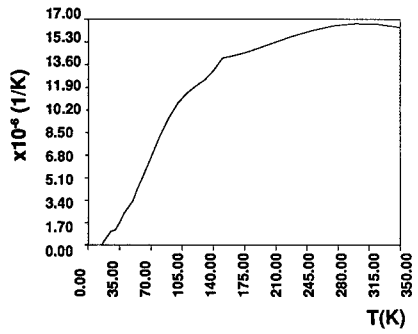


Fig. 6.17 Thermal expansion coefficient of OFHC-copper.

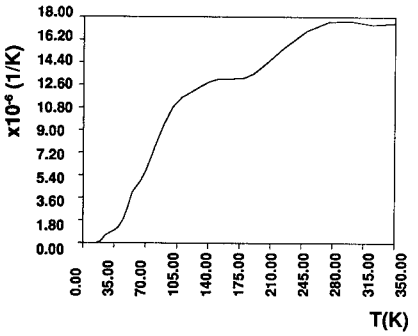


Fig. 6.18 Thermal expansion coefficient of stainless steel AISI-302.

Linear Thermal Expansion Example. Determine the change in length of a 0.5-m 6061-T6 aluminum bar after it has been cooled from 300 to 77 K (temperature of liquid nitrogen at standard atmospheric pressure).

Solution and Comments. We pick from Table 6.15 the values of the percent thermal expansion at 300 and at 77 K. Thus,

$$\Delta L = 0.5\left(\frac{0.16 - 0.395}{100}\right) = -2.055 \text{ mm} . \tag{6.17}$$

The bar has shortened 2.055 mm.

6.2.4 Radiation Exchange

All surfaces at temperatures above absolute zero radiate energy according to the Stefan-Boltzmann law:

$$E = \varepsilon\sigma T^4 = \varepsilon E_b , \tag{6.18}$$

where

E = emissive power (W/m^2) of the surface
ε = emissivity of the surface
σ = Stefan-Boltzmann constant, 5.67×10^{-8} W/m^2 K^4
T = absolute temperature (K)
E_b = emissive power of a perfect emitter or blackbody.

As depicted in Fig. 6.19, the wavelength of the radiant energy (emissive power) tends to increase as the temperature of the radiating surface decreases according to Planck's distribution law.

Other quantities that are important to understand in evaluating the radiation exchange between surfaces in a cryogenic system include:

α = absorptivity or the fraction of incident radiation absorbed by the surface
ρ = reflectivity or the fraction of incident radiation reflected from the surface
γ = transmissivity or the fraction of incident radiation transmitted through the surface
G = irradiation (W/m^2), the total rate at which radiant energy from all sources strikes the surface per unit area
J = radiosity (W/m^2), the total rate (emitted plus reflected) at which radiant energy leaves a surface per unit area
F_{12} = geometric shape factor, the fraction of radiant energy leaving surface 1 that strikes surface 2. (Geometric shape factors are available in the literature for a myriad of configurations and will not be reproduced here because of space limitations.)

Since all incident radiation must be either absorbed, reflected, or transmitted,

$$\alpha + \rho + \tau = 1 . \tag{6.19}$$

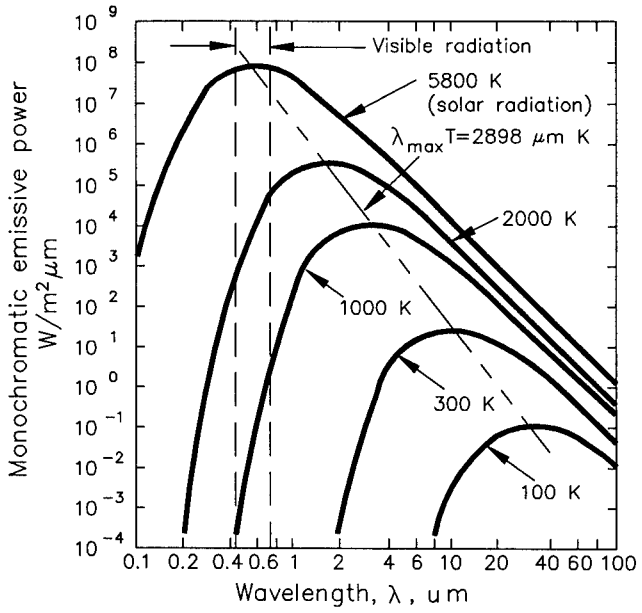


Fig. 6.19 Monochromatic emissive power of a black surface at various temperatures according to Planck's law.

For opaque surfaces where $\tau = 0$, $\alpha + \rho = 1$. In general, α depends on the spectral (wavelength) distribution of the irradiation, whereas ε is independent of irradiation. Under certain limited conditions, however, surface absorptivity and emissivity may be assumed to be independent of wavelength and direction. For such surfaces (referred to as gray diffuse), it may be shown that $\alpha = \varepsilon$.

Our discussion here will be limited to radiation exchange between two gray diffuse opaque surfaces that see each other and nothing else. Fortunately, this is often adequate for designing a cryostat. The reader is referred to any good heat transfer text (see the bibliography) for a more complete treatment of radiation.

The net rate at which radiation leaves surface 1 is given by

$$q_1 = A_1(J_1 - G_1) \ . \tag{6.20}$$

By definition,

$$\rho_1 = 1 - \alpha_1 = 1 - \varepsilon_1 \ , \tag{6.21}$$

$$J_1 = E_1 + \rho_1 G_1 = \varepsilon_1 E_{b1} + (1 - \varepsilon_1) G_1 \ . \tag{6.22}$$

Hence,

$$G_1 = \frac{J_1 - \varepsilon_1 E_{b1}}{1 - \varepsilon_1} \tag{6.23}$$

or

$$q_1 = A_1\left(J_1 - \frac{J_1 - \varepsilon_1 E_{b1}}{1 - \varepsilon_1}\right) = \frac{E_{b1} - J_1}{(1 - \varepsilon_1)/\varepsilon_1 A_1} \ . \tag{6.24}$$

The irradiation of surface 1 is given by:

$$A_1 G_1 = A_2 F_{21} J_2 + A_1 F_{11} J_1 \ . \tag{6.25}$$

Fundamental principles show that

$$A_2 F_{21} = A_1 F_{12} \ . \tag{6.26}$$

Using this well-known reciprocity relation, the net rate at which radiant energy leaves surface 1 is

$$q_1 = A_1(J_1 - G_1) = A_1(J_1 + F_{12} J_2 - F_{11} J_1) \ ; \tag{6.27}$$

but by definition for two surfaces exchanging radiation only with each other,

$$F_{11} = 1 - F_{12} \ , \tag{6.28}$$

$$q_1 = A_1(J_1 - F_{12} J_2 - J_1 + F_{12} J_1) = A_1 F_{12}(J_1 - J_2) \tag{6.29}$$

or

$$q_1 = \frac{J_1 - J_2}{(1/A_1 F_{12})} = \frac{E_{b1} - J_1}{(1 - \varepsilon_1)/\varepsilon_1 A_1} \ . \tag{6.30}$$

Thus for two diffuse gray opaque surfaces that "see" each other and nothing else, the net radiation exchange between surfaces 1 and 2 is given by

$$q_{12} = q_1 = \frac{E_{b1} - J_1}{(1 - \varepsilon_1)/\varepsilon_1 A_1} = \frac{J_1 - J_2}{(1/A_1 F_{12})} = -q_2 = \frac{J_2 - E_{b2}}{(1 - \varepsilon_2)/\varepsilon_2 A_2} \tag{6.31}$$

and represented by a potential-resistor electrical analog as shown in Fig. 6.20. Finally, with $E_b = \sigma T^4$,

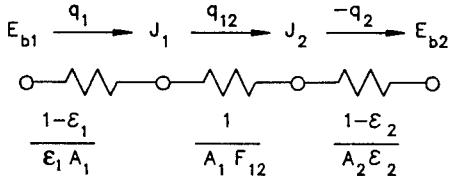


Fig. 6.20 Network analog for net radiation exchange between two diffuse gray surfaces that see each other and nothing else.

$$q_{12} = \frac{\sigma(T_1^4 - T_2^4)}{(1 - \varepsilon_1)/\varepsilon_1 A_1 + (1/A_1 F_{12}) + (1 - \varepsilon_2)/\varepsilon_2 A_2} . \tag{6.32}$$

Shown in Fig. 6.21 are the results of applying the net radiation exchange network to several common geometries.

Radiation Example 1: Radiant Heat Exchange Between Concentric Cylinders and Parallel Plates. Consider a cryostat containing a cold optics section at 80 K supported within a vacuum shell at 300 K as shown schematically in Fig. 6.22. Make an estimate of the radiation heat loads between the vacuum shell and optic section assuming all surfaces are electropolished aluminum.

Solution and Comments. Using Table 6.19 we estimate ε_2, the effective emissivity of the 300 K surface, to be 0.08 and ε_1, the effective emissivity of the 80 K surface, to be 0.04. We neglect the small difference in length of the concentric cylinders.

Exchange between the concentric cylinders is from Fig. 6.21:

$$q_{21} = \frac{\sigma A_1(T_2^4 - T_1^4)}{\dfrac{1}{\varepsilon_1} + \dfrac{1 - \varepsilon_2}{\varepsilon_2} + \dfrac{r_1}{r_2}} , \tag{6.33}$$

$$q_{12} = \frac{5.67\pi(0.50)(0.625)(3.0^4 - 0.80^4)}{\dfrac{1}{0.04} + \dfrac{1 - 0.08}{0.08} + \dfrac{0.50}{0.55}} = 12.65 \text{ W} . \tag{6.34}$$

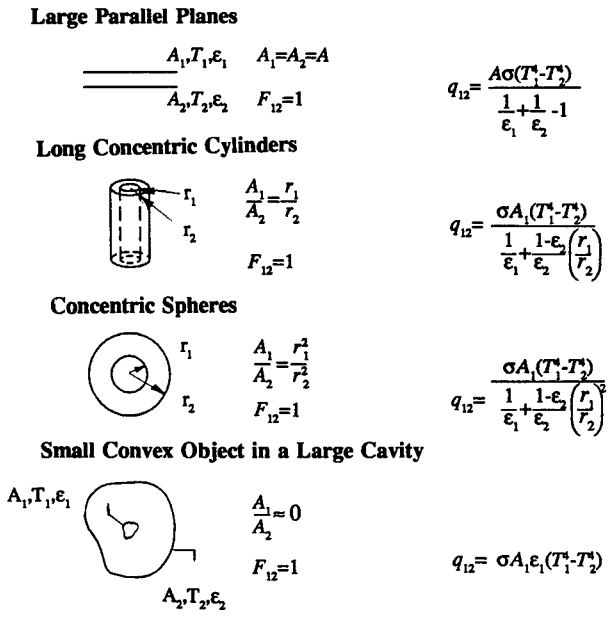


Fig. 6.21 Examples of net heat exchange between two opaque diffuse gray surfaces that see each other and nothing else.

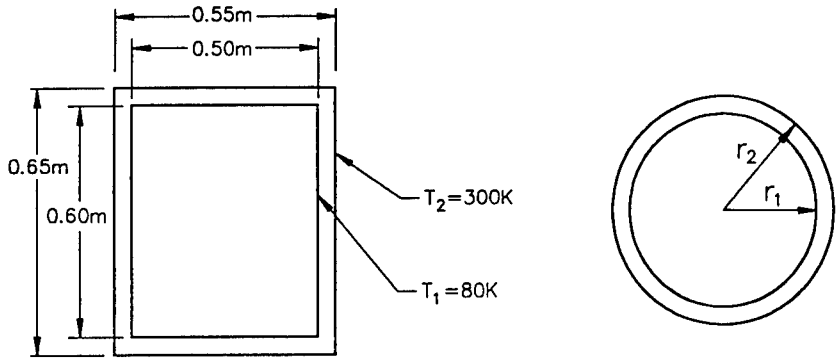


Fig. 6.22 Cryostatic schematic for radiation example.

Table 6.19 Emissivity Data for Selected Materials (from Ref. 5)

Surface ID	Emissivity at Ambient Temperature	Emissivity at Ambient to 77 K	Emissivity at 77 K to 4.2K
Al. ANOD. OP	0.98 ± 5%	0.84 ± 12%	0.75 ± 9%
Al. ANOD. OP SB	0.94 ± 5%	0.89 ± 9%	0.78 ± 9%
Al. ANOD. OP FWB		0.85 ± 8.5%	0.72 ± 4%
Al. ANOD. OP CWB		0.83 ± 8%	0.73 ± 8%
Al. ANOD. CP		0.80 ± 12%	0.76 ± 6%
Al. CLEAR ANOD.		0.78 ± 14%	0.67 ± 8%
Al. PROTECTIVE OXIDE LAYER		0.49 ± 9%	0.074 ± 7%
Al. AS FOUND		0.12 ± 16%	
Al. MECH. POLISHED		0.10 ± 9%	0.058 ± 6%
Al. ELECTROPOLISHED		0.075 ± 9%	0.036 ± 7%
St.St. AS FOUND		0.34 ± 8.5%	0.12 ± 5%
St.St. SB		0.24 ± 5%	0.14 ± 6%
St.St. MECH. POLISHED		0.12 ± 8%	0.074 ± 5%
St.St. ELECTROPOLISHED		0.10 ± 13%	0.065 ± 5%
St.St. SILVER PLATED		0.092 ± 9%	0.013 ± 9%
St.St. AS FOUND, Al. FOIL		0.056 ± 13%	0.011 ± 5%
St.St. SB, ALUMINIZED MYLAR	0.042*	0.050 ± 14%	0.18 ± 5%
Cu. AS FOUND		0.12 ± 11%	0.062 ± 14%
Cu. MECH. POLISHED		0.060 ± 9%	0.023 ± 12%
		0.022 ± 2.5%**	
Cu. ELECTROPOLISHED		0.035**	
NEXTEL on Al.	0.92 ± 5%	0.82 ± 7%	0.69 ± 6%
NEXTEL on St.St. 1/2 sample		0.84 ± 7%	0.67 ± 8%
NEXTEL on St.St.		0.80 ± 11%	0.67 ± 8%

Al. = Aluminum
Cu = Copper
St.St. = Stainless steel
ANOD = Black Anodized
CP = Closed Pore
CWB = Coarse wire brush
FWB = Fine wire brush
OP = Open Pore
SB = Shot blasted

* (Bapat, Narayankhedkar and Lukose 1990)
** (Touloukian 1970)

Treating the ends as large parallel planes, the heat exchange from Fig. 6.21 is given as

$$q_{21} = \frac{A\sigma(T_2^4 - T_1^4)}{\frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_2} - 1} = \frac{2[\pi(0.525)^2/4](5.67)(3^4 - 0.80^4)}{\frac{1}{0.04} + \frac{1}{0.08} - 1} = 5.42 \text{ W} . \tag{6.35}$$

The total radiant heat load between the vacuum shell and the optics section would thus be approximately 18.6 W.

Radiation Example 2: Radiation Shields. Estimate the rate of radiant heat exchange between the vacuum shell and the optics section of the previous example if an electropolished aluminum radiation shield is situated between as depicted schematically in Fig. 6.23.

Solution and Comment. In the radiation network also shown in Fig. 6.23, $J_{3,1}$ represents the radiosity of the shield surface facing surface 1, $J_{3,2}$ is the radiosity of the shield surface facing surface 2, and E_{b3} is the emissive power of a black surface at the shield temperature. The heat exchange is given by $(E_{b2} - E_{b1})/\Sigma R_i$ where ΣR_i is the sum of all the resistor terms. Assuming

$$\varepsilon_{31} = \varepsilon_{32} = 0.06 \quad \text{and} \quad F_{13} = F_{32} = 1 \tag{6.36}$$

results in $q_{21} = 9.5$ W.

The temperature of the shield can be estimated from

$$q_{21} = \frac{E_{b2} - E_{b3}}{\frac{1 - \varepsilon_2}{\varepsilon_2 A_2} + \frac{1}{A_2 F_{32}} + \frac{1 - \varepsilon_3}{\varepsilon_{3,2} A_3}} , \tag{6.37}$$

which results in a value of

$$E_{b3} = 282.6 \text{ W/m}^2 = \sigma T_3^4 \tag{6.38}$$

or $T_3 = 265.7$ K.

The above procedure could easily be extended to multiple shields. For the special case in which all areas are equal and all emissivities are equal, we can show that with N shields

$$(q_{21})_N = \frac{1}{N + 1}(q_{21})_0 , \tag{6.39}$$

where $(q_{21})_0$ represents the radiation heat load with no shields.

Thus one could expect that adding one shield would reduce the radiation by about 50%, two shields would provide a 67% reduction, and three a 75% reduction. Shielding is therefore widely used in cryostat design.

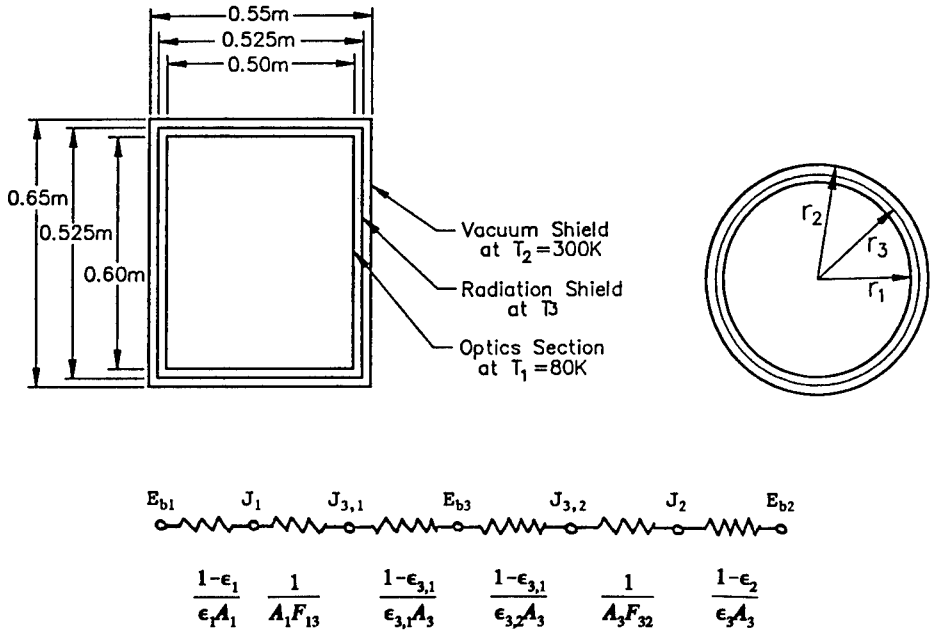


Fig. 6.23 Schematic of cryostat with radiation shield in place and associated radiation network.

Radiation Example 3: Cooled Shield. Calculate the radiant heat loads to the cold optics section if the radiation shield described in the previous example is cooled to 200 K.

Solution and Comments. The radiation heat load from the cooled shield to the optics section is now given by

$$\frac{\sigma(T_3^4 - T_1^4)}{\dfrac{1 - \varepsilon_1}{\varepsilon_1 A_1} + \dfrac{1}{A_1 F_{12}} + \dfrac{1 - \varepsilon_{3,1}}{\varepsilon_{3,1} A_3}} = 3.0 \text{ W} \ . \tag{6.40}$$

This demonstrates the importance of using cooled shields in cryostat design because reducing the shield temperature from 266 to 200 K reduces the heat load from 9.5 to 3.0 W.

6.2.5 Convection Principles

Heat transfer between a moving fluid and its bounding surface is referred to as convection. Consider the flow of a cold fluid past a warmer surface. The rate at which heat is transferred from the warm surface to the fluid depends on such parameters as velocity, temperature, viscosity, specific heat, and thermal conductivity of the fluid and characteristic dimensions of the surface.

This rather complicated convection process is usually quantified through Newton's law of cooling written as

$$q = hA(T_s - T) \ , \tag{6.41}$$

where

q = rate of heat transfer (W)
h = convective heat transfer coefficient (W/m^2 K)
A = area of fluid surface interface (m^2)
T_s = bounding surface temperature (K)
T = fluid temperature (K).

Accurately predicting values of the convective heat transfer coefficient for a given situation is a major challenge. The fluid flow boundary layer parameters are usually arranged in nondimensional groups including the Reynolds number,

$$\mathrm{Re} = \frac{\rho u_\infty L}{\mu} , \tag{6.42}$$

the Prandtl number,

$$\mathrm{Pr} = \frac{C_p \mu}{k} , \tag{6.43}$$

and the Nusselt number,

$$\mathrm{Nu} = \frac{hL}{k} , \tag{6.44}$$

where

ρ = fluid density (kg/m^3)
u_∞ = fluid velocity (m/s)
L = characteristic length (m)
μ = dynamic viscosity (N/m^2)
C_p = constant pressure specific heat (kJ/kg K)
h = convective heat transfer coefficient (W/m^2 K)
k = fluid thermal conductivity (W/m K).

The Nusselt number is typically expressed as some function of the Reynolds number and Prandtl number or

$$\mathrm{Nu} = f(\mathrm{Re},\mathrm{Pr}) . \tag{6.45}$$

A particularly useful correlation for evaluating the convective heat transfer coefficient for turbulent flow of fluids in tubes and channels is

$$\mathrm{Nu} = 0.023\mathrm{Re}^{0.8}\mathrm{Pr}^{0.4} , \tag{6.46}$$

where properties are evaluated at the average of surface and fluid temperatures. For laminar flow, Nu = 3.39 has also proven useful. These correlations have been applied to cryogenic vapors for design purposes with acceptable results. For more detail the reader should consult the literature.

Convection Example 1: Vapor-Cooled Heat Exchanger. We want to convectively remove heat from a cryogenically cooled instrument by flowing helium

vapor through a 1.386- × 0.187-in. (0.035- × 0.0048-m) rectangular channel in the base plate. The following parameters are known:

q = 12 W, imposed heat load
$\dot{m}$ = 0.4 g/s, mass flow rate of helium vapor
A_t = 16.8×10^{-5} m^2, cross-section area of flow channel
P = 3 atm, pressure of fluid
T_i = 8 K, inlet temperature of vapor
T_s = 15K, desired base plate temperature.

What length of flow channel is required?

Solution and Comments. We first check to see what change in temperature of the helium vapor is expected by writing

$$q = \dot{m} C_p (T_e - T_i) \ . \tag{6.47}$$

We evaluate the specific heat of the vapor at an average temperature of (8 K + 15 K)/2 = 11.5 K. From Table 6.26 in Sec. 6.3 we estimate the average specific heat of helium vapor to be 5.7 kJ/kg K.

$$T_e = T_i + \frac{q}{\dot{m} Cp} = 8 + \frac{12}{0.4(5.7)} = 13.3 \text{ K} \ . \tag{6.48}$$

This suggests that the mass flow rate is adequate, because the exit temperature of the vapor is less than 15 K. We then write

$$q = h(PL)\Delta T \ , \tag{6.49}$$

where

q = imposed heat load of 12 W
h = average convective heat transfer coefficient (W/m^2 s), an unknown at this point
P = wetted perimeter of the flow channel
= (1.386 + 0.187) × 2 in. = 0.08 m
L = length of flow channel (m), which is also an unknown at this point
ΔT = the effective temperature difference between the flowing fluid and the plate.

Since the temperature difference between the fluid and plate is changing as indicated in Fig. 6.24, we will evaluate ΔT as the log mean temperature difference (LMTD), defined as the temperature difference at the inlet minus the temperature difference at the exit, all divided by the natural logarithm of the ratio of these two differences:

$$\text{LMTD} = \frac{\Delta T_i - \Delta T_e}{\ln\left(\frac{\Delta T_i}{\Delta T_e}\right)} = \frac{(15 - 8) - (15 - 13)}{\ln\frac{15 - 8}{15 - 13}} = 4 \text{ K} \ . \tag{6.50}$$

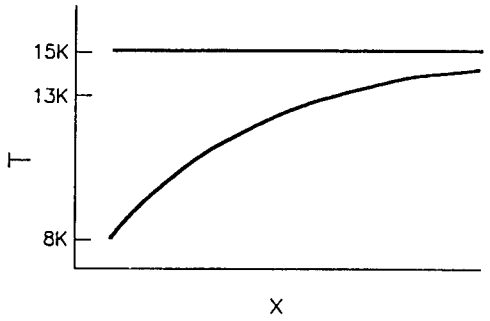


Fig. 6.24 Illustration of temperature variation in a heat exchanger. See convection example 1.

The fluid properties are evaluated at the mean temperature

$$\frac{(8 + 13)/2 + 15}{2} = 13 \text{ K} . \tag{6.51}$$

From Table 6.26 the following values are noted

density = $\rho = 11.7$ kg/m^3
viscosity = $m = 2.8$ μPa s
thermal conductivity = $k = 2.1 \times 10^{-2}$ W/m K
Prandtl number = Pr = 0.77.

We calculate the average velocity of the fluid as

$$u = \frac{\dot{m}}{\rho A} = \frac{0.4 \text{ g/s}}{(11{,}700 \text{ g/m}^3)(16.8 \times 10^{-5} \text{ m}^2)} = 0.203 \frac{\text{m}}{\text{s}} . \tag{6.52}$$

For the rectangular channel, the characteristic diameter D is taken as the hydraulic diameter, defined as four times the cross-sectional area divided by the wetted perimeter:

$$D = \frac{4(16.8 \times 10^{-5} \text{ m}^2)}{0.08 \text{ m}} = 0.0084 \text{ m} . \tag{6.53}$$

The Reynolds number is

$$\text{Re} = \frac{\rho u D}{\mu} = \frac{11.7 \text{ kg/m}^3(0.203 \text{ m/s})(0.0084 \text{ m})}{2.8 \times 10^{-6} \text{ kg/ms}} = 7100 . \tag{6.54}$$

For fluid flow in pipes or ducts it is generally assumed that when $\text{Re} \leq 2300$ the flow is usually laminar. If $\text{Re} \geq 2300$ the flow is usually assumed to be turbulent. Therefore, the flow in this example problem is turbulent, and we utilize

$$\mathrm{Nu} = \frac{hD}{k} = 0.023\ \mathrm{Re}^{0.8}\mathrm{Pr}^{0.4} \tag{6.55}$$

to evaluate the convective heat transfer coefficient:

$$h = 0.023(7100)^{0.8}(0.77)^{0.4}\left(\frac{2.1 \times 10^{-2}}{0.0084}\right)$$

$$= 62.5\ \mathrm{W/m^2\ K}\ . \tag{6.56}$$

From

$$q = hPL\Delta T \tag{6.57}$$

the length of channel required is

$$L = \frac{12\ \mathrm{W}}{(62.5\ \mathrm{W/m\ K})(0.08\ \mathrm{m})4\ \mathrm{K}} = 0.6\ \mathrm{m}\ . \tag{6.58}$$

Thus if the flow channel is more than 0.6 m (23.6 in.) long, the base plate temperature will be less than 15 K. For design purposes a channel length of about 30 in. would be appropriate.

6.2.6 Multilayer Insulation

Multilayer insulation (MLI) consists of a number of very thin highly reflective sheets such as doubly aluminized Mylar separated by a low thermal conductivity spacer material such as Dacron or silk net, all under high vacuum. Heat transfer through MLI involves several mechanisms including solid conduction through points of contact, radiation between reflective sheets, and conduction through the gas trapped between layers and in voids in the spacer materials. While heat transfer through MLI has been well investigated both analytically and with empirical correlations, a completely satisfactory predictive theoretical model has not been achieved due to the unpredictability of contact pressure, contact area, interstitial gas pressure, and the thermal properties of the material.

Examples of correlations that have proven useful for MLI blanket designers are those provided by Bell, Nast, and Wedel[6] who suggest the effective thermal conductivity (μW/m K) for silk net/double-aluminized Mylar is

$$k\ \ [\mu\mathrm{W/m\ K}] = (8.962 \times 10^{-4})(N)^{1.56}\frac{1}{2}(T_H + T_c) + \frac{(5.403 \times 10^{-6})\varepsilon(T_H^{4.67} - T_c^{4.67})}{(T_H - T_c)N}\ , \tag{6.59}$$

where N is the layer density (1/cm), ε is the room temperature emissivity, and T is in degrees Kelvin.

For Tissuglas/double-aluminized Mylar they recommend:

$$k \quad [\mu W/m\ k] = \frac{[(3.07 \times 10^{07})(T_H^2 - T_C^2) - (2.129 \times 10^{-10})(T_H^3 - T_c^3)]N^{2.91}}{(T_H - T_c)} . \quad (6.60)$$

Results from these correlations are shown in Table 6.20 in the flat plate predicted k column. Also shown in Table 6.20 are measured values of effective thermal conductivity for an MLI wrapped tank. In every instance, the measured value is significantly greater than the predicted value. These results suggest the designer should utilize an MLI correlation only as a guide, realizing that the application technique remains a major uncertainty. Several other MLI correlations have been developed and the serious designer is referred to the literature. Values of effective thermal conductivity from other researchers are shown in Table 6.21 for comparison.

In addition to application techniques, effective thermal conductivity depends on both the layer density and the total number of layers in the blanket. Figure 6.25 suggests that a density of 30 layers per centimeter is near optimum. The data in Table 6.21 suggest that increasing the number of layers in the blanket reduces the heat transfer but increases the effective thermal conductivity.

Some researchers have suggested that if the blanket thickness and/or layer density are too low, the phenomenon of *radiation tunneling* may cause an increase in effective thermal conductivity. If blanket thickness and/or layer density are too high, increased gas entrapment or increased contact area and pressure of the spacer material could cause the effective thermal conductivity to increase. The designer should thus be cautious about using blankets that are too "thick" or too "thin" or are compressed at any location.

The interstitial gas pressure may be several orders of magnitude higher than the chamber pressure, causing gas conduction to be a significant contributor to heat flow through MLI. Gas entrapment may be one of the reasons why effective thermal conductivity increases with increased blanket density and thickness. Attempts have been made to load the MLI with absorbent getter materials to reduce the interstitial pressure with mixed results. The heat flux tends to increase due to sorbent-radiation interaction and at the same time tends to decrease because of the residual gas absorption. In most applications the use of sorbents in so-called "self-pumping" MLI is thought to be thermally beneficial.

There is a surprisingly large increase in heat transfer through a MLI blanket in the vicinity of a crack. Heat loads per unit area of crack may be more than 200 times the heat load through a unit area of blanket without cracks. The crack may act as a black cavity absorbing all the incident energy. When installing blankets in a cryogenic system, patches should be used to cover the crack in every layer possible. If retrofitting an already existing blanket, concerns always exist regarding damage to the integrity of the original blanket. Certainly patches are most effective in the warmer side of the blanket where the radiation mechanism tends to dominate and should be applied at least to several outer layers. For a 30-layer blanket, four to six patches are reported to be the minimum necessary to reduce significantly the heat leak through the crack.

Table 6.20 Multilayer Insulation Data (from Ref. 6)

Insulation description	Test No.	Thickness & in.	Total layers	Layer density, number/cm	Boundary temps.,K T_H	T?	Measured effective k μW/m·K	Flat-plate predicted μW/m·K
Tissuglas/Mylar	1	1.25	137	43.3	296	135	51.9	17.3
spiral wrapped	2	1.25	137	43.3	296	152	53.6	17.3
and gored 10	3	1.25	137	43.3	296	54	38.1	12.3
layers at a	4	0.75	82	43.3	295	55	32.9	12.7
time (wrap 1)	5	0.75	82	43.3	296	136	44.9	17.3
Tissuglas/	6	1.0	110	43.3	144	50	157.5	38.1
installed as	7	1.0	110	43.3	296	50	32.9	12.6
above (wrap 2)	8	0.4	15	14.6	296	135	50.2	36.3
Silk net/Mylar	9	0.4	15	14.6	135	50	88.3	65.8
installed one	10	0.4	15	14.6	295	55	36.3	25.9
layer at a time (wrap 3)	11	0.4	15	14.6	133	50	88.3	65.8

Table 6.21 Comparison of Various Combinations of Material for MLI (from Ref. 7)

Materials shield and spacer K	Layer density, N (layers/cm)	Number of layers	Boundary temperatures T_H(K)	T_c(K)	k_{eff} (μW $m^{-1}K^{-1}$)
Aluminized	31.90	60	301.05	79.85	71.5
Mylar and	32.40	40	299.85	79.45	58.6
silk fabric	33.20	20	298.65	78.35	49.0
Aluminum	32.34	60	301.85	80.15	94.9
foil and	33.64	40	300.15	79.35	61.6
silk fabric	33.80	20	299.15	79.05	38.0
Aluminized	30.28	60	301.35	78.15	62.1
Mylar and	28.30	40	303.35	78.33	57.0
glass fabric	28.75	20	301.83	80.63	51.5
Aluminum	28.10	60	302.21	82.15	82.7
foil and	28.45	40	304.62	80.65	73.5
glass fabric	32.35	20	301.83	78.60	48.5

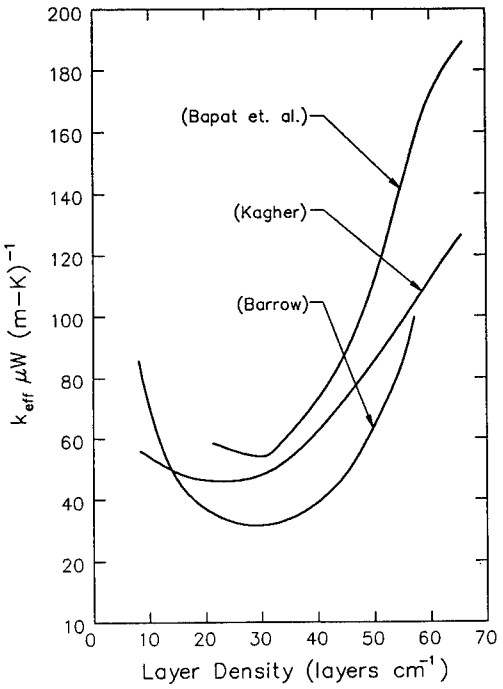


Fig. 6.25 Effective thermal conductivity of MLI as a function of layer density.

In summary, the values of the effective thermal conductivity of well-designed and -applied MLI blankets in cryogenic systems would usually fall in the range of 10 to 100 μW/m K. Heroic measures are required to obtain values at the lower end of the range while blankets performing at the upper end of this range would be considered less than optimum under typical conditions.

6.2.7 Preparing a Heat Map of the Cryostat

To track and manage the heat transfer between the warm vacuum shell and the heat sink, a heat map is a valuable design tool. An example is shown in Fig. 6.26. The heat map shows (1) the temperature levels of major components within the cryostat, (2) heat paths between temperature levels, and (3) heat rates for each path.

6.2.8 Computer Codes

When rough estimates of temperatures and heat flows are not sufficient or when the model becomes substantially complex, the serious cryostat designer

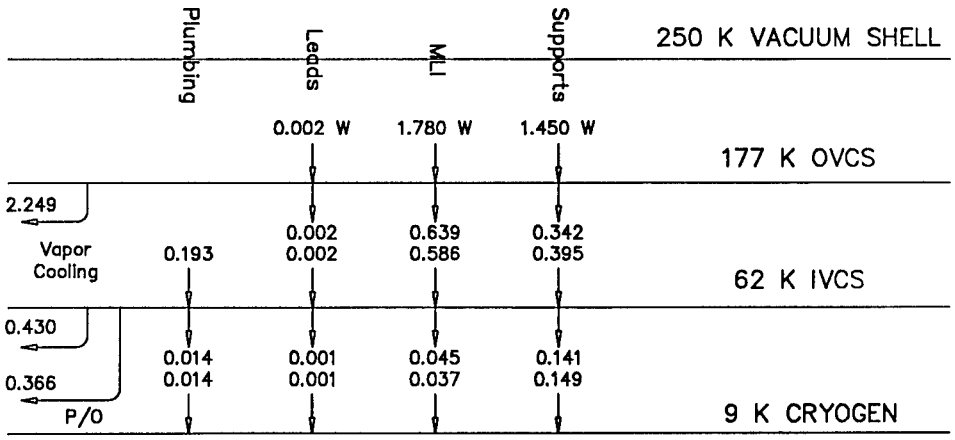


Fig. 6.26 Example of heat map. A total of 3.232 W flows through electrical leads, MLI, and structural supports from the 250 K vacuum shell to have 2.249 intercepted by the outer vapor-cooled shields (OVCS). The plumbing does not interface with the OVCS but does interface with the inner vapor-cooled shield (IVCS) where 0.430 W is absorbed by a sensible temperature increase of the cryogen vapor in a heat exchanger and 0.366 W is absorbed in a para/ortho converter also thermally connected to the IVCS. Part of the heat entering the MLI is transferred to the support structure in thermal contact with the MLI.

can turn to computer codes. Available thermal analysis computer codes allow the designer to obtain either a steady-state or transient solution to very detailed models. Although there are many such computer codes, we have chosen to present briefly the capabilities of only a selected few that are widely used in industry. The reader is encouraged to consult the specific computer code literature for detailed explanations of scope and capabilities.

Keep in mind that even though these computer codes are powerful, the integrity of the results are dependent on the accuracy of the model setup. This accuracy is increased as the designer's knowledge of the system and experience in thermal analysis increases.

SINDA. SINDA, the Systems Improved Numerical Differencing Analyzer, is a software package with capabilities that make it well suited for solving lumped parameter representations of physical problems governed by diffusion-type equations. This package is designed as a general thermal analyzer where the model consists of conductor-capacitor network representations of thermal systems.

The SINDA system consists of two primary sections: the preprocessor and the library. The preprocessor is a program that accepts problems written in SINDA and FORTRAN language. This unique preprocessor accepts "program-like" logic statements and subroutine calls (either in the SINDA library or supplied by the user), in addition to network description cards and other relevant data. This feature allows the user to tailor the program to suit a particular problem. The SINDA library contains many prewritten FORTRAN subroutines that perform a large variety of commonly needed functions for both steady-state and transient solutions. Over the years, various forms of SINDA have emerged, primarily to meet a particular need of a given aerospace

company. A very widely used and available SINDA version is SINDA 1987/ANSI, or Gaski SINDA (named after its developer). This ANSI FORTRAN 77 version of SINDA is available from Network Analysis Associates, Inc. in Fountain Valley, California.

NEVADA. NEVADA, the Net Energy Verification And Determination Analyzer, is an excellent software package for determining radiant energy exchange. Since the radiation heat transfer problem of any cryostat system or spacecraft system in general is usually quite complex, this extremely powerful software package is a very useful tool.

This package consists of four programs, RENO, VEGAS, GRID, and SPLOT, each having a particular function in solving the radiation heat transfer problem.

RENO. The basic RENO (Radiant Energy Network Option) computer code calculates the geometric view factors (F_{ij}'s), often referred to as blackbody radiation factors. The program also calculates the radiant interchange factors (B_{ij}'s), often referred to as graybody radiation factors. This program is used to best advantage in calculating F_{ij}'s and/or B_{ij}'s when the problem geometry is sufficiently complex and where the more conventional (closed form) solutions cannot readily solve the problem with the desired accuracy. Monte Carlo mathematical techniques are used to achieve these solutions.

RENO also provides a means for calculating free molecular flow conductances for complex surface geometries that hitherto could only be estimated. It allows for either diffuse or specular surface properties and allows the presence of an absorbing media between surfaces.

VEGAS. VEGAS, the Verified Earthshine, Geometric Albedo and Solar program, is a powerful tool for computing external radiation heat loads to a set of surfaces. This program is a sister program to RENO and employs the same basic Monte Carlo techniques as RENO. The unique capabilities of VEGAS include direct and reflected solar, direct and reflected albedo, and direct and reflected planetary emission. The user is allowed to specify any orbit position and spacecraft orientation.

The program output includes direct energy absorbed on each surface and total energy absorbed by each surface (directly and/or by reflection from other surfaces).

GRID. GRID, the Gebhart Radiation Interchange Determination program, was developed for the following purposes:

1. to employ the Gebhart mathematical technique to calculate radiant interchange factors (B_{ij}'s) from geometrical view factors (F_{ij}'s);
2. to adjust and finalize the interchange factors (B_{ij}'s) so that:

   $\sum_{j=1}^{n} B_{ij} = 1$;

   then
3. compute the actual nodal heat flows (G_{ij}'s) for incorporation into a thermal analyzer program (TAP)

   or

4. compute the actual nodal heat flows (G_{ij}'s) for incorporation into SINDA

   or

5. convert existing B_{ij}'s to CAL (G_{ij}) cards for input to SINDA.

Thus, GRID is the link between the NEVADA environment and the SINDA environment where the thermal analysis results are obtained.

SPLOT. SPLOT, the Surface Plot program, was developed for use with the RENO and VEGAS computer programs. SPLOT will draw as many orthogonal or perspective views of the input surface descriptions as the user has specified in the view request cards. Hence, this program is a powerful aid to the RENO or the VEGAS user because it ensures that all of the surfaces and overall geometry of a model are defined precisely as the user wants them defined.

The NEVADA software package is available from Turner Associates Consultants, Incline Village, Nevada.

SSPTA. SSPTA, the Simplified Shuttle Payload Thermal Analyzer program, was developed to aid in the evaluation of thermal design concepts of instruments to be flown in the space shuttle cargo bay. Although SSPTA was designed primarily to analyze shuttle payloads, it can easily be used to perform thermal analysis in a variety of other situations. SSPTA consists of a collection of programs used in the thermal analysis of spacecraft and have been modified for quick, preliminary analysis of payloads.

SSPTA was designed to be easy to use, and the user-required input is simple. Thus, the user is free from many of the concerns of computer usage such as disk space handling, tape usage, and complicated program control. The subprograms of SSPTA are all based on programs that have been used extensively in the analysis of orbiting spacecraft and space hardware.

Subprogram CONSHAD uses the user-supplied geometric radiation model to compute blackbody view factors, shadow factors, and a description of the surface model. The subprogram WORKSHEET uses the surface model description, optical property data, and node assignment data to prepare input for SCRIPTF. Subprogram SCRIPTF computes the inverses of the infrared (IR) and ultraviolet (UV) radiation transfer equations; it also computes the radiation coupling between nodes in the thermal model. Subprogram ORBITAL uses the shadow tables to compute incident flux intensities on each surface in the geometric model. Subprogram ABSORB uses these flux intensities combined with the IR and UV inverses to compute the IR and UV fluxes absorbed by each surface. The radiation couplings from SCRIPTF and the absorbed fluxes from ABSORB are used by subprogram TTA to compute the temperature and power balance for each node in the thermal model. Output consists of tabulated data from each of the subprograms executed during a particular analysis. Due to the modular form of SSPTA, analyses may be run in whole or in part, and new subprograms may be added by the user.

SSPTA was written in FORTRAN IV and ASSEMBLER for batch execution; however, an all–FORTRAN version of SSPTA for use on a DEC VAX-11/780 was developed in 1980. The reader should contact Computer Software Management and Information Center, Computing and Information Services, University of Georgia, Athens, Georgia, for further SSPTA details.

6.3 PROVIDING THE LOW-TEMPERATURE HEAT SINK

6.3.1 Cryogens

Every known substance can be made to exist as a solid, liquid, or gas depending on the pressure and temperature to which it is exposed. Those substances that normally exist as a gas at room temperature and pressure are called *cryogens* if liquefied or solidified. Shown in Fig. 6.27 is the pressure-specific volume temperature (P-V-T) equilibrium surface typical of almost all substances. Also shown is the P-T view of the P-V-T surface. If a substance is to exist as a liquid, the temperature must be somewhere between the triple point and the critical point. If a cryogen liquid is boiling in equilibrium with its vapor at a pressure defined by the vaporization curve in Fig. 6.27, it is often referred to as *normal boiling point* (NBP) liquid, such as NBP nitrogen. So-called "supercritical" fluids exist at a pressure greater than the critical pressure and will not boil or change phase at constant temperature when heated but will continuously increase in temperature. All substances except helium have a triple point at which all three phases can exist together in equilibrium at the same temperature and pressure. A solid cryogen existing at a pressure greater than the triple point pressure will melt upon heating, while a solid cryogen existing at a pressure less than the triple point pressure will sublimate upon heating. The temperature at which a solid cryogen sublimates may be reduced by reducing the vapor pressure above it.

The P-T diagram for helium is different from that of all other substances. As shown in Fig. 6.28, helium has a critical point and a vaporization line but no triple point. Saturated liquid helium will boil at temperatures between the critical temperature of 5.2 K and the so-called λ point at about 2.2 K, depending on the pressure. Helium existing at pressures greater than 2.26 atm and temperatures between 5.2 and roughly 10 K is called *supercritical helium*. Liquid helium existing at temperatures below ~2.2 K and to the left of the λ line in Fig. 6.28 is called *superfluid helium* because of certain remarkable properties.

Shown in Table 6.22 are properties of several liquid and vapor cryogens. Fig. 6.29 illustrates critical point, boiling point at 1 atm, triple point, and solid

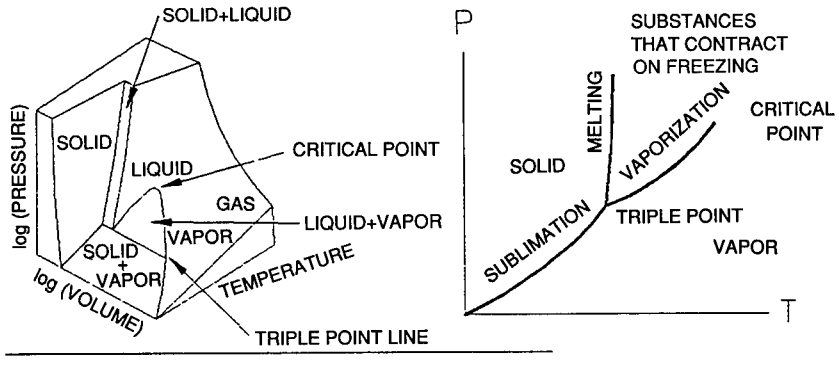


Fig. 6.27 Pressure-volume-temperature equilibrium surfaces.

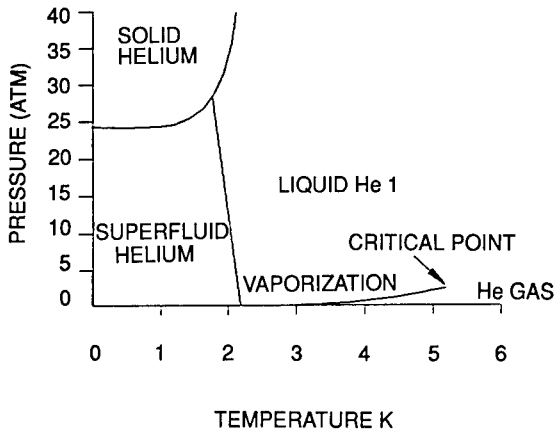


Fig. 6.28 Pressure-temperature diagram for helium.

sublimation temperature at a pressure of 0.1 Torr for a number of cryogens. Table 6.23 provides latent heat of sublimation and density values for solid cryogens.

For illustration purposes a few representative examples of thermal properties of cryogens are shown in Tables 6.24, 6.25, and 6.26. These properties are now largely available on computer disk from the National Institute of Standards and Technology in Boulder, Colorado.

Selecting a Cryogen. The drivers in cryogen selection are usually temperature, mission life, weight, volume, and system cost. If, for example, the focal plane array (FPA) in an infrared instrument must be maintained at less than 9 K, there are only three choices for cryogen: superfluid helium at ~1.6 to 2 K, NBP helium at 2.2 to 5.2 K, or supercritical helium at 5.2 to 8 K. Superfluid helium requires sophisticated fluid handling procedures that may drive up system costs. NBP helium offers simplicity and a relatively stable tank temperature and pressure, but sloshing problems may arise if the cryostat is maneuvered in a gravity or accelerating environment and phase separation is difficult in a microgravity environment. Supercritical helium offers the advantage of being a single-phase fluid that completely fills the cryo tank with no sloshing or liquid interface problem but has the disadvantage of increasing in temperature as the cryogen is expended.

As the required temperature increases, the choice of cryogens increases and the selection process requires consideration of other factors such as cryogen cooling capacity.

Cryogen Example 1: Cooling Capacity of NBP Helium. Determine the cooling capacity of 100 liters of NBP helium at 1 standard atm (101.35 kPa).

Solution. From Table 6.25:

specific volume v_f = 0.007991 m^3/kg
latent heat of vaporization h_{fg} = 20.4 kJ/kg
mass of helium = $m = V/v_f$ = 0.100 m^3/0.007991 m^3 kg
= 12.5 kg.

Table 6.22 Some Cryogens and Properties

Substance		Boiling Point at 1 atm	Critical Point	Critical Pressure	Triple Point	Triple Point Pressure	Latent Heat of Vaporization @ boil pt	Latent Heat of fusion @ triple pt	Specific Volume of sat liquid @ boil pt	Specific Volume of sat vapor @ boil pt	Specific Volume of solid @ triple pt
		K	K	MPa	K	MPa	kJ/kg	kJ/kg	m^3/kg	m^3/kg	m^3/kg
Helium	He^4	4.2	5.2	0.2290	2.18	.005102	20.5		.008000	0.058824	
Hydrogen	H_2	20.39	33.19	1.3152	13.8	.007042	448.0	57.78	.014085	0.777605	.011543
Neon	Ne	27.2	44.5	2.7155	24.54	.04319	87.0	16.04	.000833	0.105263	.000693
Nitrogen	N_2	77.37	126.1	3.3944	63.15	.01254	119.0	25.75	.001238	0.226501	.001056
Carbon Monoxide	CO	81.6	138.8	3.5464	68.13	.0154	213.5	29.85	.001232	0.230044	.000971
Fluorine	F_2	85.24	144.8	5.5729	53.48	.000252	171.5	26.80	.000661	0.171021	.000565
Argon	A	87.4	150.8	4.8636	83.80	.06871	162.7	30.35	.000719	0.198807	.000616
Oxygen	O_2	90.1	154.1	5.0764	54.34	.0001453	212.5	13.82	.000877	0.210526	.000736
Methane	CH_4	111.7	190.5	4.6407	90.68	.01174	581.0	58.41	.002353	0.568182	.001927
Krypton	Kr	120.3	209.3	5.5222	115.95	.0731	108.0	19.5	.000417	0.120048	.000345
Freon-14	CF_4	145.14	227.5	3.7490	-	-	134.8	-	.000617	0.138889	-
Ozone	O	161.3	261.1	5.5323	-	-	316.0	-	.000685	-	-
Xenon	Xe	165.3	290.0	5.8769	161.3	.0816	96.25	23.45	.000323	0.102354	.000282
Ethylene	C_2H_4	169.3	282.8	5.1574	103.99	.00012	481.0	-	.001730	0.480769	-
Nitrous Oxide	N_2O	183.6	399.7	7.2650	-	-	250.2	148.63	.000813	-	-
Ethane	C_2H_6	184.8	305.0	4.9447	89.89	-	490.0	95.10	.001779	3.125000	.001433
Acetylene	C_2H_2	189.1	309.0	6.2822	-	-	916.0	-	.001605	-	-
Freon-13	$CCLF_3$	192.0	302.0	0.3952	-	-	146.4	-	.000664	0.126582	-
Carbon Dioxide	CO_2	194.6	304.5	7.3967	216.54	.5173	574.0	180.87	.000662	-	.000588
Propylene	C_3H_6	226.1	365.0	4.5596	-	-	439.5	-	.001656	-	-
Freon-22	$CHCLF_2$	232.5	369.0	4.9345	-	-	235.0	-	.000707	0.215054	-
Ammonia	NH_3	239.8	405.0	11.2673	195.4	.00618	1363.0	-	.001464	1.113586	.00125

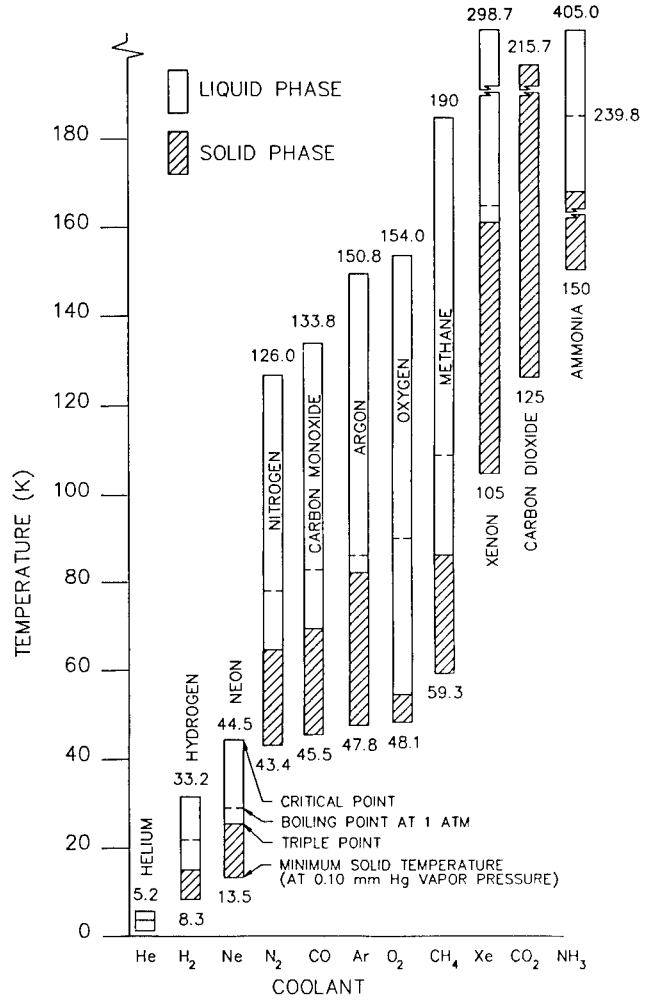


Fig. 6.29 Some cryogens.

Heat required to vaporize 12.5 kg of helium is

$$Q = mh_{fg} = 12.5 \text{ kg } (20.4 \text{ kJ/kg}) = 256 \text{ kJ} \ . \tag{6.61}$$

Cryogen Example 2: Cooling Capacity of Solid Cryogen. We desire to maintain the focal plane assembly of an optical instrument at 10 K for a mission life of 2 yr using solid hydrogen as the cryogen. The total average heat load to the cryogen is estimated to be 750 mW. Determine the required cryogen mass and volume.

Solution and Comment. The latent heat of sublimation of a solid cryogen varies only slightly with pressure. For purposes of this illustration we use the value at the triple point given in Table 6.23, which is 508.6 kJ/kg. The total heat capacity required is

$$0.750 \text{ J/s} \times 2 \text{ yr} \times 365 \text{ days/yr} \times 24 \text{ h/day} \times 3600 \text{ s/h} = 4.73 \times 10^7 \text{ J} \ . \tag{6.62}$$

Table 6.23 Operating Temperature Range (Triple Point to a Pressure of 0.1 Torr = 13 Pa), Latent Heat of Sublimation and Density of Solid Cryogens

Cryogen	Temperature Range (Kelvin)	Latent Heat of Sublimation (kJ/kg)	Density of Solid (kg/m3)	Assuming 100 mW Load (12 months) Mass (kg)	Volume (l)
Helium (liquid)	4.2 - 2.0	21.1	128.5	147.40	1147.00
Hydrogen	13.7 - 8.3	508.6	86.6	6.12	70.60
Neon	24.5 - 13.5	105.6	1490.0	29.46	19.77
Nitrogen	63.1 - 43.4	226.1	1020.0	13.76	13.49
Carbon Monoxide	68.1 - 45.5	295.5	1030.0	10.53	10.22
Argon	83.9 - 47.8	202.3	1670.0	15.38	9.21
Oxygen	54.3 - 48.1	257.0	1359.0	12.11	8.91
Methane	90.7 - 59.8	614.0	520.0	5.07	9.74
Xenon	161.3 - 105	120.5	3540.0	25.81	7.29
Carbon Dioxide	217.5 - 125	566.4	1700.0	5.49	3.23
Ammonia	195.4 - 150	1831.7	800.0	1.70	2.12

The mass and volume of solid cryogen required to sustain a heat load of 100 mW for 12 months are also shown for comparison purposes.

Table 6.24 Properties of Saturated Hydrogen (PARA)

T	P	volume, m^3/kg		enthalpy, kJ/kg			entropy, kJ/(kg K)		
K	MPa	v_f	v_g	h_f	h_{fg}	hg	s_f	s_{fg}	s_g
13.80	0.007042	0.012983	7.952	0.0	447.2	447.2	0.0	32.408	32.408
14	0.007896	0.013011	7.185	1.2	447.9	449.1	0.082	31.994	32.076
15	0.01343	0.013158	4.491	7.4	450.6	458.0	0.508	30.042	30.550
16	0.02153	0.013314	2.960	14.4	452.2	466.6	0.950	28.264	29.214
17	0.03284	0.013481	2.038	21.9	452.7	474.6	1.400	26.628	28.028
18	0.04807	0.013660	1.454	30.1	452.0	482.1	1.856	25.110	26.966
19	0.06796	0.013854	1.068	38.9	450.0	488.9	2.316	23.686	26.002
20	0.09326	0.014065	0.8045	48.3	446.8	495.1	2.781	22.337	25.118
20.28	0.101325	0.014127	0.7466	51.0	445.6	494.6	2.910	21.975	24.885
21	0.1247	0.014296	0.6185	58.4	442.0	500.4	3.251	21.047	24.298
22	0.1632	0.014550	0.4837	69.2	435.7	504.9	3.728	19.801	23.529
23	0.2094	0.014831	0.3837	80.8	427.5	508.3	4.214	18.586	22.800
24	0.2642	0.015147	0.3081	93.3	417.4	510.7	4.709	17.391	22.100
25	0.3284	0.015503	0.2496	106.7	405.1	511.8	5.216	16.202	21.418
26	0.4029	0.015911	0.2038	121.2	390.2	511.4	5.740	15.006	20.746
27	0.4885	0.016386	0.1672	137.0	372.3	509.3	6.284	13.787	21.071
28	0.5861	0.016951	0.1374	154.4	350.7	505.1	6.855	12.525	19.380
29	0.6967	0.017644	0.1129	173.7	324.4	498.1	7.467	11.186	18.653
30	0.8214	0.018532	0.09207	195.8	291.6	487.4	8.140	9.719	17.859
31	0.9615	0.019760	0.07387	222.2	248.6	470.8	8.916	8.020	16.936
32.94	1.284	0.031888	0.03189	346.5	0.0	346.5	12.536	0.0	12.536

Table 6.25 Properties of Saturated Helium-4

T	P	volume, m^3/kg		enthalpy, kJ/kg			entropy, kJ/(kg K)		
K	MPa	v_f	v_g	h_f	h_{fg}	hg	s_f	s_{fg}	s_g
2.18	0.005102	0.006837	0.8376	0.0	22.661	22.661	0.0	10.4091	10.4091
2.2	0.005395	0.006841	0.7988	0.122	22.632	22.754	0.0548	10.2875	10.3423
2.4	0.008439	0.006883	0.5468	0.720	22.824	23.544	0.3074	9.5102	9.8176
2.6	0.01250	0.006937	0.3916	1.183	23.096	24.279	0.4813	8.8831	9.3644
2.8	0.01773	0.007004	0.2904	1.692	23.260	24.952	0.6564	8.3072	8.9636
3	0.02427	0.007084	0.2213	2.229	23.329	25.558	0.8256	7.7766	8.6022
3.2	0.03230	0.007179	0.1723	2.801	23.290	26.091	0.9919	7.2781	8.2700
3.4	0.04196	0.007289	0.1365	3.434	23.109	26.543	1.1623	6.7968	7.9591
3.6	0.05339	0.007419	0.1096	4.142	22.761	26.903	1.3406	6.3226	7.6632
3.8	0.06676	0.007572	0.08892	4.933	22.229	27.162	1.5273	5.8496	7.3769
4	0.08221	0.007753	0.07271	5.811	21.491	27.302	1.7221	5.3726	7.0947
4.2	0.09990	0.007972	0.05970	6.790	20.508	27.298	1.9269	4.8829	6.8098
4.22	0.101325	0.007991	0.05883	6.869	20.422	27.291	1.9448	4.8452	6.7880
4.4	0.1200	0.008246	0.04903	7.903	19.207	27.110	2.1478	4.3652	6.5130
4.6	0.1428	0.008602	0.04001	9.217	17.446	26.663	2.3971	3.7924	6.1895
4.8	0.1684	0.009103	0.03204	10.860	14.932	25.792	2.6982	3.1108	5.8090
5	0.1971	0.009915	0.02419	13.136	10.829	23.965	3.1064	2.1659	5.2723
5.2014	0.227	0.014360	0.01436	18.622	0.0	18.622	4.1124	0.0	4.1124

Table 6.26 Thermophysical Properties of Helium-4 at 3-atm Pressure (from Ref. 8)

TEMP [K]	DENSITY [kg/m³]	PV/RT [-]	ENERGY [J/g]	ENTHALPY [J/g]	ENTROPY [J/g·K]	C_V [J/g·K]	C_P [J/g·K]	VSOUND [m/s]
0.8000	150.0	1.203	0.3299E-01	2.033	0.3717E-02	0.211E-01	0.2113E-01	261.7
1.000	150.0	0.9628	0.4391E-01	2.044	0.1572E-01	0.1056	0.1057	261.3
1.200	150.0	0.8023	0.8463E-01	2.085	0.5218E-01	0.3332	0.3332	260.7
1.400	150.1	0.6875	0.1944	2.194	0.1354	0.8083	0.8090	259.8
1.600	150.2	0.6010	0.4351	2.433	0.2937	1.649	1.653	257.9
1.800	150.5	0.5333	0.8959	2.890	0.5613	3.011	3.028	254.5
2.000	150.9	0.4784	1.720	3.707	0.9898	5.349	5.405	249.2
2.048	151.1	0.4666	2.004	3.989	1.129	6.258	6.335	247.6
2.138	151.5	0.4457	2.708	4.687	1.462	9.981	10.23	243.6
2.147	151.6	0.4437	2.809	4.788	1.509	12.57	13.08	242.8
2.149	151.6	0.4432	2.837	4.815	1.522	7.776	7.875	243.4
2.158	151.6	0.4413	2.889	4.868	1.546	4.881	4.885	244.3
2.248	151.5	0.4240	3.180	5.160	1.679	2.568	2.602	246.5
2.400	151.0	0.3984	3.522	5.509	1.829	2.026	2.119	248.6
2.700	149.7	0.3574	4.117	6.121	2.070	1.884	2.094	248.3
3.000	147.7	0.3258	4.736	6.766	2.296	1.919	2.296	245.2
3.300	145.3	0.3012	5.447	7.512	2.533	2.109	2.682	239.9
3.600	142.3	0.2819	6.269	8.377	2.784	2.273	3.088	232.8
3.900	138.8	0.2668	7.205	9.366	3.047	2.386	3.511	224.5
4.200	134.6	0.2554	8.262	10.49	3.325	2.462	3.999	214.7
4.500	129.6	0.2477	9.465	11.78	3.621	2.517	4.647	202.9
4.800	123.1	0.2444	10.88	13.31	3.951	2.565	5.667	188.2
5.000	117.5	0.2458	12.00	14.56	4.204	2.600	6.854	176.0
5.100	114.1	0.2482	12.66	15.28	4.348	2.621	7.796	168.8
5.300	104.8	0.2600	14.30	17.16	4.709	2.676	11.67	151.3
5.500	86.22	0.3045	17.24	20.72	5.367	2.790	28.12	128.0
6.000	42.21*	0.5703	25.48	32.59	7.446	3.020	13.83	122.9
6.500	32.33	0.6872	28.82	38.10	8.330	3.053	9.284	133.7
7.000	27.27	0.7565	31.33	42.33	8.958	3.067	7.841	143.1
8.000	21.51	0.8394	35.56	49.51	9.918	3.084	6.712	158.9
9.000	18.08	0.8876	39.36	55.95	10.68	3.097	6.236	172.1
10.00	15.72	0.9188	42.96	62.05	11.32	3.106	5.974	183.8
12.00	12.59	0.9557	49.86	73.68	12.38	3.119	5.695	204.1
15.00	9.804	0.9821	59.84	90.44	13.63	3.127	5.501	230.1
20.00	7.228	0.9990	76.04	117.5	15.19	3.130	5.362	266.6

Table 6.26 *(continued)*

PRESSURE = 0.300 [MPa]

TEMP [K]	$\left(\frac{T}{V}\frac{\partial V}{\partial T}\right)_P$	$\left(\frac{V}{C_V}\frac{\partial P}{\partial T}\right)_V$	$\left(\frac{P}{\rho}\frac{\partial \rho}{\partial P}\right)_T$	DIEL - 1	CONDUCT [W/m·K]	VISC [μPa·s]	THDIFF [m²/s]	PRANDTL
0.8000	0.2477E-03	1.004	0.2920E-01	0.5916E-01				
1.000	0.2809E-03	0.1815	0.2930E-01	0.5916E-01				
1.200	-0.7702E-03	-0.1310	0.2942E-01	0.5916E-01				
1.400	-0.3859E-02	-0.2300	0.2964E-01	0.5918E-01				
1.600	-0.1022E-01	-0.2570	0.3010E-01	0.5923E-01				
1.800	-0.2161E-01	-0.2568	0.3097E-01	0.5934E-01				
2.000	-0.4248E-01	-0.2440	0.3234E-01	0.5953E-01				
2.048	-0.5114E-01	-0.2416	0.3279E-01	0.5960E-01				
2.138	-0.9651E-01	-0.2617	0.3422E-01	0.5977E-01				
2.147	-0.1390	-0.2917	0.3494E-01	0.5980E-01				
2.149	-0.6023E-01	-0.2109	0.3382E-01	0.5981E-01				
2.158	-0.1158E-01	-0.6556E-01	0.3381E-01	0.5982E-01				
2.248	0.3615E-01	0.3754	0.3304E-01	0.5977E-01				
2.400	0.6128E-01	0.7448	0.3361E-01	0.5957E-01				
2.700	0.1011	1.103	0.3613E-01	0.5902E-01				
3.000	0.1500	1.309	0.4041E-01	0.5825E-01				
3.300	0.2043	1.329	0.4562E-01	0.5727E-01				
3.600	0.2713	1.323	0.5284E-01	0.5608E-01	0.1916E-01	3.940	0.4359E-07	0.6351
3.900	0.3579	1.317	0.6310E-01	0.5468E-01	0.1973E-01	3.859	0.4048E-07	0.6867
4.200	0.4769	1.309	0.7851E-01	0.5300E-01	0.2011E-01	3.747	0.3736E-07	0.7450
4.500	0.6554	1.291	0.1038	0.5097E-01	0.2030E-01	3.604	0.3373E-07	0.8247
4.800	0.9635	1.255	0.1519	0.4839E-01	0.2032E-01	3.422	0.2912E-07	0.9547
5.000	1.345	1.216	0.2172	0.4617E-01	0.2025E-01	3.271	0.2513E-07	1.107
5.100	1.660	1.190	0.2745	0.4481E-01	0.2020E-01	3.180	0.2270E-07	1.228
5.300	3.015	1.115	0.5456	0.4112E-01	0.2005E-01	2.947	0.1639E-07	1.715
5.500	9.258	0.9809	2.140	0.3375E-01	0.1962E-01	2.545	0.8090E-08	3.649
6.000	4.438	0.8069	2.157	0.1643E-01	0.1516E-01	1.970	0.2596E-07	1.798
6.500	2.625	0.7775	1.578	0.1257E-01	0.1459E-01	1.961	0.4862E-07	1.247
7.000	2.042	0.7624	1.373	0.1060E-01	0.1481E-01	2.005	0.6925E-07	1.062
8.000	1.582	0.7438	1.202	0.8354E-02	0.1576E-01	2.133	0.1092E-06	0.9086
9.000	1.386	0.7315	1.128	0.7019E-02	0.1681E-01	2.275	0.1491E-06	0.8440
10.00	1.278	0.7223	1.087	0.6101E-02	0.1783E-01	2.418	0.1899E-06	0.8098
12.00	1.164	0.7095	1.044	0.4886E-02	0.1979E-01	2.697	0.2760E-06	0.7760
15.00	1.088	0.6979	1.017	0.3803E-02	0.2252E-01	3.091	0.4175E-06	0.7550
20.00	1.037	0.6875	1.000	0.2803E-02	0.2666E-01	3.682	0.6877E-06	0.7407

The calculated mass of cryogen required is

$$\frac{4.73 \times 10^4 \text{ kJ}}{508.6 \text{ kJ/kg}} = 93 \text{ kg} \ . \tag{6.63}$$

Using a density of 86.6 kg/m^3 from Table 6.23, the volume required is

$$\frac{93 \text{ kg}}{86.6 \text{ kg/m}^3} = 1.074 \text{ m}^3 = 1074 \text{ l} \ . \tag{6.64}$$

Alternatively, we could have utilized the last two columns in Table 6.23 directly:

$$\text{mass of cryogen} = \frac{750}{100} \times \frac{24}{12} \times 6.12 = 92 \text{ kg} \ , \tag{6.65}$$

$$\text{volume of cryogen} = \frac{750}{100} \times \frac{24}{12} \times 70.60 = 1060 \text{ l} \ . \tag{6.66}$$

We could also have used the data from Table 6.22 by noting that the latent heat of sublimation is closely approximated by the sum of the latent heat of fusion and the latent heat of vaporization or

$$h_{if} + h_{fg} \times h_{ig} \ . \tag{6.67}$$

From Table 6.22, 58 + 448 = 506 kJ/kg. The density of the hydrogen ice is calculated as the reciprocal of the specific volume found in the last column of Table 6.22:

$$\begin{aligned} \frac{1}{\nu \text{ m}^3/\text{kg}} &= \rho \text{ kg/m}^3 \ , \\ \frac{1}{0.011543 \text{ m}^3/\text{kg}} &= 86.6 \text{ kg/m}^3 \ . \end{aligned} \tag{6.68}$$

Solid cryogen coolers have distinct advantages over liquid cryogens in terms of cooling capacity per unit mass. For example, the cooling capacity of 100 l or 8 kg of solid hydrogen is approximately 16 times that of 100 l or 12.5 kg of NBP helium.

Thermoacoustic Oscillation. Every designer of cryostats should be aware of the thermoacoustic oscillation (TAO) phenomenon. Under certain conditions where fill and vent lines, valved off on the warm end, run between a warm vacuum shell and a cold cryogen, acoustic waves may travel back and forth along the length of the tubes transferring heat from the warm to the cold end and rapidly dissipating the cryogen. Conditions under which this occurs include a temperature ratio of

$$T_h/T_c \geq 6 \ , \tag{6.69}$$

where T_h is the temperature in degrees Kelvin of the warm end and T_c, also

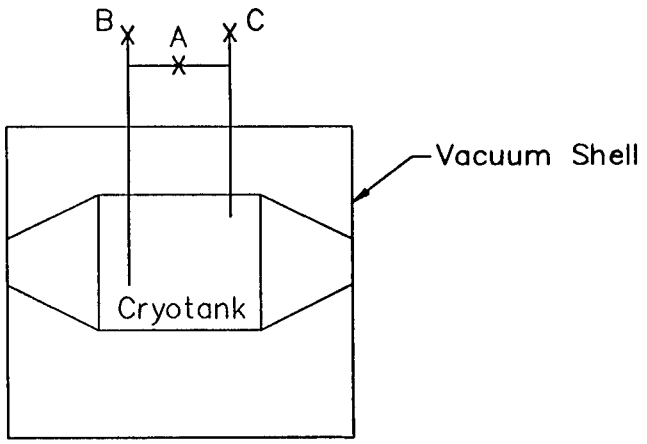


Fig. 6.30 Illustration of the cross link method of preventing thermoacoustic oscillation.

in degrees Kelvin, is the temperature of the cold end. For a room temperature vacuum shell where $T_h \approx 300$ K, T_c must be less than about 50 K for the TAO phenomenon to occur. Therefore, we would not expect TAO in a nitrogen cryostat at 77 K, but could expect it in a helium cryostat at 4.2 K, particularly one having long slender fill and/or vent lines.

The designer of cryostats with temperature ratios of $T_h/T_c \geq 6$ would be well advised to take positive steps to eliminate TAO. A simple but effective method for controlling TAO is illustrated schematically in Fig. 6.30. A cross link with a valve, A, is installed on the outside of the vacuum shell between two lines that run to the cryotank. After valves B and C are closed, valve A is opened slightly, which then allows any oscillating wave energy to dissipate rather than being reflected back down the tube. If the cross link method is not practical, another approach is to extend the length of the "warm part" of the tube by wrapping it partway around the vacuum shell before valving it off. Still another approach is to include a Helmholtz resonator or a chamber on the warm end of the line.

6.3.2 Low-Temperature Radiators

Radiating surfaces with a view to deep space have been designed to provide heat sink temperatures below 60 K for the cooling of instrument components. Such designs require that the radiating surface be thermally isolated from the platform on which it is mounted and shielded from sun, earth, and albedo.

Final designs are facilitated by computer codes with spacecraft orientation, orbit parameters, and surface geometries and properties as inputs. Presented here are some concepts and guidelines to give the designer a feeling for what is possible.

A simple energy balance on the radiating surface, as shown schematically in Fig. 6.31, gives

$$Q_{\text{net}} + Q_{\text{parasitic}} + \sum \alpha Q_{\text{incident}} = Q_{\text{total}} = \varepsilon A \sigma T^4 \ , \tag{6.70}$$

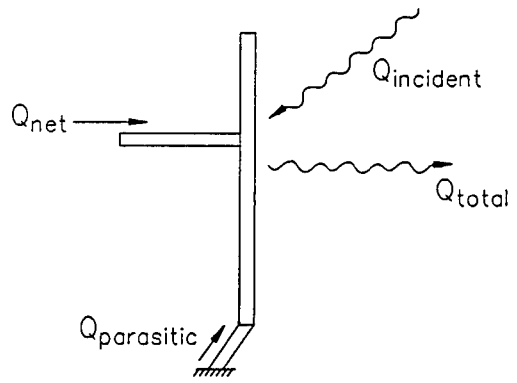


Fig. 6.31 Schematic showing heat loads to and from radiator surface.

where

Q_{net} = heat rate (W) from instrument component to be cooled by radiator
$Q_{incident}$ = radiation (W) from outside sources that is incident on the radiator
$Q_{parasitic}$ = heat (W) through supports, etc.
α = radiator absorptivity of the incident radiation
ε = effective surface emissivity
A = radiator area (m^2)
σ = Stefan-Boltzmann constant, 5.67×10^{-8} W/(m^2 K^4)
T = radiator temperature (K).

The radiator temperature can be reduced by increasing radiator surface emissivity and/or area and/or reducing the heat load, absorptivity, and incident radiation. These are the basic tasks of the designer seeking to provide radiative cooling at a given temperature level.

On a spacecraft, the area of the radiator is often limited by weight considerations and by competition from electronics boxes and other components seeking an uncluttered view to space. The upper limit on emissivity is unity for a perfectly "black" surface.

The net heat per unit area for an ideal surface radiating to deep space is

$$\left(\frac{Q}{A}\right)_{ideal} \approx \sigma T^4 \ . \tag{6.71}$$

For the actual radiator, the effective emissivity will be less than unity and the radiator will have to dissipate parasitic heat in addition to the heat imposed by the cooled instrument component:

$$\frac{Q_{total}}{A} = \varepsilon \sigma T^4 \ . \tag{6.72}$$

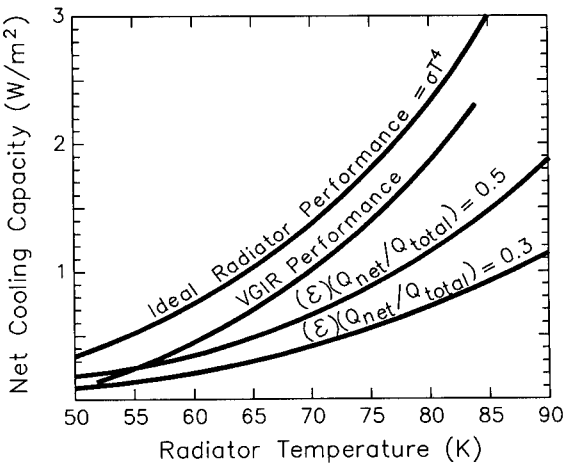


Fig. 6.32 Net cooling capacity of low-temperature space radiators. VGIR data from Ref. 9. See text for explanation of $(\varepsilon)(Q_{net}/Q_{total})$ curves.

Rewriting this as

$$\frac{Q_{total}}{A} = \frac{Q_{total}}{Q_{net}}\frac{Q_{net}}{A} = \varepsilon\sigma T^4 \ , \tag{6.73}$$

$$\frac{Q_{net}}{A} = \varepsilon\frac{Q_{net}}{Q_{total}}\sigma T^4 = \varepsilon\frac{Q_{net}}{Q_{total}}\left(\frac{Q}{A}\right)_{ideal} \ . \tag{6.74}$$

Innovative designs are producing small, lightweight radiators that are pushing the theoretical limit in performance.

Figure 6.32 compares performance of a state-of-the-art V-groove isolation radiator (VGIR) with the ideal radiator with no parasitic heat loads and no incident radiation, and also curves for values of $\varepsilon Q_{net}/Q_{total} = 0.5$ and 0.3. A comparison of the VGIR performance curve with the $\varepsilon Q_{net}/Q_{total}$ curves illustrates the point that as radiator temperatures decrease, the ratio of Q_{net}/Q_{total} will get smaller as parasitic and incident radiation increases proportionally to the net heat dissipated.

In high performance radiators, such as the VGIR, effective thermal isolation of the radiating surface from the warm spacecraft, and sun- and earthshine is accomplished by utilizing lightweight multiple radiation shields and low-thermal-conductance structural supports.

Radiator Example. Estimate the area of radiator required to dissipate 100 mW from an infrared detector operating at 65 K.

Solution and Comments. Depending on instrument configuration, a temperature difference is required between the detector and the radiator surface. A radiator temperature of 60 K will be the design goal to provide a temperature difference between detector and radiator for the heat to flow through a reasonably sized thermal link. It would also be appropriate to assume parasitic

and incident heat loads from supports, sunshades, etc., of about 50% of the net heat. Even though the radiator will probably be painted with a product having an initial emissivity greater than 0.9, it is expected that the surface will degrade over time. Therefore, for purposes of this estimate we assume a surface emissivity of 0.75. We will use Fig. 6.32 to make this estimate:

$$(\varepsilon)\left(\frac{Q_{\text{net}}}{Q_{\text{total}}}\right) = 0.75\left(\frac{1}{1.5}\right) = 0.5 \ . \tag{6.75}$$

From Fig. 6.32 at 60 K and $(\varepsilon)Q_{\text{net}}/Q_{\text{total}} = 0.5$, read

$$Q_{\text{net}}/A \sim 0.37 \text{ W/m}^2 \tag{6.76}$$

$$A = 0.100/0.37 = 0.27 \text{ m}^2 \ . \tag{6.77}$$

Alternatively, we could calculate it using

$$\frac{Q_{\text{net}}}{Q} = \varepsilon \frac{Q_{\text{net}}}{Q_{\text{total}}} \sigma T^4$$

$$= 0.75\left(\frac{1}{1.5}\right)5.67(0.60)^4 = 0.367 \ . \tag{6.78}$$

6.3.3 Cryogenic Refrigerators

The use of cryogenic refrigerators to provide the low-temperature heat sink in cryostats as the advantage over expendable cryogens of potentially long-term service without replenishing.

Operating temperature, cooling capacity, power requirements, mass, reliability, lifetime, and vibration control are important issues in refrigerator selection, particularly for applications in space-deployed surveillance systems.

The refrigerator coefficient of performance (β) is defined as

$$\beta = \frac{Q_L}{W} \ , \tag{6.79}$$

where Q_L is the heat transferred from the refrigerated space to the cold junction of the refrigerator in watts and W is the power supplied to the refrigerator in watts. Refrigerator efficiency is sometimes expressed in terms of the specific power S_p, defined as

$$S_p = \frac{1}{\beta} \ . \tag{6.80}$$

Refrigerator specific mass S_m, defined as the ratio of the mass of a refrigerator to the refrigeration produced, is expressed as

$$S_m = \frac{m_I}{Q_L} \ , \tag{6.81}$$

where m_I is the mass of the refrigerator in pounds or kilograms.

Cryocooler technology continues to advance as mass, power input, vibration levels, and cold tip temperatures are reduced, while increasing expected operating life and reliability. The Air Force Space Technology Center (AFSTC) is supporting the development of a standard spacecraft cryocooler (SSC) designed for a 10-yr lifetime with a reliability greater than 0.95%, providing 2 W of cooling at 65 K with a total input power of less than 80 W and a total mass under 15 kg. Cool down time is to be 5 min or less. So-called "tactical" cryocoolers are generally produced in larger quantities, at much lower cost, with less precision and lower reliability and expected life than are cryocoolers intended for use in space. Tactical cryocoolers of various kinds are used extensively in weapons guidance systems.

Shown in Table 6.27 is a summary of cryogenic refrigerators, together with performance parameters. Several of these refrigerators are further described in the following subsections.

Stirling Cycle Coolers (SCC). The British Aerospace mechanical cryogenic cooler, developed by Oxford University and Rutherford Appleton Laboratory, is a small split-cycle refrigerator with three main components: compressor, displacer, and electronics. The compressor is 200 mm in length with a diameter of 120 mm and a mass of 3.0 kg. The displacer has a length of 190 mm, a diameter of 75 mm, and a mass of 0.9 kg. The electronics section is the largest and heaviest with rectangular dimensions of 225 × 230 × 150 mm and a weight of 4.5 kg.

The compressor and displacer are driven by loudspeaker-type linear motors. The device uses spiral arm springs with high radial stiffness, enabling clearance seals to be used. The absence of rubbing parts makes the device capable of a very long life, with a design goal of 3 to 5 yr. Life testing commenced in January 1986 and after 14,220 hours, there was no degradation of performance.

This cooler can supply 0.8 W of cooling at 80 K with a power consumption of 35 W. Cool down time[10] to 80 K is less than 5 min. In 1992 the cooler was flying on the Improved Stratospheric and Mesospheric Sounder (ISAMS) experiment on the NASA Upper Atmosphere Research Satellite, and on the European Remote Sensing (ERS-1) satellite. Stirling cycle coolers based on the Rutherford Appleton technology are now being built in the United States by TRW, Hughes, Lockheed, and Ball Aerospace. TRW has developed and tested a miniature SCC capable of 250 mW of cooling at 65 K with a power input of 18 W.

Creare, Inc., is also developing a 65 K SSC under contract for the AFSTC. The cooler is a double-acting, flexure-bearing, split Stirling cycle design with linear electromagnetic drives for the expander and compressors. Pistons are replaced by flat metal diaphragms for both sweeping and sealing the working volumes. This cooler is designed to provide 2 W of cooling at 65 K with a target specific power of 30 W/W and a target mass of 14 kg including electronics. This results in an input power of 60 W and a specific mass of 7 kg/W. The cooler is also designed to have a 10-yr lifetime with a 95% reliability. Two test unit prototypes were being built and tested in 1992.

The Ricor micro IDCA cryocooler model K506B SCC, manufactured by Ricor Limited of Israel, can deliver 0.37 W of cooling at 85 K, 0.59 W at 110 K, and 1.18 W at 150 K with an input power of 12 W. Ricor is a major supplier of cryocoolers to the Israeli Defense Forces.

Table 6.27 Types of Cryogenic Refrigerators and Summary of Performance Parameters

Refrigerator ID	Type	T (K)	Heat Lift (W)	Power In (W)	Mass (kg)	Specific Power (W/W)	Specific Mass (kg/W)	Lifetime
ABG Semca	Stirling	80	0.25	--	--	--	--	--
Aerojet Electro-Systems								
C/CH_4	Sorption	130	4	--	--	--	--	5-10 yr*
	Sorption	137	1.97	151	--	76.6	--	500+ hr
advanced	Sorption	25*	1*	195*	--	195	--	--
A. D. Little	Brayton	60	40	2670	210	66.75	5.25	--
		12	1.5	2670		1780		
3-Stage PFC								
1st Stage	Brayton	70*	80.0*	--	--	--	--	50,000 hr*
2nd Stage	Brayton	25*	9.0*	--	--	--	--	same
3rd Stage	Brayton	8.5*	2.5*	--	--	--	--	same
Air Products								
CS 308	mod Solvay + J-T	4.2	1.7	9000	344	5294.1	202.4	--
CS 208L	mod Solvay	20	12	6300	280	525	23.33	--
CS 208R	mod Solvay	20	8	6300	280	787.5	35	--
CS 204SL	mod Solvay	10	8	3200	100	400	12.5	--
CS 204	mod Solvay	20	4	3200	105	800	26.25	--
CS 202	mod Solvay	20	2.25	1700	75	755.6	33.3	--
CS 201	mod Solvay	20	0.6	1700	71	2833.3	118.3	--
CS 108	mod Solvay	77	100	6300	316	63	3.16	--
CS 104	mod Solvay	77	60	3200	103	53.3	1.72	--
CS 102	mod Solvay	77	30	1700	75	56.7	2.5	--
CS 308L	mod Solvay + J-T	4.2	1	9000	350	9000	350	--
CS 304	mod Solvay + J-T	4.2	0.5	4800	170	9600	340	--
CS 302	mod Solvay + J-T	4.2	0.25	--	127	--	508	--
AiResearch								
851310	J-T	80	2.5	500	14.3	200	5.72	--
IR TECH/ASD	Brayton	20	20	4000	136	200	6.8	--
	Brayton	20	5	2200	90.3	440	18.1	--
	VM	80	0.25	--	--	--	--	--

Table 6.27 *(continued)*

Refrigerator ID	Type	T (K)	Heat Lift (W)	Power In (W)	Mass (kg)	Specific Power (W/W)	Specific Mass (kg/W)	Lifetime
Aisin	Stirling	20	15	6600	240	440	16	--
		100	150	6600		44		
Astronautics Corp. of America	M-C	5	0.4	--	--	--	--	--
	AMR	1.8	0.1	--	--	--	--	--
		4.7						
	Pulse-Tube	80*	1*	--	--	--	--	--
Ball Aerospace								
Single-stage	Stirling	80	0.8	60	12	75	15	10 yr*
		80	3	60	12	20	4	
Two-stage	Stirling	30	0.3	75	15	250	50	10 yr*
J-T	J-T	65	0.65	100	--	153.8	--	>5 yr*
		65	3.5	210		60		
Blazers High Vacuum Corp.	G-M	12	5	--	--	--	--	--
		65	80					
British Aerospace/ Oxford University	Stirling	80	0.8	35	8.4	43.75	10.5	3-5 yr*
CEN Grenoble	ADR	2.1	0.92	--	--	--	--	--
Centre d'Estudes Nucléaires, France	VM	50	1*	250	--	250	--	50,000 hr*
		190	5*	250		50		
Centre of Advanced Technology, India	G-M	30	2.6+	--	--	--	--	--
		30	2.4++					
Creare								
Baseline	Brayton	65	5*	200	50*	40*	10	10 yr*
Advanced	Brayton	65	5*	134*	20*	26.8	4	10 yr*
65 K SSC	Stirling	65	2*	60	14*	30*	7	10 yr*
Cryogenic Consultants Ltd.								
R400	G-M+J-T	4.2	0.75	5000	125	6666.7	166.7	--
R700	G-M+J-T	4.2	1.5	8500	242	5666.7	161.3	--

(continued)

Table 6.27 *(continued)*

Refrigerator ID	Type	T (K)	Heat Lift (W)	Power In (W)	Mass (kg)	Specific Power (W/W)	Specific Mass (kg/W)	Lifetime
Cryogenic Laboratory Beijing, China	Pulse-Tube	62	0	188	40.4	--	--	--
		77	2.5	188	40.4	75.2	16.16	
	Pulse-Tube	49	0	--	--	--	--	--
		65	6.6					
		77	12					
Cryomech								
GB07	G-M	16	15	8500	242	566.7	16.1	--
GB04	G-M	12	4	5000	125	1250	31.25	--
		20	9	5000		555.6		
GB03	G-M	20	5	5000	125	1000	25	--
AL05	G-M	40	50	5000	125	100	2.5	--
		80	120	5000		41.7		
AL03	G-M	30	2	700	21.3	350	10.65	--
		80	9	700		77.8		
AL01	G-M	80	20	1200	36	60	1.8	--
		27	3	1200		400		
Cryosystems								
21	G-M	10	0.25	1500	55.5	6000	222	--
LTS 1020	G-M+J-T	4.5	3	8000	--	2666.7	--	--
		77	15	8000		533.3		
CTI-Cryogenics								
CM-1	Stirling	77	0.9	90	1.47	100	1.63	--
CM-2	Stirling	80	1	50	1.72	50	1.72	--
CM-3	Stirling	77	3.5	180	2.95	51	0.84	--
CM-4	Stirling	80	1	60	2	60	2	--
CM-5	Stirling	80	0.3	30	1.13	100	3.77	--
21	G-M	15	0.3	1500	6.5	5000	21.67	--
		77	15	1500		100		
22	G-M	15	0.2	1500	6.5	7500	32.5	--
		77	7.5	1500		200		
350	G-M	15	1.5	1500	15	1000	10	--
		77	19	1500		78.9		
1020	G-M	15	0.6	5000	15	8333.3	25	--
		77	37	5000		135.1		
1050	G-M	15	1.7	5000	15	2941.2	8.82	--
		77	62	5000		80.6		
SP 77A	Stirling	80	0.8	140	2.5	175	3.13	--

Table 6.27 *(continued)*

Refrigerator ID	Type	T (K)	Heat Lift (W)	Power In (W)	Mass (kg)	Specific Power (W/W)	Specific Mass (kg/W)	Lifetime
VM-1	VM	80	0.8	370	4.8	462.5	6	--
120	G-M	80	8	830	15.8	103.8	1.98	--
		26	1	830		830		
Design Study(ONR)	Stirling	10	0.05	250	--	5000	--	--
CVI Inc. TM	G-M	77	75	5000	--	66.7	--	--
		20	8	5000		625		
David Taylor Research Center	M-C	8.5*	1*	--	--	--	--	10 yr*
Galileo Corp.	Stirling	80	0.25	30	1.5	120	6	--
Garrett-AiResearch								
2-stage	Brayton	90	20	5000	177	250	8.85	--
		20						
General Pneumatics								
COBRA	Claude	80	0.25	--	--	--	--	--
anit-clogging	J-T	90	0.74###	--	--	--	--	--
			1.96###					
			2.6###					
Goddard Space Flight Center	ADR	0.065	--	--	--	--	--	--
Hitachi								
II	Claude	4.5	30	--	165	--	5.5	--
III	Claude	4.5	5	--	45	--	9	--
H. R. Textron	Stirling	80	0.25	--	--	--	--	--
Hughes								
Hi Cap	VM	75	12*	2700	--	225	--	20,000 hr*
		33	10*	2700		270		
		11.5	0.3*	2700		9000		
SIRE	VM	75	8.3	--	--	--	--	--
		26.5	1.9					
		11.5	0.15					
77 Ms-1A###	Stirling	77	2	200	6.58	100	3.29	--

(continued)

Table 6.27 *(continued)*

Refrigerator ID	Type	T (K)	Heat Lift (W)	Power In (W)	Mass (kg)	Specific Power (W/W)	Specific Mass (kg/W)	Lifetime
		20	1					
IR&D	Stirling	65	0.373	40.6	12.3[#]	108.9	33.0	--
65 K SSC	Stirling	65[*]	2[*]	72	13	36[*]	6.5	10 yr[*]
7004H[#]	Stirling	65	1	50	1.81	50	1.81	2000 hr
7022H[#]	Stirling	65	0.75	50	1.59	66.7	2.12	2500 hr
7044H[#]	Stirling	65	1.5	75	1.81	50	1.21	2000 hr
7050H[#]	Stirling	65	0.5	50	1.59	100	3.18	4000 hr
SADA[#]	Stirling	72	1.75	60	2.27	34.3	1.3	4000 hr
Indian Institute of Technology	Solvay	8.5	0	--	--	--	--	--
		15.5	1					
International Cryogenic Enterprises (ICE)								
MC111C-011	Stirling	77	0.15	5.4	0.43	36	2.87	--
		80	0.15	4.8		32		
		85	0.15	4.3		28.7		
		140	0.13	3.9	0.43	30	3.31	
		160	0.13	3.4		26.2		
		180	0.13	2.7		20.8		
		190	0.13	2.4		18.5		
Japanses National Railways	Claude	4.4	30	--	140	--	4.67	--
	Claude	4.4	5	--	40	--	8	--
	Stirling+ J-T	4.4	8	--	33.9	--	4.24	--
Jet Propulsion Lab								
C/CH_4	Sorption	140	2	160	--	80	--	--
CH_4+PCO	Sorption	70	1.7	306	--	180	--	--
CH_4+PCO+LH_2	Sorption	14	1.5	645	--	430	--	--
C/Xe+C/Kr+PCO	Sorption	65	1	56	--	56	--	--
gas-gap heat switch	Sorption	65	0.5	--	--	--	--	--
fluid-loop regenerative	Sorption	65	1	--	--	--	--	--
C/Xe+C/Kr+PCO + Liquid Hydride	Sorption	14	--	--	--	306	--	--

Table 6.27 *(continued)*

Refrigerator ID	Type	T (K)	Heat Lift (W)	Power In (W)	Mass (kg)	Specific Power (W/W)	Specific Mass (kg/W)	Lifetime
C/Xe+C/Kr+PCO+LH_2 + Solid Hydride	Sorption	10	--	--	--	706	--	--
$LaNi_5$ hydride	Sorption	27	0.17	75	--	441	--	>1000 hr
DSN maser cooler	G-M+J-T	2.5	0.7	--	--	--	--	--
		4.3	3					
Laboratoire de Radioastronomie, Paris, France	Pulse-Tube	57	0	--	--	--	--	--
		72	12					
	Pulse-Tube	85	0	--	--	--	--	--
		100	4					
L'Air Liquide								
RC 400	open J-T	77	2	--	--	--	--	--
RCF 30-4	G-M	80	30	1500	99	50	3.3	--
		20	4	1500		375		
RH 820	Stirling	80	1	--	--	--	--	--
RH 520	Stirling	80	0.5	--	--	--	--	--
Leybold-Heraeus, Cologne, Germany	G-M	80	10	--	--	--	--	8000 hr*
		20	2					
		14	0					
RG580/RW3	G-M	20	3.75	1800	64	480	17.1	--
		80	37.5	1800		48		
RG580/RW6	G-M	20	6.3	4000	160	634.9	25.4	--
		80	100	4000		40		
RG1040	G-M	20	12.5	4000	160	320	12.8	--
		80	43	4000		93.0		
RG210/RW3	G-M	20	2.5	1800	64	720	25.6	--
		80	15	1800		120		
Los Alamos National Laboratory	ADR	4.3	0.66	--	--	--	--	--
Lucas Aerospace/ Lockheed Missiles and Space Company								
CCS 4000	Stirling	65	4	--	--	--	--	90,000 hr*
CCS 1000+++	Stirling	65	2	100	13	50	6.5	same
CCS 500+++	Stirling	80	0.8	60	13	75	16.25	same
CCS 250	Stirling	65	0.25	--	--	--	--	same

(continued)

Table 6.27 *(continued)*

Refrigerator ID	Type	T (K)	Heat Lift (W)	Power In (W)	Mass (kg)	Specific Power (W/W)	Specific Mass (kg/W)	Lifetime
Magnavox								
MX045L#	Stirling	85	0.2	30	1.13	150	5.65	2500 hr
		85	0.35	35	1.13	100	3.23	same
MX7011	Stirling	80	1	--	--	--	--	--
MX7040	Stirling	80	0.25	30	1.2	120	4.8	--
MX7043	Stirling	77	0.88	55	2.09	62.5	2.38	2500 hr
		80	1	55		55		
MX7045	Stirling	85	0.25	25	1.1	100	4.4	--
MX7048#	Stirling	80	0.4	65	0.998	162.5	2.5	2500 hr
MX7049#	Stirling	80	1.75	100	2.72	57.1	1.55	2500 hr
MX7051#	Stirling	80	0.75	60	1.36	80	1.81	2500 hr
MX7058#	Stirling	80	1	60	1.81	60	1.81	4000 hr
HD1033C/UA	Stirling	80	1	50	1.7	50	1.7	--
Magnetic Bearing	Stirling	65	5	160	83.9	32	16.78	3-5 yr*
Mechanical Technology Incorporated (MTI)	Stirling	80	0.25	3	--	12	--	--
MIT Cryogenic Engineering Lab	SVC	40	1	--	--	--	--	--
Mitsubishi								
Three-Stage	G-M	3.3	0	2500	--	--	--	--
		4.2	0.02	2500		125000		
		5.2	0.05	2500		50000		
Two-Stage	G-M+J-T	4.4	5	8000	--	1600	--	--
Three-Stage	Stirling + J-T	4.4	5	5000	--	1000	--	--
Mitsubishi/Electro-technical Lab	VM	80	1.8	210	6	117	3.33	--
MMR Technologies	J-T	70	13.6	--	--	--	--	--
K7701,K7702,K770T	J-T	77	0.25	--	1	--	4	--
R101	J-T	90	3	--	0.2	--	0.07	--
NASDA of Japan/Hitachi								
BBM1	Stirling	80	2.5	--	--	40*	4*	8000 hr*
NBS gap regenerator	Stirling	6.9	0.001	--	--	--	--	--

Table 6.27 *(continued)*

Refrigerator ID	Type	T (K)	Heat Lift (W)	Power In (W)	Mass (kg)	Specific Power (W/W)	Specific Mass (kg/W)	Lifetime
NIST								
Lead Regenerator	VM	15	1.08	--	--	--	--	--
Brass Regenerator	VM	15	0.34	--	--	--	--	--
2nd stage	Pulse-Tube	44	0	--	--	--	--	--
NIST/Los Alamos National Laboratory	TA + Pulse-Tube	90	0	3000	--	--	--	5-10yr*
		120	5	3000		600		
Osaka Oxygen cryomini D	mod Solvay	20	2.5	2400	84	960	33.6	--
Philips								
Rhombic-Drive	Stirling	90	0.3	30	7.2	100	24	>12,300 hr
		140	1.5	30		20		
UA 7044/00	Stirling	80	1	55	1.8	55	1.8	--
UA 7011	Stirling	80	1	--	--	--	--	--
UA 7039/00	Stirling	80	0.25	30	1.5	120	6	--
	Stirling	64.6	5	200	--	40	--	--
triple expansion	Stirling	10*	0.05*	250*	25*	5000	500	--
Quantum Technology Corp. 100	open J-T	4.2	0.002	--	--	--	--	--
Ricor, Israel								
K405	VM	80	1	150	3.8	150	3.8	--
K413G	Stirling	80	0.4	40	3.8	100	9.5	--
K505	Stirling	80	0.25	30	1.35	120	5.4	--
K506B##	Stirling	85	0.37	12	--	32.4	--	6165 hr
		85	0.5	14.7		29.4		
		110	0.59	12		20.3		
		110	0.5	10.5		21.0		
		150	1.18	12		10.2		
		150	0.5	7		14		
K526##	Stirling	80	0.25	30	0.907	120	3.63	>1000 hr
K526S##	Stirling	80	0.5	30	0.907	60	1.81	>1000 hr
Rutherford Appleton Laboratory, UK 2 stage	Stirling	20	0.06	85**	--	1417	--	--
		30	0.32	85**		265.6		

(continued)

Table 6.27 *(continued)*

Refrigerator ID	Type	T (K)	Heat Lift (W)	Power In (W)	Mass (kg)	Specific Power (W/W)	Specific Mass (kg/W)	Lifetime
Shanghai Institute of Technical Physics	J-T	27	4	--	8.6	--	2.15	--
S. H. E. Corp 4K, four-stage cryocooler	Stirling	7	0.001	--	--	--	--	--
Sterling Federal Systems Inc.	ADR	2	0.1	--	--	--	--	--
Stirling Technology Company (STC)								
Technology Demonstration Model (TDM)	Stirling	80	2.5	72[**]	--	28.8	--	500+ hr
			1	46[**]		46		
			2	62[**]		31		
			3	97[**]		32.3		
			4	127[**]		31.75		
		70	1	62[**]		62		
			2	107[**]		53.5		
			2.8	139[**]		49.6		
		60	1	137[**]		137		
Prototype Preliminary Design (PPD)	Stirling	80[*]	1.12[*]	31.3[*]	<5[*]	27.9	4.46	--
Telefunken AG	Stirling	80	0.25	--	--	--	--	--
Texas Instruments (TI)								
Split[##]	Stirling	80	0.33	--	1.27	--	3.81	--
Magnetic linear[##]	Striling	80	1	40	1.81	40	1.81	4000 hr
Tohoku University, Japan	Stirling	15	0.05	150	24	3000	480	--
		10	0					
Toshiba								
	ADR	0.3	--	--	--	--	--	12 hr
	Claude	4.4	4	--	--	--	--	--
Toshiba/Tokyo	G-M	5.15	1	2800	--	2800	--	--

Table 6.27 *(continued)*

Refrigerator ID	Type	T (K)	Heat Lift (W)	Power In (W)	Mass (kg)	Specific Power (W/W)	Specific Mass (kg/W)	Lifetime
Institute of Technology		6.35 10	2 5	2800 2800		1400 560		
TRW								
Space Cooler 3-stage	Stirling	65	0.25	14	1.145	56	4.58	10 yr*
stage one	Pulse-Tube	91.8	5.5	--	--	--	--	--
stage two	Pulse-Tube	39	0.724	--	--	--	--	--
stage three	Pulse-Tube	11.5	0.056	--	--	--	--	--
University of California at Berkeley	ADR	0.1	--	--	--	--	--	38 hr
University of Tokyo/ Osaka Oxygen Indust.	Solvay	15	1	--	--	--	--	--
USSR/SIAME Optimized	Stirling	80	6.3	205	---	32.54	--	--

ADR = Adiabatic Demagnetization Refrigerator
AMR = Active Magnetic Regenerator
G-M = Gifford-McMahon
J-T = Joule-Thomson
M-C = Magneto-Caloric
SVC = Saturated Vapor Compression
TA = Thermoacoustic
VM = Vuilleumier

* = designed
\- = input power to compressor
\-- = modified, Stirling, split component units
\+ = 75 rpm
++ = 100 rpm
+++ = this system uses 2 refrigerators to cancel out vibrations
# = compressor and expander modules
= primarily used as a tactical cooler
= approximately 1 W of cooling at 90 K for each 0.01 grams/s of argon

Ricor's tactical coolers have been extensively tested in the field and under battle conditions. Life testing of three Ricor cryocooler production units started on January 10, 1989. The test conditions included ambient temperature cycling between −30°C to + 50°C; three on/off segments per 20-h cycle; and 32 total hours of random vibration, 5 *g* between 20 and 2000 Hz, during the first 1500 h of operation. The failure criteria for the tests were an inability to maintain a cold tip temperature below 90 K with a 50-mW electrical heat load, a cool down time greater than 10 min, and a required input power greater than 12 W. The three units had a mean time to failure (MTTF) of 6165 h.

The Magnavox magnetically suspended, linear, Stirling-cycle refrigerator, developed by Philips Laboratories, was specifically designed for a 3- to 5-yr lifetime in spaceborne applications. The cooler technology was transferred to Magnavox electro-optical systems where it continues to be developed.

This cyrocooler uses closed-loop controlled, moving magnet, linear motors for the compressor and expander. The moving elements are supported by active contactless magnetic bearings with a clearance seal of 20 μm. The magnetic bearings are controlled by signals from fiber optic sensors, which detect the radial position of the shafts. The frequency, phase, stroke, and offset of the compressor and expander are controlled by signals from high-bandwidth LVDTs. An active counterbalance also uses a moving magnet linear motor and magnetic bearings to reduce vibration. A magnetic spring in the expander and gas springs in the compressor and counterbalance enhance the efficiency of these three active members. The magnetically suspended 24-kg cryocooler will deliver 5 W of cooling at 65 K with an operating input power of 160 W and a standby power of 16.

Stirling Technology Company has developed and tested a technical demonstration model (TDM) cryocooler that will provide 2.5 W of cooling at 80 K with 72 W of input power, 1 W of cooling at 70 K with 62 W of input, 2 W at 70 K with 107 W of input or 2.8 W of cooling at 70 K with 139 W of input. One watt of cooling at 60 K requires 137 W of input power.

The Magnavox Electro-Optical Systems MX 7043 mini-cryocooler is a free-displacer type, closed-cycle, Stirling-principle cryogenic cooler. The 2-kg cooler will deliver 0.88 W of cooling at 77 K with a power input of 55 W. The cooler has a guaranteed lifetime of 2500 h.

The Scientific Industrial Association of Microcryogenic Engineering in Omsk, Russia, has developed a cryocooler supplying 6.3 W of cooling at 80 K with an input power of 205 W.

Ball Aerospace Systems Group licensed Rutherford Appleton Laboratory technology in 1990 with the intent of improving the performance of the Oxford-designed split Stirling cycle cryocooler. The cooler uses diaphragm springs and clearance seals in order to reduce wearing surfaces. Nonfatigued springs were developed by limiting stress levels below the fatigue limit by a factor of 10 and by not exciting the harmonic frequency of the springs. To avoid gas contamination, the cooler minimized nonmetallic materials and was designed for thermal-vacuum bakeout. The cooler was hermetically sealed by replacing O rings with brazing. The single-stage cryocooler can provide 0.8 to 3 W of cooling at 80 K with an input power of 60 W, including electronics. The cooler has a mass of approximately 12 kg and is designed for a 10-yr lifetime.

The Ball two-stage Stirling cryocooler supplies 0.3 W of cooling at 30 K with an input power, including electronics, of 75 W. The cooler has a mass of approximately 15 kg and is designed for a 10-yr lifetime.

Lucas Aerospace Limited and Lockheed Missiles & Space Company, Inc., are working on a joint project to develop a series of split Stirling cycle cryocoolers. So far there are four models: the CCS 4000, CCS 1000, CCS 500, and CCS 250. They can supply, respectively, 4, 1, 0.5, and 0.25 W of cooling at 65 K. The target lifetime is 90,000 h running with 1000 on/off cycles of operations.

Scientists at Rutherford Appleton Laboratory are developing a two-stage Stirling cycle cooler for space use. The cooler can produce 0.06 W of cooling at 20 K and 0.32 W of cooling at 30 K with a total input power to the compressors of 85 W.

Brayton Cycle Cryocoolers. Creare, Inc., is developing a small single-stage, reverse turbo-Brayton cryocooler for NASA/Goddard Space Flight Center. The cooler's turbine has a diameter of 3.175 mm with a shaft rotation speed of 510,000 rpm. The working fluid will be neon, which has a boiling point of 27.09 K. The cooler will be designed[11] to deliver 5 W of cooling at 65 K and have a specific power of 40 W/W, which means the input power will be 200 W. The cooler is designed to weigh less than 50 kg and have a lifetime of about 10 yr.

Vuilleumier Cycle Cryocoolers. Hughes Aircraft Corporation developed a three-stage Vuilleumier-cycle cryocooler called Hi Cap to deliver 12 W of cooling at 75 K, 10 W of cooling at 33 K, and 0.3 W of cooling at 11.5 K, all with a maximum input power of 2700 W. With the Vuilleumier cycle, most of this input power is in the form of heat. The designed lifetime of the cooler is intended to be about 20,000 h.

Mitsubishi Electric Company, located in Amagaskai, Japan, and the Electrotechnical Laboratory, located in Tsukuba, Japan, have developed a small Vuilleumier refrigerator for spaceborne instrumentation. This cooler will deliver 1.8 W of cooling at 80 K with a total power input of 210 W and a cooler weight of 6 kg.

The Centre d'Etudes Nucléaires in France was developing in 1992 a Vuilleumier refrigerator for spaceborne instrumentation. The project is being supported by CNES with L'Air Liquide and Aerospatiale as industrial partners. The cooler design specifications are 1 W of cooling at 50 K and 5 W of cooling at 190 K with a power input of 250 W. The design lifetime will be 50,000 h.

Gifford-McMahon Cycle Cryocoolers. The Leybold-Heraeus Company in Cologne, Germany, has successfully developed, built, and tested a split cold-head Gifford-McMahon refrigerator. The first stage of the refrigerator supplies 10 W of cooling at 80 K, the second stage supplies 2 W of cooling at 20 K, and the ultimate temperature must be less than 14 K. A requirement of this cooler is to have reliable operation with no maintenance for 8000 h.

Toshiba Research and Development Center and the Department of Applied Physics at the Tokyo Institute of Technology have developed a Gifford-McMahon refrigerator using rare earth compounds as a regenerator matrix. An Er_3Ni compound can produce a no-load temperature of 5.15 K, 1 W of cooling at

6.35 K, 2 W of cooling at 7.2 K, and 5 W of cooling at 10 K. The input power is 2800 W.

The Central Research Laboratory at Mitsubishi Electric Corporation has developed a three-stage Gifford-McMahon refrigerator in which they have successfully liquefied helium. The third stage of this cooler can produce a no-load temperature of 3.3 K, 20 mW of cooling at 4.2 K, and 50 mW of cooling at 5.2 K, all with an input power to the compressor of approximately 2500 W. It takes about 200 min for the temperature of the third stage to reach 4.2 K.

The Centre of Advanced Technology in Indore, India, has built a two-stage Gifford-McMahon refrigerator that is operated on the gas balancing principle. The cryocooler supplies 2.6 W of cooling at 30 K with a speed of 75 rpm and 2.4 W of cooling at 30 K with a speed of 100 rpm. It takes 75 min to obtain a temperature change from room temperature to 26 K at a speed of 75 rpm but it only takes 65 min with a speed of 100 rpm.

Joule-Thomson Cryocoolers. Joule-Thomson cryocoolers are open-system coolers that rely on the adiabatic expansion of high-pressure gases to produce low temperatures in a very short period of time—on the order of minutes.

Ball Aerospace Systems Group has been developing a Joule-Thomson cryocooler since 1982. This cooler uses a fourth-generation advanced breadboard compressor, which incorporates oil lubrication for long life, circulating fluid for contamination control, and a zero-*g* contamination control system. This compressor has demonstrated no degradation after 4000 h of operation with 100 start/stops. The cooler has a patented thermoelectric cooler for a first stage and dual-pressure Joule-Thomson second and third stages. The cold head has no moving parts and a liquid is produced in the final stage. This cryocooler can provide 0.65 to 3.5 W of cooling at 65 K with a power consumption of 100 to 210 W. This gives a specific power of 153.8 W/W with 0.65 W of cooling and 100 W of input power and a specific power of 60 W/W with 3.5 W of cooling and 210 W of input power. The cooler is designed for a lifetime greater than 5 yr with a reliability greater than 0.95%.

The Shanghai Institute of Technical Physics has developed a fast cool down Joule-Thomson mini-cryocooler utilizing a two-phase valve and a directly wound fin tube to provide 4 W of cooling at 27 K with a liquid nitrogen consumption of 1.5 l/h. The cooler uses a working fluid of H_2 of Ne + LN_2 with a pressure range of 0.1 to 12 MPa. The cryostat weight is 8.6 kg with a no load cool down time of 0.5 to 1 min.

MMR Technologies in Mountain View, California, has developed a two-stage fast cool-down Joule-Thomson refrigerator using nitrogen gas and a nitrogen-hydrocarbon gas mixture as the refrigerant. The cooler uses a venturi pump to reduce the operating temperature to less than 70 K. This cooler provides 13.6 W of cooling for cool down from 300 to 70 K in 10 s.

Sorption Cryocoolers. The Jet Propulsion Laboratory (JPL) in Pasadena, California, has done extensive work on the development of sorption refrigeration. A sorption cryocooler is a type of Joule-Thomson closed-cycle cryocooler in which the mechanical compressors are replaced with adsorption compressors. In the sorption refrigeration process, low-pressure gas is absorbed by physical absorption on the surface of a material or by chemical absorption within a solid material, called the *sorbent*. The sorbent is then heated with thermal

energy to about 100 to 200°C to desorb the gas at high pressure. The pressurized gas is passed through precooling stages and then expanded in a Joule-Thomson valve where it is partially liquefied and cooled to the final low temperature. The low-pressure boil-off gas is then reabsorbed by the sorbent. Sorption refrigeration has the potential lifetime of decades with virtually no vibration.

Charcoal/methane (C/CH_4) physisorption refrigerators operate between 110 and 150 K and deliver 2 W of cooling with a specific power of 80 W/W at 140 K. Lower temperatures can be achieved by cascading one type of sorption refrigerator after another. The CH_4 + praseodymium cerium oxide (PCO) chemisorption refrigerator will operate between 55 and 90 K and supply 1.7 W of cooling with a specific power of 180 W/W at 70 K. The next step is a CH_4 + PCO + liquid hydrogen (LH_2) hydride chemisorption refrigerator operating between 14 and 30 K and delivering 1.5 W of cooling. A CH_4 + PCO + LH_2 hydride + solid hydrogen (SH_2) hydride chemisorption refrigerator will operate between 7 and 10 K and a Ch_4 + PCO + LH_2 hydride + mechanical helium system refrigerator will operate between 4 and 5 K.

In more recent studies at JPL, a three-stage carbon/xenon (C/Xe) + carbon/krypton (C/Kr) + PCO refrigerator has produced 1 W of cooling at 65 K with only 56 W of input power. By cascading sorption refrigerators, a 65 K refrigerator + LH_2 hydride cooler has a temperature of 15 K with a specific power of 250 W/W. A 15 K + SH_2 hydride refrigerator has a temperature of 10 K with a specific power of 400 W/W.

As of February 1990, two PCO compressors had each accumulated more than 9600 h and 15,700 cycles of operation without any noticeable signs of degraded performance. Also, in autumn of 1989, four C/Kr compressors began 24-h continuous operation.

Pulse-Tube Cryocoolers. Orifice pulse-tube refrigerators have been developed by Mikulin et al.[12]; at the National Bureau of Standards in Boulder, Colorado; and at Xi'an Jiaotong University in Xi'an, China. These refrigerators reached a low temperature of 105, 60, and 42 K, respectively.[21]

The Cryogenic Laboratory in Beijing, China, has developed a compact, coaxial pulse-tube refrigerator especially for cooling electronic devices. This refrigerator achieved a minimum temperature of 62 K and will produce 2.5 W of cooling at 77 K. The total weight of the cooler is approximately 40.4 kg and the compressor/pressure wave generator requires 188 W of input power.

Laboratoire de Radioastronomie in Paris, France, has developed a hybrid pulse-tube refrigerator that consists of both regular and orifice pulse tubes. This refrigerator achieved a no-load temperature of 57 K and will supply 12 W of cooling at 72 K.

Adiabatic Demagnetization Cryocoolers. Group P-10 at the Los Alamos National Laboratory has produced a Carnot-cycle magnetic refrigerator using gadolinium gallium garnet (GGG) as the magnetic material refrigerant. This refrigerator will supply 0.66 W of cooling at approximately 4.3 K while rejecting the heat at 15 K.

The University of California at Berkeley has developed a small adiabatic demagnetization refrigerator (ADR) to cool bolometric infrared detectors on the Multiband Imaging Photometer of the Space Infrared Telescope Facility. One such refrigerator has achieved a low temperature of 0.1 K with a hold time of greater than 38 h.

The Energy Science and Technology Laboratory of the Toshiba Research and Development Center in Kawasaki, Japan, is developing an ADR for cooling infrared detectors. Using a magnetic material of manganese ammonium sulphate, a temperature below 0.3 K was maintained for more than 12 h.

6.4 MECHANICAL DESIGN

The mechanical design of a cryogenic system is an interdisciplinary effort. There are multiple areas of design that relate to experience and are not rigorously supported by analytical equations of engineering. The integrity of bonded joints at cryogenic temperatures, the diameters and length of cryogen lines to reduce thermoacoustical vibrations, the fatigue of burst disks when subjected to cyclic loadings at required fills, the reduction of thermal resistances at bolted joints, and the outgassing of various materials are examples of problems that relate to the art rather than the science of cryogenic design. Fracture control, stability of elastic structures, finite element stress analysis, dynamic modeling, and other rather sophisticated disciplines may be required to design dewars properly where a high risk of failure is involved. The complete spectrum of design and the documentation of years of experience cannot be covered herein. This section includes only the fundamentals that relate to a basic design.

6.4.1 Supply Tank

A cryogenic system usually contains a storage tank that provides a reservoir of cryogen in the solid, liquid, or supercritical phase. The cryogenic hold time of the system depends on the size of the reservoir and the rate at which the parasitic and thermally controlled heaters boil off the cryogen. Conduction rods penetrating into the tank, conduction straps anchored to the wall of the tank, and/or vapors flowing from the tank provide the cooling required throughout the system. The design of the storage tank may depend on several variables, such as volume, geometry, weight restraints, fill and vent procedures, space available, safety lines, plumbing, etc. The structural design, however, depends primarily on the volume, shape, and pressure differential. At present, most cryogenic tanks are made from aluminum. Composite materials have the potential of significant weight reduction, yet leakage and related diffusion problems through the wall thickness have limited the use of nonmetallic tanks. The purpose of this section is to provide the basic criteria to assist in the design of a metal storage tank.

Spherical Tanks. A spherical tank is structurally the most efficient shape in terms of stress and buckling resistance to withstand internal or external pressure differentials. For a given volume, the sphere results in the least weight and wall thickness compared to other geometries. In fact, the membrane wall stress in a thin-walled sphere is only one-half the burst stress in the wall of a cylindrical tank with the same pressure, wall thickness, and radius. Spherical tanks are most commonly used for large pressure differentials, larger than is common for most cryogenic systems.

The surface of a spherical tank may include plumbing fixtures, attachment plates for sensors, or other structures that result in localized stress concentration due to geometry discontinuities. The localized stress field surrounding

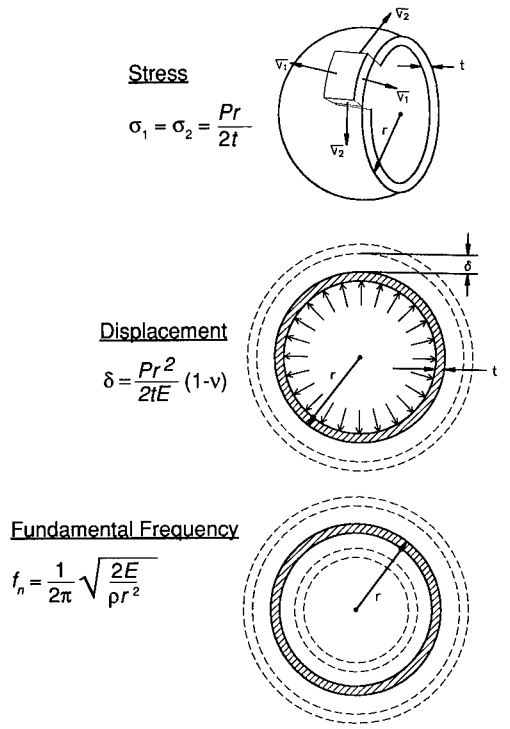


Fig. 6.33 Thin-wall spherical tank internal or external pressure.

a discontinuity must be superimposed on the membrane stresses for a thin-walled sphere.

As shown in Fig. 6.33, the wall stress in a thin-walled spherical tank is equal in all directions in the tangent plane and it is given as

$$\sigma = \frac{Pr}{2t} , \tag{6.82}$$

where σ is the stress, P the pressure, r the radius, and t the wall thickness.

The radial displacement of a spherical tank as shown in Fig. 6.33 is

$$\delta = \left(\frac{Pr^2}{2tE}\right)(1 - \nu) , \tag{6.83}$$

where δ is the radial displacement, P the pressure, r the radius, t the thickness, E the modulus of elasticity, and ν is Poisson's ratio. The equation is valid for a perfect sphere, but becomes an approximation if plates and fixtures attached to the shell cause nonsymmetric radial displacements.

The continuous sphere has an infinite number of natural frequencies of vibration. The lowest or fundamental frequency is typically the most critical. As shown in Fig. 6.33, the fundamental natural frequency of a thin-walled sphere is

$$f_n = \frac{1}{2\pi}\left(\frac{2E}{\rho r^2}\right)^{1/2}, \tag{6.84}$$

where f_n is the frequency in Hertz, E the modulus of elasticity, r the radius, and ρ the mass density. This frequency relates to symmetric radial displacements. Sphere attachment points may alter the fundamental frequency and induce other modes.

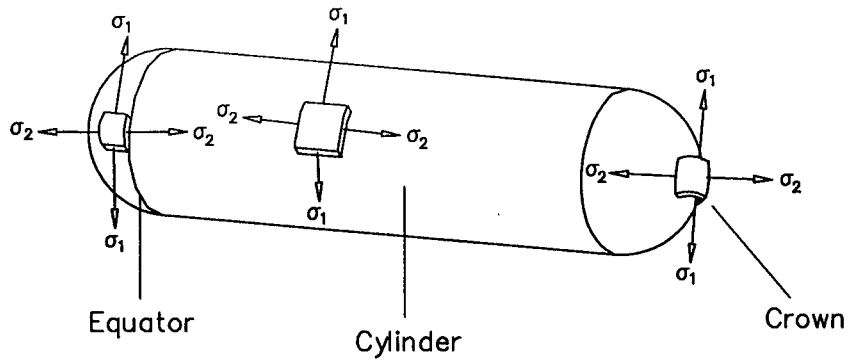
Cylindrical Tanks. Cryogenic tanks frequently have pressure differentials of less than 45 psi. Consequently, static pressures are often not the dominant design variable. The geometry and volume of the space available for the reservoir, the ease of fabrication, the attachment of cold fingers, fill and vent requirements, tank orientations relative to required modes of operation, tank suspension systems, and fracture control may be more significant than the pressure differentials. Cylindrical tanks with different end conditions are commonly used.

Hemispherical Ends. The hemisphere is structurally the most efficient end shape for a cylindrical tank, because bending stresses are essentially eliminated in thin-walled hemispheres. A circular flat plate in the end of a cylinder is subject to bending stresses and is structurally the most inefficient end shape. Torospherical and ellipsoidal end plates experience some localized bending and are somewhere between the hemispherical and flat ends in terms of structural efficiency.

The stresses in a thin-walled cylindrical tank with hemispherical ends are shown in Fig. 6.34. The burst or tangential stress in the cylinder, σ_1, is the maximum stress. The longitudinal stress, σ_2, and the stresses at the equator and crown are only one-half the burst stress. The stress is constant throughout the end hemispheres.

Stress concentration occurs at the equator. The radial expansion of an unrestrained cylinder is not equal to the unrestrained expansion of a hemisphere when subjected to the same pressure. Since the head and the cylinder are connected, the two must experience the same displacement. The stresses in pressure cylinders associated with this so-called "discontinuity" can be evaluated by using the theory of beams on elastic foundations. In a typical pressure vessel made from a cylinder attached to hemispherical heads with similar materials and wall thicknesses, the maximum longitudinal stress increases from $0.5\,Pr/t$ to $0.646\,Pr/t$ due to the discontinuity at the equator. The maximum tangential or burst stress increases from Pr/t to $1.032\,Pr/t$. Since the burst stress usually dominates in design, we can see that the increase due to the discontinuity at the equator is in the neighborhood of 3%. Although slight, the additional stress at the equator is motivation to move the welded joint away from the equator and into the cylinder. This is usually done by including a cylindrical stub on the hemispherical head that is actually welded to the tank cylinder.

Ellipsoidal Heads. Ellipsoidal heads are commonly used in cryogenic systems. The reduced tank length is sometimes an advantage that overshadows the loss in structural efficiency. The derivations of stress equations for ellipsoidal heads are found in several texts[13,14] and are summarized in Fig. 6.35. The longitudinal stress σ_2 is independent of the ellipsoid minor diameter b at the equator.



σ_1 = burst pressure r = radius

σ_2 = longitudinal pressure t = wall thickness

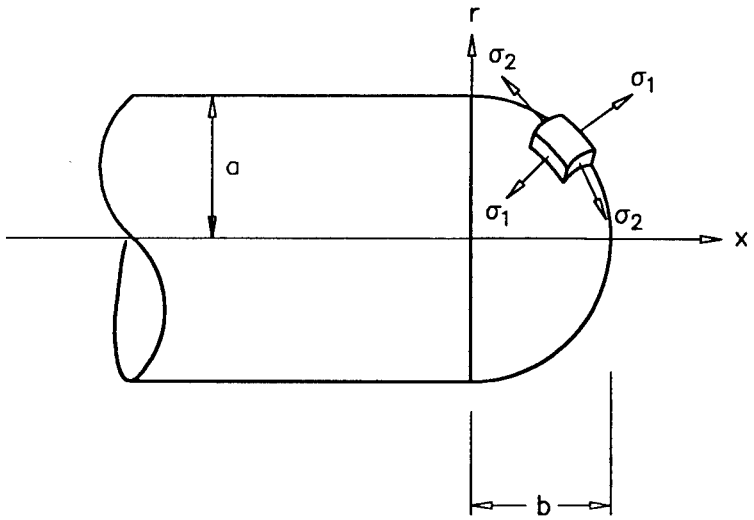
Cylindrical Shell	Equator	Crown
$\sigma_1 = \frac{Pr}{t}$	$\sigma_1 = \sigma_2 = \frac{Pr}{2t}$	$\sigma_1 = \sigma_2 = \frac{Pr}{2t}$
$\sigma_2 = \frac{Pr}{2t}$		

Fig. 6.34 Cylindrical tank hemispherical heads.

The stresses σ_1 and σ_2 are equal at the crown and increase as a/b increases or as the ellipsoid flattens. At the equator, the burst stress or tangential stress σ_1 is equal to the longitudinal stress at $a = b$; however, σ_1 is reduced as the ellipse flattens. At $a/b = \sqrt{2}$, the tangential stress σ_1 is zero at the equator. For $a/b > \sqrt{2}$, the hoop stress is negative. Localized buckling may occur at the equator for values of $a/b > \sqrt{2}$. At $a/b = 2$, the compressive hoop stress at the equator is equal to the tension stress at the crown and is often used as a practical limit for the flatness of the head. The stress relationships as a function of major and minor axes are conveniently summarized by Harvey[15] as shown in Fig. 6.36.

The shearing stress depends on the ratio of a/b. At values of $a/b < \sqrt{2}$, the maximum shearing stress is at the crown of the head. For $a/b > \sqrt{2}$, the maximum shearing stress shifts from the crown to the equator.[15]

There is also a stress concentration at the equator where the radial expansion of the cylinder is restricted by the ellipsoidal end. A bending stress results that must be combined with the membrane stresses. Timoshenko[14] and others have shown that the ellipsoidal bending stress to be added to the membrane stress increases by a factor of a^2/b^2 over the bending stress in a hemispherical end. The combined membrane and bending stress near the equator in the longitudinal direction, σ_2, becomes



σ_1 = tangential, hoop, or burst stress

σ_2 = meridional or longitudinal stress

$$\sigma_1 = \frac{P}{2tb}\,[a^4-2r^2(a^2-b^2)]\,[a^4-r^2(a^2-b^2)]^{\frac{-1}{2}}$$

$$\sigma_2 = \frac{P}{2tb}\,[a^4-r^2(a^2-b^2)]^{\frac{1}{2}}$$

Ellipse equation $\left(\frac{x}{b}\right)^2+\left(\frac{r}{a}\right)^2 = 1$

Equator $\quad \sigma_1 = \frac{Pa}{t}\left(1-\frac{a^2}{2b^2}\right) \qquad \tau_{max} = \frac{Pa}{4t}\left(\frac{a^2}{b^2}-1\right)$

$$\sigma_2 = \frac{Pa}{2t}$$

Crown $\quad \sigma_1 = \sigma_2 = \frac{Pa^2}{2bt} \qquad \tau_{max} = \frac{P}{4}\left(\frac{a^2}{bt}+1\right)$

Fig. 6.35 Cylindrical tank ellipsoidal head.

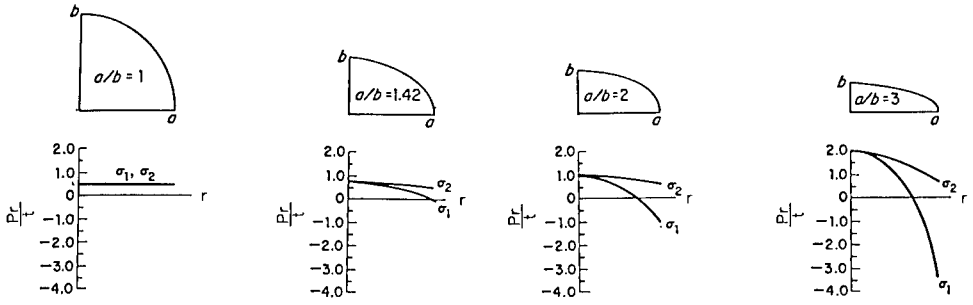


Fig. 6.36 Stresses in ellipsoidal heads.

$$(\sigma_2)_{\max} = \left(1 + 0.293\frac{a^2}{b^2}\right)\frac{Pr}{2t} . \tag{6.85}$$

The combined membrane and bending stress near the equator in the tangential direction becomes

$$(\sigma_1)_{\max} = \left(1 + 0.032\frac{a^2}{b^2}\right)\frac{Pr}{t} . \tag{6.86}$$

The tangential stress dominates and the maximum stress is located at $x = 1.44\sqrt{rt}$ measured from the equator longitudinally along the cylinder.

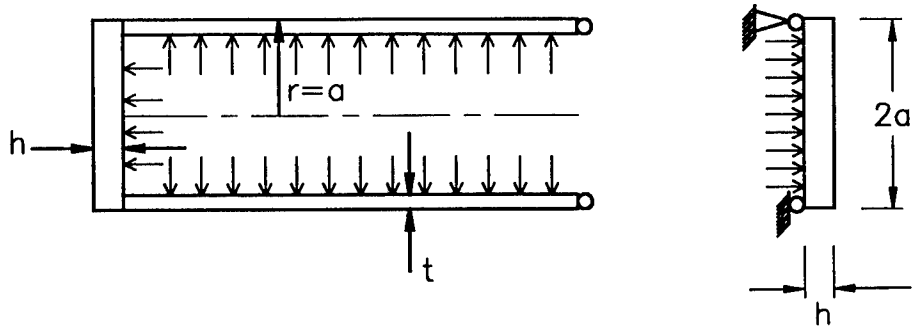
Flat-Head Cylinders. Cryogenic tanks made from cylinders with round flat end plates are structurally inefficient, yet are often used. Shells such as thin-walled cylinders and spheres are often considered as membrane structures in that internal stresses are only tension or compression and are uniform throughout the wall thickness. The circular end plates of a cylinder are subjected to bending stresses and are not considered as membranes. The bending stresses in the end plates are the most critical in the design of the tank with flat ends.

The stresses in the walls of a cylindrical tank with flat end plates are the same as for hemispherical or ellipsoidal ends. At a short distance from the ends and beyond, the burst or tangential stress is $\sigma_1 = Pr/t$, whereas the longitudinal stress is $\sigma_2 = Pr/2t$.

The design of the end plates is governed by classical plate theory. A decision must be made as to the characterization of the plate edge condition. The actual end conditions usually fall between the extremes of simply supported and clamped. Since the end plates are usually much thicker than the wall of the cylinder, the end plate boundary conditions can be approximated by assuming zero moment resistance to bending of the plate, which is essentially a simply supported edge. It is noted that zero moment at the plate edge is usually conservative in the determination of plate stresses, but not in the prediction of cylinder wall stresses. A reasonable approach to envelop the maximum stresses is to assume a simply supported condition for the end plate and then clamped conditions for the cylinder end.

The equations for design of the end plate assuming simply supported edges are shown in Fig. 6.37. The stresses, displacement, and natural frequency for an end plate assuming fixed or clamped edge conditions are given in Fig. 6.38.

The bending moments at the connection of the cylinder to the end plate result in stresses that must be combined with the membrane stresses in the cylinder. These bending stresses may be several times greater than the membrane stresses, depending on the rigidity of the end plate. If the end plate is sufficiently rigid to provide a cylinder end that is "fixed" or "built-in," then there is no cylinder expansion in the radial direction of the end plate. The stress in the cylinder at the fixed end plate is bending only. It can be shown[13] that the maximum bending stress in the cylinder at the fixed end is $\sigma = 1.82\ Pr/t$ and is directed longitudinally. In other words, the stress concentration factor to account for bending is 1.82 times the maximum membrane stress at $\sigma_1 = Pr/t$. Note that the bending stress is longitudinal at the end of the fixed cylinder, whereas the maximum membrane stress is circumferential and away from the end.



Circular End Plate
Simply Supported (Zero End Movement)

$$\sigma_{max} = \frac{3(3+\nu)Pa^2}{8h^2} \quad \bullet \text{ center} \qquad D = \frac{Eh^3}{12(1-\nu^2)}$$

$$\delta_{max} = \frac{(5+\nu)Pa^4}{64(1+\nu)D} \quad \bullet \text{ center} \qquad f_n = \frac{4.98}{2\pi a^2}\sqrt{\frac{D}{\rho h}}$$

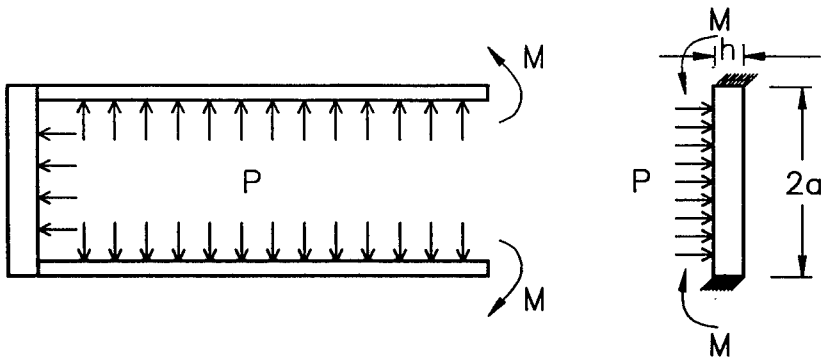
Fig. 6.37 Simply supported flat-head cylinder.

If the end plate is too thin to provide a relatively rigid cylinder end support, then end plate deformations induce larger moments at the cylinder ends. A finite element analysis is then encouraged. For example, an aluminum cylinder with a radius of 5 in., a wall thickness of 0.0625 in., and flat end plates at a thickness of 0.25 in. resulted in bending stresses at the cylinder-plate interface at approximately eight times the membrane stress in the cylinder.

Buckling of Cylinders. Cryogenic thin-walled cylindrical shells may be subjected to vacuum loadings that result in external pressures acting on the wall of the cylinder. A cryogenic system usually consists of an outer cylindrical shell that is subjected to atmospheric pressure externally and vacuum conditions on the inside. In space operations, the outside or external pressure may reduce to a near zero absolute pressure, yet launch conditions and ground operations require a differential of 1 atm on the outer shell.

The internal cryogen reservoir may never experience a resultant external pressure under normal operating conditions. There are anomalies, however, such as a loss of vacuum between the cryogen tank and the outer shell, that may result in a net external pressure on the reservoir. Consequently, it is usually advisable for both interior and exterior cylindrical shells to be designed to withstand external pressures and prevent cylinder buckling.

Thin-walled cylinders subjected to external pressure may collapse at wall stresses well below yield stresses. The conventional stress analysis is not adequate to prevent failure by buckling. The criteria to predict buckling has resulted from years of testing and analysis. Several equations provide reasonable estimates; however, the user must realize that the results have limited accuracy. A safety factor of 5 is commonly prescribed. In other words, the design should be capable of five times the expected external pressure.



Circular End Plate
Clamped Edges (Moment Resistance)

$$\sigma_{max} = \frac{3}{4}\frac{Pa^2}{h^2} \quad @\ boundary\ r = a$$

$$\delta_{max} = \frac{Pa^4}{64D}$$

$$D = \frac{Eh^3}{12(1-\nu^2)}$$

$$f_n = \frac{10.22}{2\pi a^2}\sqrt{\frac{D}{\rho h}}$$

Cylinder Wall Bending Stress

$$M = \frac{Pa^2}{8}$$

$$\sigma = \frac{6M}{h^2}$$

Fig. 6.38 Clamped end condition for flat-head cylinder.

An equation used to predict buckling in thin-walled cylinders was provided by Windenburg[16] as

$$P_c = \frac{2.60\,E(t/D)^{2.5}}{L/D - 0.45(t/D)^{1/2}}, \tag{6.87}$$

where P_c is the external pressure that will collapse the cylinder, D is the diameter, t is the thickness of the wall of the cylinder, and L is the center-to-center length between ends or stiffeners that maintain the circular cylinder wall. No axial loads are considered. An obvious source of error is related to the capability of the stiffener to provide a circular configuration.

If the length L is greater than $1.11D\sqrt{D/t}$, then the critical pressure for collapse is independent of the length and the buckling equation suggested by Harvey[15] is

$$P_c = \frac{2E}{(1-\nu^2)}\left(\frac{t}{D}\right)^3, \tag{6.88}$$

where E is the modulus of elasticity and ν is Poisson's ratio.

Fracture Control. Materials and welded joints used in the fabrication of pressure vessels may include embedded porosity, flaws, and/or cracks that could grow under cyclic loadings and result in structural failure at stresses well below the material yield strength. An attempt to limit the growth of potential flaws and/or cracks and prevent failure in a prescribed cyclic load environment is commonly called *fracture control.* The critical concerns are usually one of the following: (1) What are the critical flaw sizes that will grow to failure at the expected operational stress levels? (2) What are the initial flaw sizes and where are they located and how are they oriented? (3) Will these initial flaws grow to failure when subjected to the cyclic load environment for an expected service life of the vessel?

A thorough fracture mechanics analysis of a pressure vessel is beyond the scope of this chapter. Several reference books provide the fundamental equations to complete a fracture analysis. Perhaps the most acceptable approach is to utilize a standard software package called FLAGRO4, which is available through NASA and allows variables such as initial crack size, geometry, operational stresses, load cycles, and unstable flaw growth to be related.

Some screening equations have been proposed that tend to reduce the need for fracture analysis and suggest that the conventional ductile stress analysis is adequate. One such screening equation is provided by the Department of Defense (DOD) as follows:[17]

$$\frac{K_{IC}}{\sigma} \geq 2\alpha\sqrt{t} \ , \tag{6.89}$$

where

K_{IC} = material plane strain fracture toughness
σ = applied maximum stress
α = proof test factor (minimum = 1.1)
t = material thickness.

A study was completed by Goddard Space Flight Center to suggest that the so-called "DOD screening equation" is not always conservative in the limiting case where the equal sign is met.[18] Nonetheless, as the K_{IC}/σ ratio increases when compared to $2\alpha\sqrt{t}$, the need for additional fracture analysis on metal tanks decreases. The value of K_{IC} usually increases as the yield strength of a material decreases. Usually, thin-walled aluminum tanks used in cryogen systems will leak before they burst. The size of the crack required for unstable growth would allow leakage and would reduce the pressure to avoid a sudden rupture. The sophistication of the fracture analysis required is a judgment decision that may depend on several variables.

Pressure Vessel Applications. A simple application will demonstrate the use of the equations relating to the design of a pressure vessel. Assume a thin-walled cylindrical shell with a flat end and an ellipsoidal end as shown in Fig. 6.39. The membrane stresses in the cylinder wall are calculated as follows:

$$\sigma_1 = \frac{Pr}{t} = \frac{60(5)}{0.0625} = 4800 \text{ psi} \ , \tag{6.90}$$

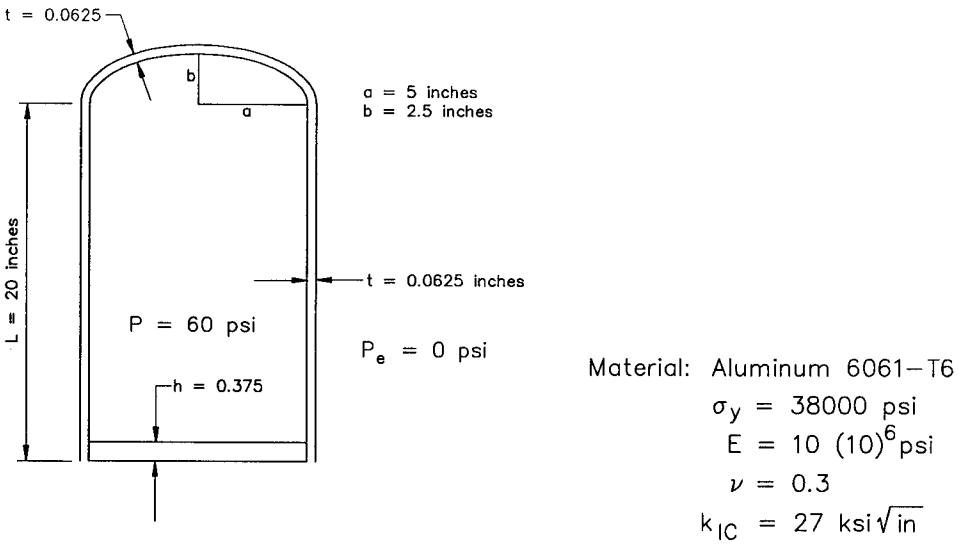


Fig. 6.39 Pressure vessel application: cylindrical tank with flat head.

$$\sigma_2 = \frac{Pr}{2t} = \frac{60(5)}{(2)0.0625} = 2400 \text{ psi} \ . \tag{6.91}$$

To calculate the maximum stress in the flat end plate, simply supported end conditions are assumed. From Fig. 6.37 the equations follow as:

$$\sigma_{\max} = \frac{3(3 + \mu)\, Pa^2}{8h^2}$$

$$= \frac{3(3 + 0.3)(60)(5)^2}{8(0.375)^2} = 13{,}200 \text{ psi} \ , \tag{6.92}$$

where the maximum stress is at the center. The center transverse deflection of the end plate is calculated as follows:

$$\delta = \frac{(5 + \mu)\, Pa^4}{64(1 + \mu)D} \qquad D = \frac{Eh^3}{12(1 - \mu^2)} \ , \tag{6.93}$$

$$D = \frac{(10)(10)^6(0.375)^3}{12(1 - 0.3^2)} = 48{,}292 \text{ lb in.} \ , \tag{6.94}$$

$$\delta = \frac{(5 + 0.3)(60)(5)^4}{64(1 + 0.3)48292} = 0.049 \text{ in.} \ . \tag{6.95}$$

The membrane stresses in the ellipsoidal head are characterized by Fig. 6.40, where $a/b = 2.0$. The stresses are suggested to be maximum at the crown and the equator. Using the equations from Fig. 6.35, the equator stress is

$$\sigma_1 = \frac{Pa}{t}\left(1 - \frac{a^2}{2b^2}\right) = \frac{(60)(5)}{0.0625}\left[1 - \frac{5^2}{2(2.5)^2}\right] = -4800 \text{ psi} \ , \tag{6.96}$$

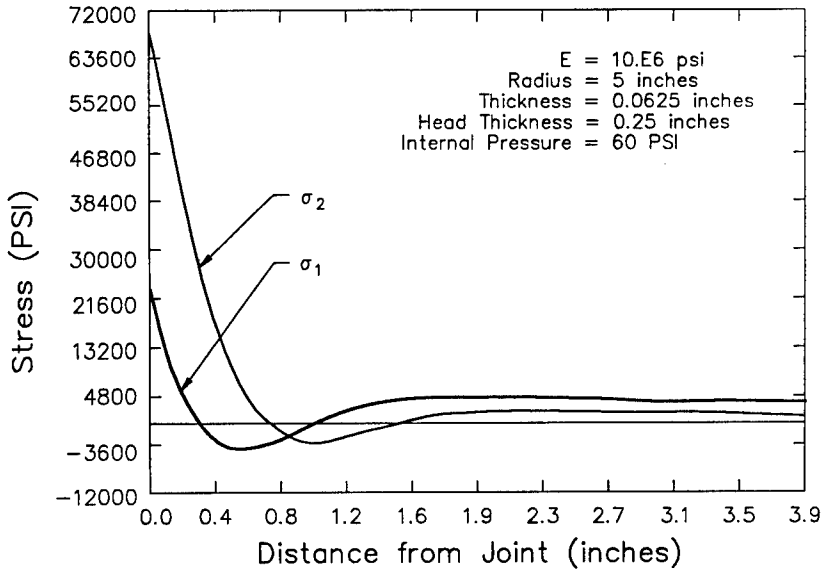


Fig. 6.40 Inner surface stresses in cylindrical tank with a flat head.

$$\sigma_2 = \frac{Pa}{2t} = \frac{(60)(5)}{2(0.0625)} = 2400 \text{ psi} , \tag{6.97}$$

$$\Gamma_{\max} = \frac{Pa}{4t}\left(\frac{a^2}{b^2} - 1\right) = \frac{(60)(5)}{4(0.0625)}\left(\frac{5^2}{2.5^2} - 1\right) = 3600 \text{ psi} . \tag{6.98}$$

The stresses at the crown are calculated as

$$\sigma_1 = \sigma_2 = \frac{Pa^2}{2bt} = \frac{60(5)^2}{2(2.5)(0.0625)} = 4800 \text{ psi} , \tag{6.99}$$

$$\Gamma_{\max} = \frac{P}{4}\left(\frac{a^2}{bt} + 1\right) = \frac{60}{4}\left[\frac{5^2}{2.5(0.0625)} + 1\right] = 2415 \text{ psi} . \tag{6.100}$$

The ellipsoidal head limits the expansion of the cylinder and bending stresses are induced near the equator. The membrane stresses combined with the bending stresses may be calculated as

$$\begin{aligned}(\sigma_2)_{\max} &= \left(1 + 0.293\frac{a^2}{b^2}\right)\frac{Pr}{2t} \\ &= \left[1 + 0.293\left(\frac{5}{2.5}\right)^2\right]\frac{Pr}{2t} \\ &= 2.172\frac{Pr}{2t} = 2.172\frac{(60)(5)}{2(0.0625)} = 5213 \text{ psi} ,\end{aligned} \tag{6.101}$$

$$(\sigma_1)_{\max} = \left(1 + 0.032\frac{a^2}{b^2}\right)\frac{Pr}{t} = 1.128\frac{Pr}{t} = 5414 \text{ psi} \ . \tag{6.102}$$

Note that σ_2, the stress in the longitudinal direction, was increased 217% by superimposing the cylinder bending stresses. The circumferential stress σ_1 was increased 12.8% due to bending. Since the membrane stresses longitudinally are only half the circumferential, the combined membrane and bending stresses in the circumferential direction still dominate.

The flat head must also be evaluated at the equator. If one assumes that the end plate is sufficiently rigid that the cylinder end is considered fixed and rigid, then the equation follows as:

$$\sigma = 1.82\frac{Pr}{t} = 1.82\frac{(60)(5)}{0.0625} = 8736 \text{ psi} \ , \tag{6.103}$$

which is 1.82 times the maximum membrane stress and is longitudinal at the end rather than circumferential.

As noted earlier, a designer must be cautious about the assumption that the end plate is sufficiently rigid to provide a fixed joint for the cylinder end. For example, if the end plate were reduced to a thickness of 0.25 in., then the center deflection could be calculated as 0.166 in., which is 3.38 times the center deflection of the 0.375-in. plate. The deflection varies as the cube of the plate thickness. An end plate with a 5-in. radius and a calculated center deflection of 0.166 in. would alert a designer to flexibility and violation of the rigid plate assumption. The stress in the cylinder wall attached to an end plate only 0.25 in. thick would still be predicted as

$$\sigma = 1.82\frac{Pr}{t} = \frac{1.82(60)(5)}{0.0625} = 8736 \text{ psi} \tag{6.104}$$

if the rigid plate assumption were erroneously made. A finite element model of the same cylinder predicts stresses near 66,000 psi as shown in Figs. 6.40 and 6.41. The maximum stresses in the cylinder would be roughly 7.5 times the stresses predicted with the hand analysis. Note, however, that the plate stresses at the center vary as the square of the plate thickness. An end plate with stresses significantly higher than the cylinder membrane stresses will result in relative plate rigidity sufficient to use the hand calculations for cylinder design. The most common error is to design an end plate that is too thin.

The cylinder can be evaluated to determine the external pressure required for buckling. The critical buckling pressure may be estimated as

$$P_c = \frac{2.60E(t/D)^{2.5}}{L/D - 0.45(t/D)^{1/2}}$$

$$= \frac{2.60(10)(10)^6(0.0625/10)^{2.5}}{20/10 - 0.45(0.0625/10)^{1/2}} = 40.8 \text{ psi} \ . \tag{6.105}$$

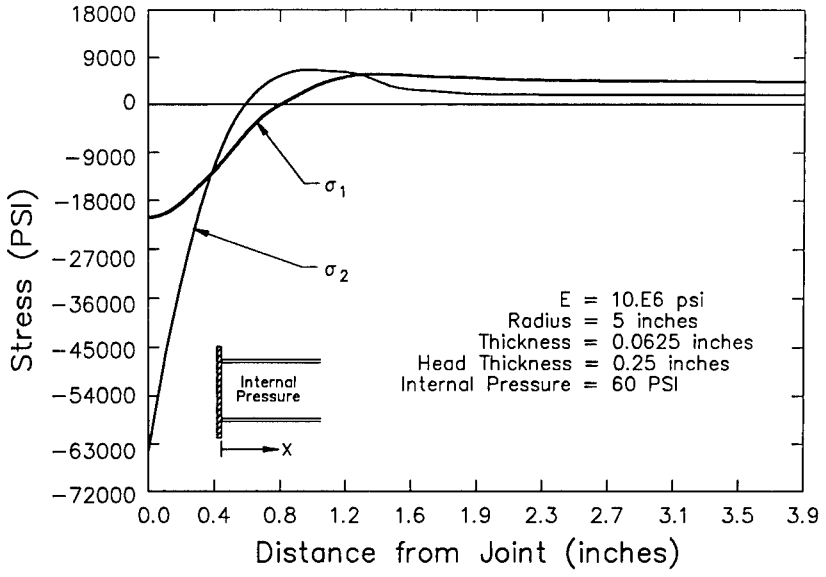


Fig. 6.41 Outer surface stresses in cylindrical tank with a flat head.

Since 40.8 psi/15 psi = 2.72, the suggested buckling safety factor of 5 is not met. Consequently, the cylinder wall thickness should be increased if an external pressure differential of 1 atm is anticipated.

The pressure vessel can be subjected to the screening criteria for fracture control. The cylinder and ellipsoidal end can be evaluated as follows:

$$\frac{K_{IC}}{\sigma} \geq 2\alpha\sqrt{t}$$

$$= \frac{27{,}000}{8736} = 3.09 \ , \tag{6.106}$$

$$2\alpha\sqrt{t} = 2(1.1)\sqrt{0.0625} = 0.55 \ . \tag{6.107}$$

Since $3.09 > 0.55$, the walls of the cylinder are not likely to fail from unstable flaw growth.

The end plate screening criteria can be applied as follows:

$$\frac{K_{IC}}{\sigma} = \frac{27{,}000}{13{,}200} = 2.04 \ , \tag{6.108}$$

$$2\alpha\sqrt{t} = 2(1.1)\sqrt{0.375} = 1.34 \ . \tag{6.109}$$

Again, since $2.04 > 1.34$, the end plates screen out of fracture mechanics. The conventional stress analysis for ductile materials should be adequate. It is interesting to note that an increase in the end plate thickness decreases the stress, yet the need for fracture analysis increases as the thickness increases.

As the risk of a failure increases, the need for a more sophisticated fracture analysis also increases.

6.4.2 Suspension System

A typical cryogenic system or dewar will consist of a cryogen supply tank and an instrument or sensor package that must be cooled. Both the supply tank and the sensor must be thermally isolated and yet mechanically suspended. Unfortunately, these two objectives counter each other. Thermal isolation typically requires long and thin structural suspension members to reduce the conduction of heat, whereas the pursuit of structural integrity encourages shorter and larger suspension systems. The trade-off for added structural capability is usually a reduction in hold time for the dewar. In dewar design, it is imperative that the load environment be realistically prescribed such that unnecessary so-called "structural conservatism" is avoided and cryogenic hold time is not sacrificed. Several structural configurations can be used to support tanks and sensors. In some designs, the tank has been supported by tension straps anchored to the outside vacuum shell of the dewar. Structurally, the tension straps are more efficient than suspension systems that must withstand bending stresses. The tension-only loading of the straps allows reduced cross sections and thereby increases the resistance to thermal conduction. The COBE dewar designed and built by Ball used tension straps for support of the supply tank. Sensors were then mounted to the tank directly to avoid heat exchangers, plumbing, and thermal conduction straps.

Suspension systems using concentric cylinders have been used frequently in dewars. Recent upper atmospheric experiments, such as CIRRIS-1A, SPIRIT II, and EXCEDE as designed by Utah State University, have used concentric cylinder suspension systems. SPIRIT III and CLAES as designed by Lockheed Missile & Space also use concentric cylinders. Obviously, a multiplicity of different suspension systems cannot be treated in detail herein. Since the concentric cylinder suspension structure is perhaps the most commonly used, it will be discussed in more detail.

A typical dewar schematic using concentric cylinders for a suspension system for both the supply tank and sensor is shown in Fig. 6.42. The outermost fiberglass cylinder is bonded to the base support structure and continues longitudinally to a "floating" aluminum ring. A second fiberglass cylinder is bonded to the same ring and extends in the reverse direction to a second "floating" aluminum ring. A third fiberglass cylinder extends downward longitudinally and connects to a center ring on the aluminum tank. The conduction heat path is lengthened by the use of multiple cylinders.

Suspension System Stresses and Displacements. The weights of the fiberglass cylinders are usually negligible when compared with the supply tank or sensor. The system can be modeled as an end weight on a cantilevered beam as shown in Fig. 6.43. The bending stresses from transverse accelerations usually dominate and provide the criteria for the structural design. The bending stress is maximum where the moment is maximum, which usually occurs at the base for either a single- or multiple-cylinder system. The bending stress at any location can be calculated using the conventional flexural stress formula:

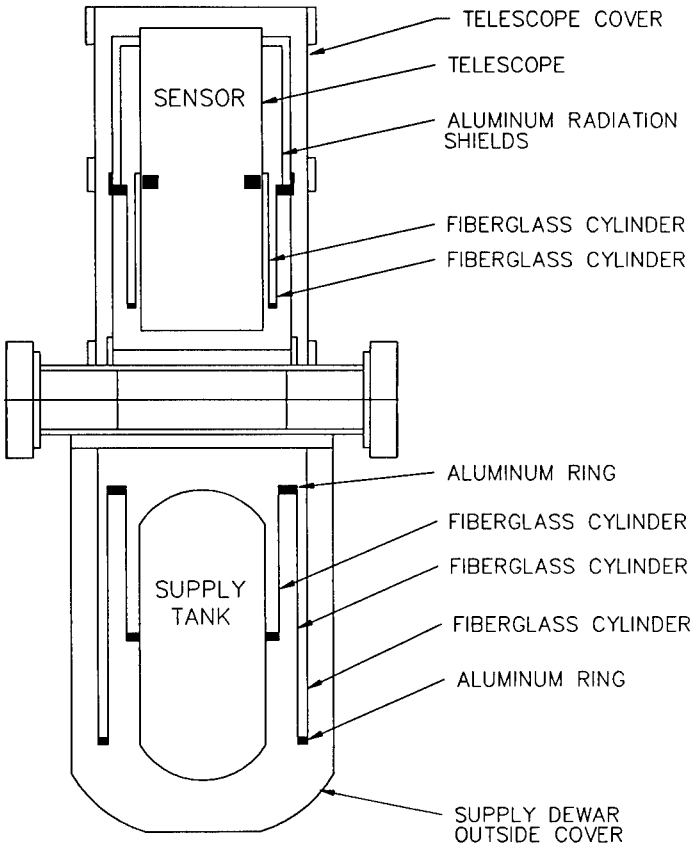


Fig. 6.42 CIRRIS-1A suspension system: concentric cylinders.

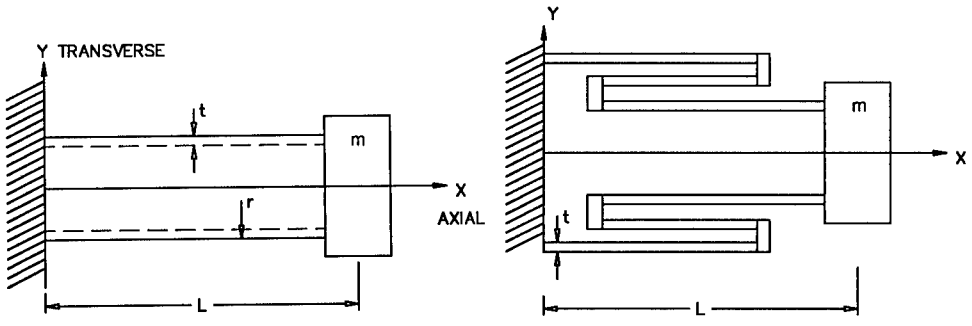


Fig. 6.43 Typical cantilevered cylinders.

$$\sigma = \frac{Mr}{I} = \frac{ma_y Lr}{\pi r^3 t} , \tag{6.110}$$

where M is the moment at the point of interest, r is the radius of the cylinder, and I is the moment of inertia. The transverse acceleration is A_T, the end mass is m, and the length of the cylinder from the base to the mass center of the end load is L.

The axial stress can be calculated as

$$\sigma = \frac{ma_x}{A} = \frac{ma_x}{2\pi rt} , \tag{6.111}$$

where a_x is the axial acceleration and A is the cross-sectional area. Usually, the axial stresses are negligible when compared to the bending stresses. Potential buckling of the thin-walled cylinders when subjected to combined axial and bending loads does present a potential failure mode that should be considered independently. The transverse end deflection can be calculated as

$$\delta = \frac{ma_y L^3}{3EI} , \tag{6.112}$$

where E is the modulus of elasticity. The axial deflection can be determined as

$$\delta = \frac{ma_x L}{AE} \tag{6.113}$$

and is usually negligible. The end deflection may be of significance with regard to optical alignment, radiation shield impacts, insulation short circuits, or calibration in vertical and horizontal planes in a gravity field.

Suspension System Vibrations. A continuous structure has an infinite number of natural frequencies and corresponding displacement patterns or mode shapes. The lower frequencies and mode shapes are usually of structural importance. The continuous structure may be modeled as a discrete mass system in order to approximate the lower frequencies. The fundamental frequency can be approximated by a single discrete mass and one equivalent spring. Usually, the damping is sufficiently low that the fundamental frequency can be determined without the complexity added by including damping coefficients. The intent of these hand calculations is to provide initial estimates on material dimensions in the iterative stages of design. Eventually, the frequency and mode shapes can be either measured or determined analytically from more sophisticated finite element models. The single degree of freedom discrete mass models are depicted in Fig. 6.44. The transverse fundamental frequency f_n of a cantilever beam of mass m_B with an end mass m_O can be estimated by assuming bending stiffness only as

$$f_n = \frac{1}{2\pi}\left(\frac{3EI}{m}\right)^{1/2} , \tag{6.114}$$

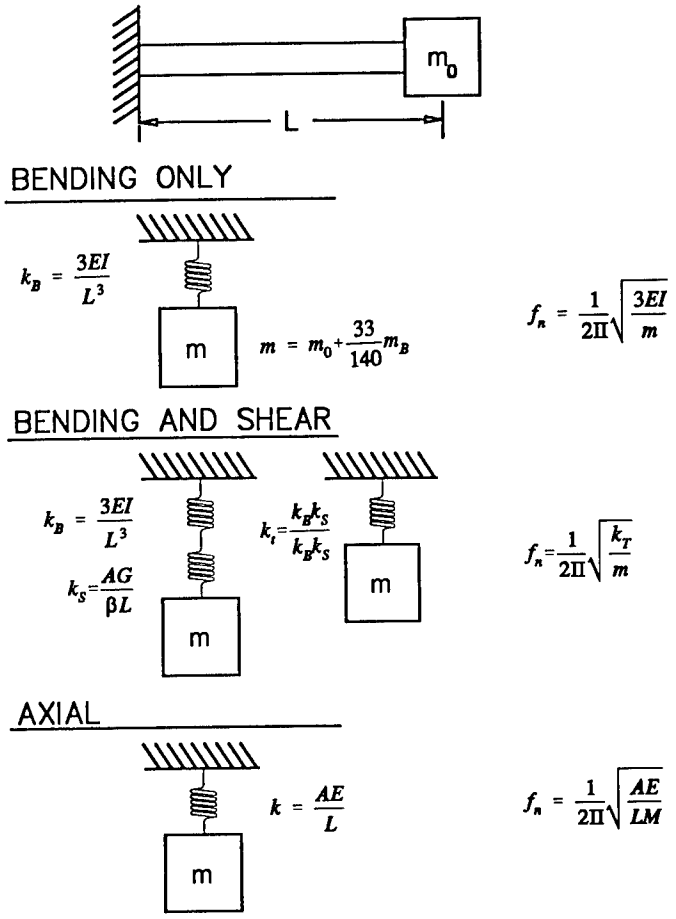


Fig. 6.44 Cantilever cylinder natural frequencies.

where $m = m_O + (33/140)\, m_B$, E is the modulus of elasticity, and I is the moment of inertia.

Suspension systems are often sufficiently short such that the shearing stiffness will affect the natural frequency. The bending spring constant $k_B = 3EI/L^3$ must be added in series with the shearing spring constant $k_S = AG/\beta L$ to obtain an equivalent spring $k_B k_S/(k_B + k_S)$. The constant β relates to geometry and is equal to 2.0 for a thin cylinder. The shear modulus is G. The improved transverse frequency estimate is then

$$f_n = \frac{1}{2\pi}\left(\frac{k_T}{m}\right)^{1/2}, \tag{6.115}$$

$$k_T = \frac{k_B k_S}{k_B + k_S} \,. \tag{6.116}$$

The axial natural frequency is usually significantly higher than transverse frequencies and is given as

$$f = \frac{1}{2\pi}\left(\frac{AE}{Lm}\right)^{1/2} \,. \tag{6.117}$$

Suspension Systems and Transmissibility. The stiffness and damping of a suspension system affect the displacements, accelerations, and forces that are transmitted. Expressions can be derived to describe the vibration environment of an instrument suspended from a base support. The key variables are the natural frequency of the suspension system, $\omega_n = \sqrt{k/m}$ and the harmonic forcing frequency of the excitation ω. The damping ratio $\zeta = c/2\omega_n$ is also included where c is the damping coefficient. Approximations can obviously be made by setting $\zeta = 0$ and evaluating an undamped system.

A suspension system fixed at the base with a harmonic forcing function applied to the mass is shown in Fig. 6.45. The equation of motion can be derived by summing forces and is given as

$$m\ddot{x} + c\dot{x} + kx = F \sin\omega t \,, \tag{6.118}$$

where $x(t)$ is the dynamic absolute displacement of the mass. The static displacement of the mass if the load F were placed on the spring is F/k. The dynamic amplification of the displacement can be shown to be[19]

$$\frac{X}{F_1/k} = \frac{1}{\left[\left(1 - \frac{\omega^2}{\omega_n^2}\right)^2 + \left(2\zeta\frac{\omega}{\omega_n}\right)^2\right]^{1/2}} \,, \tag{6.119}$$

where X is the maximum dynamic displacement of the mass.

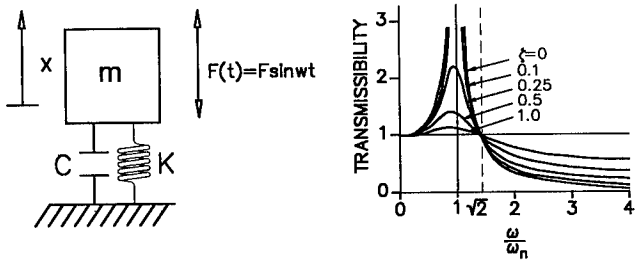


Fig. 6.45 Suspension transmissibility force applied to instrument mass.

The amplitude of the force transmitted through the damper and the spring to the base can be shown to be

$$F_{TR} = X(k^2 + c^2\omega^2)^{1/2} \ . \tag{6.120}$$

The force transmitted through the suspension system to the base is defined as the transmission ratio and can be derived as

$$TR = \left[\frac{1 + \left(2\zeta \frac{\omega}{\omega_n}\right)^2}{\left(1 - \frac{\omega^2}{\omega_n^2}\right)^2 + \left(2\zeta \frac{\omega}{\omega_n}\right)^2} \right]^{1/2} \ . \tag{6.121}$$

The transmission ratio TR is plotted as a function of frequency ratio in Fig. 6.45. Note that the forces could be axial or transverse, simply by specifying k as either the axial or transverse stiffness.

In cryogenic systems, the sensor is usually mounted to a base support that is subjected to external excitations. Figure 6.46 shows a discrete mass model for either axial or transverse vibrations of a sensor mounted to a base through a suspension system. The absolute displacement of the mass of the sensor is x_2, whereas the harmonic excitation applied to the base is $x_1(t)$. Applying Newton's second law, the equation of motion can be written[20]

$$m\ddot{x}_2 + c\dot{x}_2 + kx_2 = kx_1 + c\dot{x}_1 \ . \tag{6.122}$$

The ratio of the mass displacement over the base displacement or excitation is called the *amplitude ratio*:

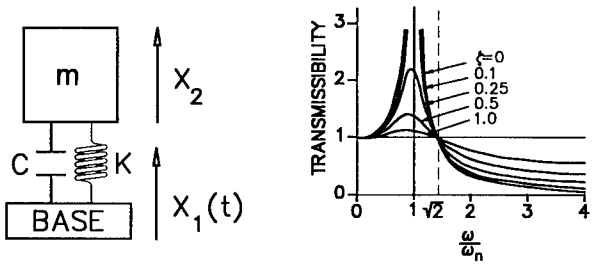


Fig. 6.46 Suspension transmissibility displacement/acceleration applied to base.

$$\frac{X_2}{X_1} = \frac{\left[1 + \left(2\zeta\frac{\omega}{\omega_n}\right)^2\right]^{1/2}}{\left[\left(1 - \frac{\omega^2}{\omega_n^2}\right)^2 + \left(2\zeta\frac{\omega}{\omega_n}\right)^2\right]^{1/2}} . \tag{6.123}$$

If the motion is harmonic, then $x_{2\max} = \omega^2 X_2$ and $x_{1\max} = \omega^2 X_1$, and it follows that

$$\frac{\ddot{x}_{2\max}}{\ddot{x}_{1\max}} = \frac{\omega^2 X^2}{\omega^2 X_1} = \frac{\left[1 + \left(2\zeta\frac{\omega}{\omega_n}\right)^2\right]^{1/2}}{\left[\left(1 - \frac{\omega^2}{\omega_n^2}\right)^2 + \left(2\zeta\frac{\omega}{\omega_n}\right)^2\right]^{1/2}} . \tag{6.124}$$

Therefore, the acceleration ratio is the same as the displacement ratio for a base excitation. The results are graphed in Fig. 6.46.

Once the acceleration of the mass is known, then the force transmitted F_T can be determined by multiplying by m. The force transmitted to the sensor from a base acceleration is then

$$\frac{F_{T\max}}{\ddot{x}_{1\max}} = \frac{m\ddot{x}_{2\max}}{\ddot{x}_{1\max}} = \frac{m\left[1 + \left(2\zeta\frac{\omega}{\omega_n}\right)^2\right]^{1/2}}{\left[\left(1 - \frac{\omega^2}{\omega_n^2}\right)^2 + \left(2\zeta\frac{\omega}{\omega_n}\right)^2\right]^{1/2}} . \tag{6.125}$$

The governing equations show that the vibration environment of the sensor and/or supply tank depends on the ratio of the natural and forcing frequencies. In particular, high-frequency high-forcing functions result in large ratios of ω/ω_n and are attenuated through the suspension systems. Consequently, a designer may reduce the acceleration loadings on an instrument by proper design of the natural frequency of the suspension systems. For example, if $\omega/\omega_n = 3$ and ζ were assumed to be zero, then the acceleration transmitted from the base to the instrument would be reduced by a factor of 0.125.

6.5 DESIGN LOADS

The load environment of a dewar must be defined in order to select materials and determine structural dimensions. The design of a dewar is often an iterative process, and the final loads are not precisely known until completion of the project. Nonetheless, equivalent static loads must be approximated early in the design process that will eventually envelop the resultant of all static-, sinusoidal-, random-, shock-, and acoustic-type forces. As mentioned, an overly conservative estimate of loads tends to increase wall thicknesses in the suspension system and the reduced thermal resistance reduces the cryogenic hold

time of the dewar. Frequently, the required acceptance and qualification tests set the load maximums. Conservatism in the test specifications, coupled with margins of safety in the structural design, contribute to structural integrity, but the reduced thermal efficiency pays the penalty.

6.5.1 Static Loads

The forces placed on a dewar system due to mass acted on by gravity are small compared to the dynamic loads. Dead weight forces may be negligible with regard to material stress; however, they may influence calibration of a sensor, such as a telescope that includes optical mirrors. A designer should avoid structural configurations that result in displacements under gravity loads that are difficult to predict and/or that are nonrepeatable.

The structural design of a dewar is based on equivalent static loads. These loads are estimated by multiplying the component accelerations in mutually orthogonal directions by the total mass of a structured element. Since the actual mass of a component is distributed over a region, the point of application for the lumped mass force must be selected with proper judgment. Placement of the equivalent static force at the mass center of structural components is frequently done.

6.5.2 Harmonic Loads

The actual working load environment may include steady-state, periodic, and aperiodic accelerations. The most likely exposure to harmonic accelerations, however, is during acceptance and/or qualification testing. A common approach is to determine natural frequencies by using a sine sweep at an acceleration level of 1 g or less. Structural testing is then done by subjecting the cryogenic system or individual components to a sine sweep or a sine dwell test at a higher acceleration. Since a sine sweep passes through a spectrum of frequencies, resonances will repeatedly occur and localized accelerations on system components significantly greater than anticipated may occur due to transmissibility. For example, a sine dwell at 6 g applied to the base of a suspension system may result in multiples of 6 g on the sensor at near resonance frequencies. It is more likely that the equivalent static loads will envelop the loads induced by a sine dwell test. If a dwell frequency different than a natural frequency is selected, then the localized amplifications due to resonance will not provide unexpected severe environments. A sine dwell test also allows distributed mass loading to occur compared to concentrated point loads on a static test.

The designer of the system must anticipate the type of harmonic testing, if any, that will be placed on the dewar. The possibility of dynamic load amplifications relating to transmissibility may well result in accelerations that exceed the static design loads and cause failure.

6.5.3 Random Loads

Vibrations that are not periodic and have no repeatable pattern to amplitude or frequency are called *random vibrations*. The response of a structure subjected to random vibrations must be treated using statistical methods. The purpose

of this section is to provide a workable approach to allow one to calculate an equivalent static load for a specified random environment.

A typical random vibration environment is provided in Fig. 6.47. The excitation is called a power spectral density (PSD) curve, even though spectral density would be more descriptive. The PSD level is plotted as a function of frequency. The units of the PSD curve are usually g^2/Hz, however, other units are possible.

A random excitation that is constant for all frequencies such that the PSD curve is flat is called *white noise*. The spectral density is a logarithmic plot. The specific PSD level for a given frequency on an upslope or a downslope must be calculated using equations for log-log curves. The PSD level for an upslope is given as

$$\text{PSD}_2 = \text{PSD}_1\left(\frac{f_2}{f_1}\right)^m , \tag{6.126}$$

where PSD_1 is the spectral density at the beginning of the upslope at frequency f_1, f_2 is the frequency where PSD_2 is required, and m is the product of 0.3322 times the upslope in units of dB/octave.

The PSD level for a downslope can be calculated from

$$\text{PSD}_3 = \text{PSD}_2\left(\frac{f_2}{f_3}\right)^m , \tag{6.127}$$

where PSD_2 is the spectral density at the beginning of the downslope at frequency f_2. The coefficient $m = 0.3322$ times the downslope, where the downslope is expressed in dB/octave.

Using the PSD curve provided in Fig. 6.47, the PSD value on an upslope at a frequency of 35 Hz could be calculated as

$$\text{PSD}_2 = \text{PSD}_1(f_2/f_1)^m \quad \text{for } m = 0.3322(12) = 3.986 , \tag{6.128}$$

$$\text{PSD} = 0.01(35/20)^{3.986} = 0.093\ g^2/\text{Hz} , \tag{6.129}$$

and on the downslope at $f = 1500$ Hz,

$$\text{PSD} = 0.3(1390/1500)^{3.986} = 0.2214\ g^2/\text{Hz} . \tag{6.130}$$

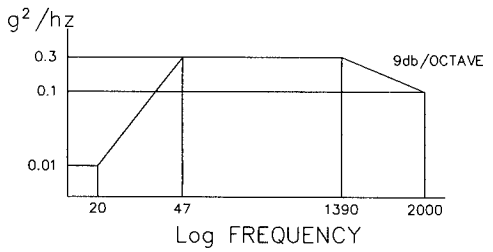


Fig. 6.47 Typical acceleration spectral density.

A term referred to as root mean square (rms) acceleration is defined by multiplying the PSD by the frequency bandwidth and taking the square root. The rms acceleration is the square root of the area under the PSD curve. In the case of white noise, the same rms acceleration could be obtained by a large PSD and a short frequency interval or a small PSD and a large frequency interval. The rms acceleration should not be used to estimate an equivalent static load. For example, a white noise PSD of 0.1 g^2/Hz over a frequency range of 90 Hz would result in 3 g rms. The same rms acceleration would correspond to only 0.0045 g^2/Hz over a frequency band of 2000 Hz. Although the rms acceleration would be the same, the larger PSD would result in a much larger equivalent static force.

An equivalent static force can be approximated that represents the effect of a random vibration on an instrument. In Fig. 6.48, an instrument is placed on a shaker table and subjected to a random vibration defined by some PSD curve. An acceleration that represents the combined response of the mass to the random vibration can be determined. The effective force can then be determined using Newton's law.

The response acceleration for a single degree of freedom model subjected to a random excitation at its base is provided by the so-called "Miles equation" as

$$R = \lambda\left[\frac{\Pi}{2} f_n(\text{PSD})Q\right]^{1/2}, \tag{6.131}$$

where λ is a probability factor, f_n is the fundamental frequency of the model, PSD is the power spectral density (g^2/Hz) at the fundamental frequency, and Q is the dynamic magnification factor at resonance ($2\delta = 1/Q$). If $\lambda = 3$, then there is only a 0.3% probability of the response exceeding R. If $\lambda = 2$, the probability of exceeding R is 4.6%, and at $\lambda = 1$ the probability of exceeding R is 31.7%. Typical values of Q commonly used are in the range of 10 to 25. The Miles equation is usually quite conservative. A paper by Thampi and Vidyasagar[21] discusses the results of computer models versus the Miles equation.

The acceleration response computed from the Miles equation can then be used with Newton's law to calculate an equivalent force to be placed on the structure. This force should then be combined with other forces as necessary.

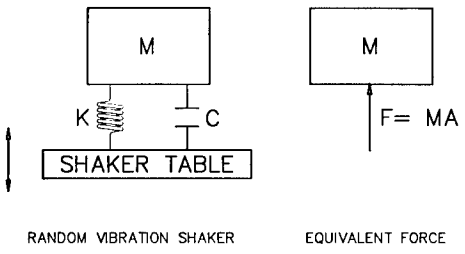


Fig. 6.48 Equivalent shaker force from random vibration.

6.5.4 Random Force Applications

The application of the Miles equation will be demonstrated. Assume that a power spectral density spectrum is given as follows:

Frequency	Level
20 Hz	0.02 g^2/Hz
20–150	+3 dB/Oct
150–190	0.15 g^2/Hz
9000–2000	−12 dB/Oct

Also, assume that $\lambda = 3$, $Q = 10$, and $f_n = 100$. The response acceleration can then be completed as follows:

$$R = \lambda\left[\left(\frac{\Pi}{2}\right)(f_n)(\text{PSD})Q\right]^{1/2} . \tag{6.132}$$

Since the natural frequency is on the upslope of the PSD curve, the PSD level at f_n can be computed as

$$\text{PSD}_2 = \text{PSD}_1\left(\frac{f_2}{f_1}\right)^m \qquad \text{for } m = 0.3322(3) = 1.0$$

$$= 0.02(100/20)^{1.0} = 0.1\ g^2/\text{Hz} . \tag{6.133}$$

Upon substitution, the response acceleration follows as

$$R = 3\left[\left(\frac{\Pi}{2}\right)(100)(0.1)(10)\right]^{1/2} = 37.6\ g . \tag{6.134}$$

This response acceleration can be used with Newton's law to calculate an equivalent static force that would act on the instrument subjected to the random vibration. In other words, the random vibration environment applies an equivalent static load equal to the product of mass times acceleration, where the acceleration is R.

References

1. *Thermophysical Properties of Matter—The TPRC Data Series,* Vol. 1, CINDAS/Purdue University, West Lafayette, IN (1970).
2. R. Barron, *Cryogenic Systems,* Series in Mechanical Engineering, p. 469, McGraw-Hill, New York (1966).
3. *Thermophysical Properties of Matter—The TPRC Data Series,* Vol. 4, CINDAS/Purdue University, West Lafayette, IN (1971).
4. *Thermophysical Properties of Matter—The TPRC Data Series,* Vol. 12, CINDAS/Purdue University, West Lafayette, IN (1975).
5. W. Obert, J. R. Coupland, D. P. Hammond, T. Cook, and K. Harwood, "Emissivity measurements of metallic surfaces used in cryogenic applications," JET Joint Undertaking, Abingdon, Oxon, UK (1988).
6. G. A. Bell, T. C. Nast, and R. K. Wedel, "Thermal performance of multilayered insulation applied to small cryogenic tankage," in *Advances in Cryogenic Engineering,* K. D. Timmerhaus, R. P. Reed, and A. F. Clark, Eds., Vol. 22, pp. 272–282, Plenum Press, New York (1977).

7. S. L. Bapat, K. G. Narayankhedkar, and T. P. Lukose, "Experimental investigators of multilayer insulation," *Cryogenics* **30,** 700–719 (Aug. 1990).
8. V. D. Arp and R. D. McCarty, "Thermophysical properties of helium-4 from 0.8 to 1500 K with pressure to 2000 MPa," NIST Technical Note 1334, National Technical Information Service, Springfield, VA (Nov. 1989).
9. S. Bard, "Development of a high-performance cryogenic radiator with V-groove radiation shields," *Journal of Spacecraft* **24**(3), 193–197 (May–June 1987).
10. "Mechanical cryogenic coolers for space applications," Company Brochure, British Aerospace plc, Bristol, UK (n.d.).
11. "Single-stage turbo cooler," photocopy of overhead (n.d.).
12. E. I. Mikulin, A. A. Tarasov, and M. P. Shrebyonock, "Low temperature expansion pulse tubes," *Advances in Cryogenic Engineering,* R. W. Fast, Ed., Vol. 29, pp. 629–637, Plenum Press, New York (1984).
13. A. H. Burr, *Mechanical Analysis and Design,* Elsevier Science Publishing, New York (1981).
14. S. Timoshenko and Woinowsky-Krieger, *Theory of Plates and Shells,* McGraw-Hill, New York (1959).
15. J. F. Harvey, *The Theory and Design of Pressure Vessels,* Van Nostrand Reinhold, New York (1985).
16. D. F. Windenburg, "Vessels under external pressure," *Mechanical Engineering* (Aug. 1937).
17. "Fracture controls requirements for DOD shuttle payloads SD," YV-0068, U.S. Air Force Space Transportation (1981).
18. "Comparison of screening methods for determining safe-life in fracture control of space transportation systems," memorandum 731-0004-83, 6SFC.
19. R. F. Steidel, *An Introduction to Mechanical Vibrations,* John Wiley & Sons, New York (1989).
20. F. S. Tse, I. E. Morse, and R. T. Hinkle, *Mechanical Vibrations Theory and Applications,* Allyn and Bacon, Needham Heights, MA (1978).
21. S. K. Thampi and S. V. Vidyasagar, "Random vibration analysis of space flight hardware using NASTRAN," GE Government Services.

Bibliography

Thermal Properties

Barron, R., *Cryogenic Systems,* Series in Mechanical Engineering, p. 469, McGraw-Hill, New York (1966).

Batty, J. C., *Thermal Conductivity and Thermal Expansion Data for Cryogenic Solids,* n.p. (n.d.), Department of Mechanical and Aerospace Engineering, Utah State University.

Kasen, M. B., G. R. MacDonald, D. H. Beekman, Jr., and R. E. Schram, "Mechanical, electrical and thermal characterization of G-10CR and G-11CR glass-cloth/epoxy laminates between room temperature and 4 K," National Bureau of Standards, Boulder, CO.

Khalil, A., and K. S. Han, "Mechanical and thermal properties of glass-fiber reinforced composites at cryogenic temperatures," in *Advances in Cryogenic Engineering,* R. P. Reed and A. F. Clark, Eds., Vol. 28, pp. 243–251, Plenum Press, New York (1981).

Touloukian, Y. S., R. W. Powell, C. Y. Ho, and P. G. Klemens, "Thermal conductivity—metallic elements and alloys," in *Thermophysical Properties of Matter—The TPRC Data Series,* Vol. 1, CINDAS/Purdue University, West Lafayette, IN (1970).

Touloukian, Y. S., and E. H. Buyco, "Specific heat—metallic elements and alloys," in *Thermophysical Properties of Matter—The TPRC Data Series,* Vol. 4, CINDAS/Purdue University, West Lafayette, IN (1971).

Touloukian, Y. S., R. K. Kirby, R. E. Taylor, and P. D. Desal, "Thermal expansion—metallic elements and alloys," in *Thermophysical Properties of Matter—The TPRC Data Series,* Vol. 12, CINDAS/Purdue University, West Lafayette, IN (1975).

Vendell, E. W., D. G. Frodsham, and O. Lin, "A supercritical helium cooling system for a shuttle-borne infrared telescope," *Proceedings of 1983 Space Helium Dewar Conference,* 15–28 (1983).

Radiating Surface Properties

Anderson, C. C., and M. M. Haflar, "Calorimetric measurements of thermal control surfaces at geosynchronous orbit," *Journal of Thermophysics* **2**(2), 145–150 (Apr. 1988).

Babelot, J. F., and M. Hoch, "Investigation of spectral emissivity data of some metals and nonmetals in the wavelength range 400–15000 nm, and their total emissivity," *High Temperatures High Pressures* **21,** 79–84 (1989).

Corlett, R. C., "Direct Monte Carlo calculation of radiative heat transfer in vacuum," *Journal of Heat Transfer* **88,** 376–382 (Nov. 1966).

Emery, A. F., and A. Abrous, "Effects of specularly reflected radiation on spacecraft temperatures and thermal gradients," *Journal of Spacecraft and Rockets,* **24,** 122–126 (Mar.–Apr. 1987).

Feingold, A., and K. G. Gupta, "New analytical approach to the evaluation of configuration factors in radiation from spheres and infinitely long cylinders," *Journal of Heat Transfer* **92,** 69–76 (Feb. 1970).

Hesketh, P. J., B. Bebhart, and J. N. Zemel, "Measurements of the spectral and directional emission from microgrooved silicon surfaces," *Journal of Heat Transfer* **1109,** 680–686 (Aug. 1988).

Obert, W., J. R. Coupland, D. P. Hammond, T. Cook, and K. Harwood, "Emissivity measurements of metallic surfaces used in cryogenic applications," JET Joint Undertaking, Abingdon, Oxon, UK (1988).

Turner, W. D., and T. J. Love, "Directional emittance of a two-dimensional ceramic coating," *AIAA Journal* **9**(9), 1849–1853 (Sep. 1971).

Zaitsev, V. A., I. V. Gorbatenko, and M. A. Taimarov, "Emissivity of steels and alloys in the spectral region 2–13 μm," translated from *Inzhenerno-Fiziches Kii Zhurnal* **50**(4), 620–625 (Apr. 1986).

Convection

Bishop, E. H., "Heat transfer by natural convection of helium between horizontal isothermal concentric cylinders at cryogenic temperature," *Journal of Heat Transfer* **110,** 109–115 (Feb. 1988).

Clausin, A. M. "Convective heat transfer research using a cryogenic enviroment," *Cryogenic* **30,** 335–340 (Apr. 1990).

Lipkea, William H., and George S. Springer, "Heat transfer through gases contained between two vertical cylinders at different temperatures," *International Journal of Heat Mass Transfer* **11,** 1341–1350 (1968).

McCarthy, J. R., and H. Wolf, "Forced convection heat transfer to gaseous hydrogen at high heat flux and high pressure in a smooth, round, electrically heated tube," *ARS Journal* **30**(4), 423–425 (Apr. 1960).

McLead, Andrew E., and Eugen H. Bishop, "Turbulent natural convection of gases in horizontal cylindrical annuli at cryogenic temperatures," *International Journal of Heat and Mass Transfer* **32**(10), 1967–1978 (1989).

Ogata, H., and S. Sato, "Measurements of forced convection heat transfer to supercritical helium," Central Research Laboratory, Hitachi Ltd., Tokyo.

Taylor, Maynard F., "Experimental local heat-transfer and average friction data for hydrogen and helium flowing in a tube at surface temperatures up to 5600°R," NASA TN D-2280, National Aeronautics and Space Administration, Washington, DC (Apr. 1964).

Tsao, M. K., "Temperature distribution and power loss of a gas-cooled support for cryogenic container," *Cryogenics* **14**(5), 271–275 (May 1974).

Multilayer Insulation

Adelberg, M., "Effective thermal conductivity and multilayered insulation," in *Advances in Cryogenic Engineering,* K. D. Timmerhaus, Ed., Vol. 12, pp. 272–282, Plenum Press, New York (1960).

Bapat, S. L., K. G. Narayankhedkar, and T. P. Lukose, "Experimental investigators of multilayer insulation," *Cryogenics* **30,** 711–719 (Aug. 1990).

Bapat, S. L., K. G. Narayankhedkar, and T. P. Lukose, "Performance prediction of multilayer insulation," *Cryogenics* **30,** 700–710 (Aug. 1990).

Bell, G. A., T. C. Nast, and R. K. Wedel, "Thermal performance of multilayered insulation applied to small cryogenic tankage," in *Advances in Cryogenic Engineering,* K. D. Timmerhaus, R. P. Reed, and A. F. Clark, Eds., Vol. 22, pp. 272–282, Plenum Press, New York (1977).

Black, I. A., A. A. Fowle, and P. E. Glaser, "Development of high-efficiency insulation," in *Ad-*

vances in Cryogenic Engineering, K. D. Timmerhaus, Ed., Vol. 5, pp. 181–188, Plenum Press, New York (1959).

Halaczek, T. L., and J. Rafalowicz, "Heat transport in self-pumping multilayer insulation," *Cryogenics* **26,** 372–376 (June 1986).

Hnilicka, M. P., "Engineering aspects of heat transfer in multilayer reflective insulation and performance of NRC insulation," *Advances in Cryogenic Engineering,* Vol. 5, pp. 199–208, Plenum Press, New York (1959).

Kropschot, R. H., J. E. Schrodt, M. M. Fulk, and B. J. Hunter, "Multiple-layer insulation," in *Advances in Cryogenic Engineering,* Vol. 5, pp. 189–198, Plenum Press, New York (1959).

Lin, H. C., "A thermal model of a shuttle-borne helium-cooled infrared telescope," MS thesis, Utah State University (1985).

Matsuda, A., and H. Yoshikiyo, "Simple structure insulating material properties for multilayer insulation," *Cryogenics* **20,** 135–138 (Mar. 1980).

Mikhalchenko, R. S., V. F. Getmanets, N. P. Pershin, and Yu. V. Butozskii, "Study of heat transfer in multilayer insulations based on composite spacer materials," *Cryogenics* **23,** 309–311 (June 1983).

Mikhalchenko, R. S., V. F. Getmanets, N. P. Pershin, and Yu. V. Butozskii, "Theoretical and experimental investigations of radiative heat transfer in multilayer insulations," *Cryogenics* **25,** 275–278 (May 1985).

Scurlock, R. G., and B. Saull, "Development of multilayer insulations with thermal conductivities below 0.1 μWcm^1K^1," *Cryogenics* **16,** 303–311 (May 1976).

Shu, Q. S., "Systematic study to reduce the effects of cracks in multilayer insulation, part 1: theoretical model," *Cryogenics* **27,** 249–256 (May 1987).

Shu, Q. S., R. W. Fast, and H. L. Hart, "Systematic study to reduce the effects of cracks in multilayer insulation, part 2: experimental results," *Cryogenics* **27,** 298–311 (June 1987).

Siahpush, S. A., "Solar radiation trapping in multilayer insulation blankets," MS thesis, Utah State University (1987).

Tien, C. L., and G. R. Cunnington, "Cryogenic insulation heat transfer," *Advances in Heat Transfer* **9,** 349–417 (1973).

Solid Cryogen Coolers

"Cryogenics," company brochure, Ball Aerospace Systems Group (1988).

Donabedian, Martin, "Cooling systems," Chap. 15 in *The Infrared Handbook,* W. L. Wolfe, G. J. Zissis (Eds.), Environmental Research Institute of Michigan, Ann Arbor, MI (Revised 1985).

Naes, L. G., W. J. Horsley, C. S. Ngai, D. C. Read, and T. C. Nast, "Design and performance analysis of the CLAES Ne/CO_2 cryostat," *Proceedings of the SPIE* **973,** 369–377 (1988).

Nakano, G. H., L. F. Chase, J. R. Kilner, W. G. Sandie, G. J. Fishman, W. S. Paciesas, R. E. Lingenfelter, and S. E. Woosley, "SONGS—a high resolution imaging gammaray spectrometer for the space station," *Proceedings of the SPIE* **1159,** 165–176 (1989).

Nast, T. C., "Status of solid cryogen coolers," in *Advances in Cryogenic Engineering,* R. W. Fast (Ed.), Vol. 31, pp. 835–849, Plenum Publishing, New York (1985).

"Proposal for SPIRIT III cryogenic support system sensor, Vol. I technical," submitted to Space Dynamics Laboratory, Utah State University, Lockheed Missiles & Space Company (1989). Available from SDL MSX Library, Logan, UT.

Thermoacoustic Oscillation

Daney, D. E., P. R. Ludtke, and M. C. Jones, "An experimental study of thermally-induced flow oscillations in supercritical Helium," *Journal of Heat Transfer,* **101**(1), 9–14 (Feb. 1979).

Gorbachev, S. P., A. L. Korolev, V. K. Mafyushchenkov, and V. A. Sysoev, "Simple method for elimination of thermoacoustic oscillations in cryogenic tubes," Balashikha Scientific-Industrial Organization of Cryogenic Machine Construction, translated from *Pribory, Tekhnika Eksperimenta,* No. 1, pp. 220–221 (Jan–Feb. 1986).

Haycock, R. H., "Techniques for eliminating thermal-acoustical oscillations in cryogenic instrumentation," *Proceedings of the SPIE* **245,** 143–148 (1980).

Merkli, P., and H. Thomann, "Thermoacoustic effects in a resonance tube," *Journal of Fluid Mechanics* **70** (part 1), 161–177 (1975).

Ozoe, H., N. Safo, and S. W. Churchill, "The effect of various parameters on thermoacoustic convection," *Chemical Engineering Communication* **5,** 203–221 (1980).

Rott, N., "Thermoacoustics," *Advances in Applied Mechanics* **20,** 135–175 (1980).

Yazaki, T., A. Tominaga, and Y. Narahara, "Large heat transport due to spontaneous gas oscillation induced in a tube with steep temperature gradients," *Journal of Heat Transfer* **105,** 889–889 (Nov. 1983).

Radiator

Bard, Steven, "Advanced passive radiator for spaceborne cryogenic cooling," *Journal of Spacecraft* **21,** 150–155 (Mar.–Apr. 1984).

Bard, Steven, "Development of a high-performance cryogenic radiator with V-groove radiation shields," *Journal of Spacecraft* **24**(3), 193–197 (May–June 1987).

Bobco, R. P., and B. L. Drolen, "Engineering model of surface specularity: spacecraft design implications," *Journal of Thermophysics* **3**(3), 289–296 (July 1989).

"Current development in NASA cryogenic cooler technology," in *Advances in Cryogenic Engineering,* R. W. Fast, Ed., Vol. 33, pp. 800–801, Plenum Press, New York (1987).

Furukawa, Masao, "Design and off-design performance calculations of space radiators," *Journal of Spacecraft* **18**(6), 515–519 (Nov.–Dec. 1981).

"History, status, and future applications of spaceborne cryogenic systems," in *Advances in Cryogenic Engineering,* Vol. 27, pp. 1008–1011, Plenum Press, New York (1981).

Rutledge, S. K., D. L. Motes, and P. E. Paulsen, "The effects of atomic oxygen on the thermal emittance of high temperature radiator surfaces," NASA Lewis Research Center, Cleveland, OH.

Salazar, R. P., and N. Evans, "A study of a 65K radiative cooler for the advanced moisture and temperature sounder," paper presented at AIAA 16th Thermophysics Conference, Palo Alto, California, June 23–25, paper AIAA 81–1101 (1981).

Wright, J. P., "Development of a 5 watt 70°K passive radiator," paper presented at AIAA 15th Thermophysics Conference, Snowmass, Colorado, July 14–16, 1980, paper AIAA-80-1512.

Refrigerators

"Ball cryocoolers," company brochure, Ball Corporation, Boulder, CO (n.d.).

Barclay, J. A., W. F. Stewart, W. C. Overton, R. J. Candler, and O. D. Harkleroad, "Experimental results on a low-temperature magnetic refrigerator," in *Advances in Cryogenic Engineering,* R. W. Fast (Ed.), Vol. 31, pp. 743–752, Plenum Press, New York (1985).

Bradhsaw, T. W., J. Delderfield, S. T. Werrett, and G. Davey, "Performance of the Oxford miniature Stirling cycle refrigerator," in *Advances in Cryogenic Engineering,* R. W. Fast (Ed.), Vol. 31, pp. 801–809, Plenum Press, New York (1985).

British Aerospace plc, "Mechanical cryogenic coolers for space applications," company brochure, Bristol, UK (n.d.).

"Creare single stage reverse brayton," overhead, n.p. (n.d.).

David, Marc, and Jean-Claude Maréchal, "How to achieve the efficiency of a Gifford-MacMahon cryocooler with a pulse tube refrigerator," *Cryogenics* **30** (supplement), 262–265 (Sep. 1990).

Forth, H. J., R. Heisig, and H. H. Klein, "Gifford-McMahon refrigerator with split cold head," in *Refrigeration for Cryogenic Sensors: Proceedings of the Second Biennial Conference on Refrigeration for Cryogenic Sensors and Electronic Systems,* NASA Goddard Space Flight Center, Greenbelt, Maryland, December 7–8, 1982, Max Gasser (Ed.), pp. 305–313, National Aeronautics and Space Administration, Washington, DC (1982).

Fujita, T., T. Ohtsuka, and Y. Ishizaki, "Japanese activities in refrigeration technology," in *Refrigeration for Cryogenic Sensors: Proceedings of the Second Biennial Conference on Refrigeration for Cryogenic Sensors and Electronic Systems,* NASA Goddard Space Flight Center, Greenbelt, Maryland, December 7–8, 1982, Max Gasser (Ed.), pp. 33–46, National Aeronautics and Space Administration, Washington, DC (1982).

Garrett, Steven, "Cool sound," *Discover* **11,** 25 (Dec. 1990).

Gresin, A. K., Y. O. Prousman, A. V. Smirnov, and L. A. Babi, "Analytical and experimental study of the small Stirling Cryocoolers," *Cryogenics* **30** (supplement), 241–246 (Sep. 1990).

Jones, J. A., "$LaNi_5$ hybride cryogenic refrigerator test results," in *Refrigeration for Cryogenic*

Sensors: Proceedings of the Second Biennial Conference on Refrigeration for Cryogenic Sensors and Electronic Systems, NASA Goddard Space Flight Center, Greenbelt, Maryland, December 7–8, 1982, Max Gasser (Ed.), pp. 357–373, National Aeronautics and Space Administration, Washington, DC (1982).

Jones, J. A., "Sorption cryogenic refrigeration—status and future," in *Advances in Cryogenic Engineering,* R. W. Fast (Ed.), Vol. 33, pp. 869–878, Plenum Press, New York (1987).

Jones, J. A., "Status of sorption cryogenic refrigeration," JPL invention report NPO-17349/6859, Jet Propulsion Laboratory, Pasadena, CA (1988).

Jones, J. A., S. Bard, H. R. Schember, and J. Rodriguez, "Sorption cooler technology development at JPL," *Cryogenics* **30,** 239–245 (Mar. 1990).

Keung, C., P. J. Patt, M. Starr, and R. McFarlane, "Performance of a prototype, 5 year lifetime, Stirling cycle refrigerator for space applications," paper presented at 6th International Cryocooler Conference, Plymouth, Massachusetts, October 25–26, 1990.

Kuriyama, T., R. Hakamada, H. Nakagome, Y. Tokai, M. Sahashi, O. Horigami, R. Li, O. Yoshida, K. Matsumoto, and T. Hashimoto, "Development of a 5K GM refrigerator using rare earth compounds as regenerator matrix," in *Cryogenics and Refrigeration: Proceedings of International Conference,* Zhejiang University, Hangzhou, China, May 22–26, 1989, Chan Goubang, Thomas M. Flynn (Eds.), pp. 91–96, Pergamon Press, New York (1989).

Lindale, E., and D. Lehrfeld, "Life test performance of a Philips rhombic-drive refrigerator with bellows seals," in *Refrigeration for Cryogenic Sensors: Proceedings of the Second Biennial Conference on Refrigeration for Cryogenic Sensors and Electronic Systems,* NASA Goddard Space Flight Center, Greenbelt, Maryland, December 7–8, 1982, Max Gasser (Ed.), pp. 197–213, National Aeronautics and Space Administration, Washington, DC (1982).

"Linear resonant split Stirling cooler MX 7043," company brochure, Magnavox Government and Industrial Electronics Company.

Ludwigsen, Jill, and Mark Fraser, "SPAS III tactical cooler survey," Nichols Research Corporation, Albuquerque, NM (1991).

"Mechanical cryogenic cooling systems for space applications," Company Brochure, Lucas Aerospace Limited and Lockheed Missiles & Space Company, Palo Alto, CA (1990).

Orlowska, A. H., T. W. Bradshaw, and J. Hieatt, "Closed cycle coolers for temperatures below 30K," *Cryogenics* **30,** 246–248 (Mar. 1990).

Paugh, Robert L., "New class of microminiature Joule-Thomson refrigerator and vacuum package," *Cryogenics* **30,** 1079–1083 (Dec. 1990).

Peterson, I., "A new, sound way to refrigerate," *Science News* **122,** 358 (Dec. 4, 1982).

Price, Kenneth, "A nonwearing, low vibration, 65 K Stirling cryocooler," in *Proceedings of the Fourth Interagency Meeting on Cryocoolers,* Plymouth, Massachusetts, October 24, 1990, Geoffrey Green, Margaret Knox (Eds.), pp. 255–261, David Taylor Research Center, Bethesda, MD (1991).

Radebaugh, Ray, James Zimmerman, David R. Smith, and Beverly Louie, "A comparison of three types of pulse tube refrigerators: new methods for reaching 60 K," in *Advances in Cryogenic Engineering,* R. W. Fast (Ed.), Vol. 31, pp. 779–789, Plenum Press, New York (1985).

Ravex, A., G. Claudet, and P. Rolland, "A Vuilleumier refrigerator for long-life spaceborne applications," *Cryogenics* **30** (supplement), 277–281 (Sep. 1990).

"Ricor Micro IDCA Cryocooler Model K506B," Company Brochure, Kollmorgen Corporation (1990).

Riggle, Peter, "Stirling cryocooler with extremely low vibration SBIR, phase II final Report," prepared for NASA Goddard Research Center, Greenbelt, MD (Mar. 1992).

Russo, S. C., "Stirling cycle 65 K standard spacecraft cryocooler development," in *Proceedings of the Fourth Interagency Meeting on Cryocoolers,* Plymouth, Massachusetts, October 24, 1990, Geoffrey Green, Margaret Knox (Eds.), pp. 235–251, David Taylor Research Center, Bethesda, MD (1991).

Sherman, Allen, "History, status, and future applications of spaceborne cryogenic systems," in *Advances in Cryogenic Engineering,* R. W. Fast (Ed.), Vol. 27, pp. 1007–1029, Plenum Press, New York (1982).

Shimko, Martin, Creare, Inc., Hanover, NH, private communication (Nov. 6, 1990).

Shimko, Martin A., W. Dodd Stacy, and John A. McCormick, "Stirling cryocooler test results and design model verification," in *Proceedings of the 25th Intersociety Energy Conversion Engineering Conference,* Reno, Nevada, August 12–17, 1990, Paul A. Nelson, William W. Schertz, Russel H. Till (Eds.), American Institute of Chemical Engineers, New York (1990).

Smith, J. L., Jr., G. Y. Robinson, and Y. Iwasa, "Survey of the state-of-the-art of miniature cryocoolers for superconductive devices," NRL memorandum 5490, Naval Research Laboratory, Arlington, VA (1984).

Stolfi, F., M. Goldowsky, J. Ricciardelli, and P. Shapiro, "A magnetically suspended linearly driven cryogenic refrigerator," in *Refrigeration for Cryogenic Sensors: Proceedings of the Second Biennial Conference on Refrigeration for Cryogenic Sensors and Electronic Systems,* NASA Goddard Space Flight Center, Greenbelt, Maryland, December 7–8, 1982, Max Gasser (Ed.), pp. 263–303, National Aeronautics and Space Administration, Washington, DC (1982).

Swift, G. W., "Thermoacoustic engines," *Journal of the Acoustic Society of America,* **84,** 1145–1180 (Oct. 1988).

Thirumaleshwar, M., and R. M. Pandey, "Two stage Gifford-McMahon cycle cryorefrigerator operated by gas balancing principle," *Cryogenics* **30**, 100–104 (Feb. 1990).

Timbie, P. T., G. M. Bernstein, and P. L. Richards, "An adiabatic demagnetization refrigerator for SIRTF," *IEEE Transactions on Nuclear Science* **36,** 898–902 (Feb. 1989).

"Ultra-reliable split mini-cooler MX 7043 series," application notes, Magnavox Government and Industrial Electronics Company (1988).

Wang, Junjie, WenXiu Zhu, Pingsheng Zhang, and Yuan Zhou, "A compact co-axial pulse tube refrigerator for practical application," *Cryogenics* **30** (supplement), 267–271 (Sep. 1990).

Werrett, S. T., G. D. Peskett, G. Davey, T. W. Bradshaw, and J. Delderfield, "Development of a small Stirling cycle cooler for spaceflight applications," in *Advances in Cryogenic Engineering,* R. W. Fast (Ed.), Vol. 31, pp. 791–799, Plenum Press, New York (1985).

Wheatley, John, T. Hofler, G. W. Swift, and A. Migliori, "Understanding some simple phenomena in thermoacoustics with applications to acoustical heat engines," *American Journal of Physics* **53,** 147–62 (Feb. 1985).

Williams, Brian G., "Extended life refreezable solid cryogenic cooler," MS Thesis, Department of Mechanical and Aerospace Engineering, Utah State University (1991).

Xie, J. K., "A fast cool-down J-T minicryocooler," in *Advances in Cryogenic Engineering,* R. W. Fast (Ed.), Vol. 29, pp. 621–627, Plenum Press, New York (1983).

Yazawa, T., A. Sato, and J. Yamamoto, "Adiabatic demagnetization cooler for infrared detector," *Cryogenics* **30,** 276–280 (Mar. 1990).

Yoshimura, Hideto, and Masakuni Kawada, "Small Vuilleumier cooler," in *Advances in Cryogenic Engineering,* R. W. Fast (Ed.), Vol. 33, pp. 837–844, Plenum Press, New York (1987).

Yoshimura, Hideto, Masashi Nagao, Takashi Inaguchi, Tadatoshi Yamada, and Masatami Iwamoto, "Helium liquefaction by a Gifford-McMahon cycle cryogenic refrigerator," *Review of Scientific Instrumentation* **60,** 3533–3536 (Nov. 1989).

Zhu, Shaowei, Peiyi Wu, Zhonggi Chen, Wenxiu Zhu, and Yuan Zhou, "A single stage double inlet pulse tube refrigerator capable of reaching 42K," *Cryogenics* **30** (supplement), 257–261 (Sep. 1990).

CHAPTER 7

Image Display Technology and Problems with Emphasis on Airborne Systems

Lucien M. Biberman
Institute for Defense Analyses
Alexandria, Virginia

Brian H. Tsou
Armstrong Laboratory, Crew Systems Directorate
Wright Patterson Air Force Base, Ohio

CONTENTS

7.1 INTRODUCTION

Computer-related technologies have made the development and manufacture of displays a "greater than ten billion dollar a year industry."[1] In the past few years a resurgence in display technology has occurred, accompanied by improvements in the understanding of vision in display applications—although that understanding seems to have been confined to a precious few people. Much of the work on understanding display requirements from the observer's point of view has been driven by the need for better cockpit information handling in both commercial and military aircraft. Although this chapter provides a broad, general collection of information, much of its impetus stems from military applications. Thus, the chapter is slanted in that direction, both in the text and, perhaps more noticeably, in the references and bibliography. A bimodal distribution is apparent in the dates of the references and bibliographical material, many of the items being either from the late 1960s to the early 1970s or from the mid-1980s to 1990.

There is a prevailing opinion that the presentation of data to an observer via a display monitor is a simple function of geometry and geometrical optics, and that the electronic technology involved is a simple matter solved by good development planning, use of good materials, and good engineering sense.

Unfortunately, the perception of displayed information is largely a perceptual problem for the human observer, and it now seems that factors originating between the human's eyes and brain are more significant than those relating to the geometric optics of human eyes (Secs. 4.3.1 and 4.3.2 in Ref. 1). Increasing the display resolution at the expense of necessary field rates or frame rates confuses the human perceptual system with a situation it misinterprets. A human observer perceives a high-resolution, high-contrast image in which there is motion, presented on a display at a low frame rate, as being of low resolution. This is due to motion effects between frames and a temporal sampling rate that is too low.[2]

Image sampling is a major problem yet to be resolved, in the sense that poor sampling diminishes information transfer yet at the same time reduces costs considerably. Photographic or cinematic imagery that presents an image sampled only by the grain size of the film used is far different from the one-dimensionally band-limited imagery of television, and it is even more different from the two-dimensionally sampled imagery of the modern forward-looking infrared (FLIR) device and its digital two-dimensionally sampled liquid crystal display (LCD). The LCD has received increasing attention of late because of its ability to preserve contrast and readability under extremely high ambient illumination such as that often encountered in an airborne environment. This ability is due not to inherent brightness but to filtering by a double-twisted nematic design using two polarizers.

Among cockpit displays, collimated "infinity" displays such as the head-up display (HUD) are quickly becoming complex display media that provide not only aiming information but also critical flight information and warnings to pilots in high-performance military as well as commercial aircraft. HUDs are even beginning to be found in automobiles. In recent years, however, concern has arisen about the possible contribution of HUDs to spatial disorientation. This area of concern deserves further investigation. Some pertinent sources of

further information are listed in the Bibliography under the section on "HUDs and Disorientation."

More recently, the helmet-mounted display (HMD) has been used in much the same way as the HUD, but the HMD affords an expanded capability, since by virtue of its being helmet mounted it does not constrain the pilot to a particular viewing direction. However, a number of HMD design variables interact with vision and visual perception and are not well understood. These variables include field of view, resolution, vibration, luminance, contrast, adaptation, ocular vergence and accommodation, distortion, registration, and color. Some of these parameters interact with each other to limit the direct applicability of existing guidelines. The accumulated operational experience with the U.S. Army AH-64 helicopter's Integrated Helmet and Display Sight System (IHADSS) has been documented by Rash, Verona, and Crowley.[3]

The choice of specifications for a display terminal of a sensor system is not different in principle from the choice of specifications for detectors, amplifiers, or data processors. The designer must first establish the needs or requirements for the overall system output and ensure that the display can provide gain and bandwidth sufficient to fill the data throughput requirements. The designer must then ensure that the inevitable introduction of noise into the subsystem does not degrade system utility below the design requirement. Finally, the critical problem of coupling the sensor system to the observer must be treated with greater care than usual, since this last interface can and often does degrade real systems far more than most other related major hardware design decisions. Once the designer has considered these primary issues, he can proceed to examine the available devices that may fill the matrix of parameters established by overall system requirements. However, considerable attention must be given to the special problems caused by the introduction into the system of a human as a critical data processing element.

Section 7.2 reviews the factors that govern the transfer of information across the display/observer interface and then examines the technologies for display implementation in Sec. 7.3. Section 7.4 discusses standards and calibrations, and Sec. 7.5 provides some caveats and personal opinions.

A generalized step-by-step procedure for designing a display system follows in Sec. 7.6. This is not presented as a series of highly specific algorithms but rather as a series of questions concerning what needs the display must serve and what the operating conditions will be. These factors are dealt with sequentially, and sometimes recursively, until a convergence to an acceptable result is achieved.

A significant portion of the material in this chapter comes from two sources: *The Infrared Handbook*,[4] Chapter 18, and IDA Document D-713, *Proceedings of the Sensor Display Workshop*.[5]

7.2 DISPLAY PERFORMANCE REQUIREMENTS

7.2.1 General

Display performance requirements are primarily determined by the characteristics of a human operator and the visual task to which he has been assigned.

The environment seriously affects the observer's abilities to do even simple tasks.

It is clear that the information contained in a sheet of microfiche is degraded in the process of projection. Yet, in spite of these losses, an observer can obtain an understanding of the microfiche content by using a projector or "reader," even though the microfiche contains more signal and less noise than its poorly displayed projection on a granular screen. The microfiche image size is such a bad mismatch to the eye that it must be enlarged by the projector. This situation is akin to the mismatch in the display of information in many electro-optical systems. Though the signal may well be there, it may also be nearly useless because of its form or size, not because of its content. Errors are often made by designers when they provide excellent displays that require too much time to scan visually. Trade-offs are very important, since display size may determine the overall utility of an otherwise good sensor system.

7.2.2 The Display and the Observer

The detail discrimination threshold of the eye, i.e., visual acuity, has been investigated exhaustively. Some recent work, cited by Levine,[1] has shown that the effective acuity of the overall eye-brain combination is more in the retina and the various synapses on the way to the brain than in the geometrical optics of the normal eye.

Various kinds of acuity such as minimum detectable, minimum separable, vernier, and stereo have been qualified and defined. Minimum separable visual acuity applies in the case of shape recognition in which, generally, closely spaced image details must be discerned. It is known to vary as a function of adaptation level, image brightness, contrast, exposure time, image motion, vibration, spectral characteristics, angular position of the target relative to the line of sight, etc. Visual acuity is usually defined in terms of arbitrary regular test patterns with generally sharp edges, although some studies have been conducted with sine-wave patterns.

The published acuity data are statistics representing specified performance levels (usually 50% detection probability). Thus they provide information in a probabilistic sense rather than in a deterministic sense. Therefore, in any specific instance, visual performance may fall far short of, or exceed, predictions based on published data. In general, standard visual acuity data are modified by field factors to obtain realistic operator performance estimates under operational conditions. Unmodified data can be used to establish average expected limits of performance under ideal conditions.

Minimum contrast threshold visual acuity curves from Patel[6] and from Murch and Virgin[7] are plotted in Fig. 7.1. The data in Fig. 7.1(a) are for a sine-wave test pattern with an average brightness of 100 foot-lamberts (fL) and viewed at 25 in. This curve neglects image motion, exposure time, wavelength, and vibration effects. The visual acuity curve sets the lower limit on useful system contrast. To be visually discernible, an image detail must exceed the threshold modulation of Fig. 7.1(a). The maximum usable resolution of a sensor display system (for a specified viewing distance) is indicated by the point at which the modulation transfer function (MTF) of the system crosses the corresponding visual acuity modulation threshold. Figures 7.1(b) and (c) use different units of measurement to illustrate the same phenomenon.

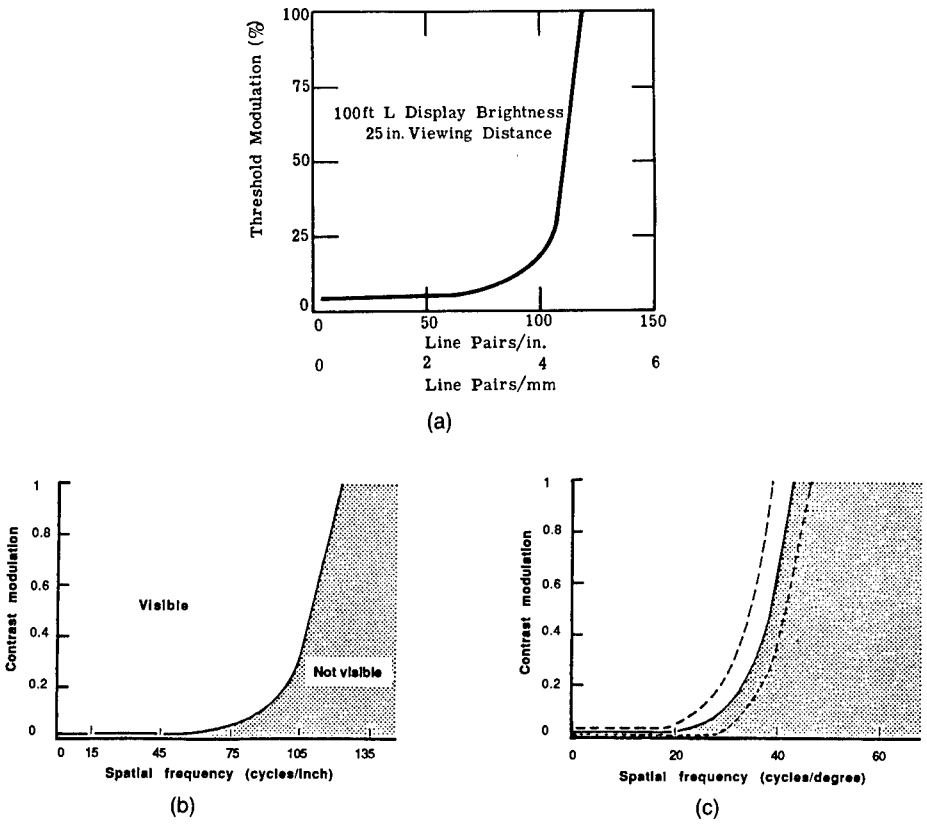


Fig. 7.1 (a) Visual acuity threshold modulation.[6] (b) Contrast sensitivity of human vision for vertical and horizontal gratings (in cycles per inch) viewed from a distance of 18 in.[7] (c) Contrast sensitivity function of human vision (in cycles per degree) of visual angle subtended by vertical and horizontal gratings. The broken lines represent 90% population limits.[7]

Discrimination of imagery detail differs from visual acuity measurements in that it requires detection of discontinuities characterized by diffuse edges and irregular brightness distributions. This topic is covered rather completely in the first six chapters of Ref. 8. A more recent report by Task[9] evaluates several measures of image quality and is recommended to those who wish to pursue this subject further. Related material is listed in the Bibliography under the heading "Observer: Eye, Performance, and Performance Measures."

The display contrast must be sufficient to provide a clearly visible image when the ambient illumination is as high as 10,000 foot-candles (fc) or more. From a display design standpoint, sunlight shining directly on the display [e.g., cathode-ray tube (CRT) phosphor] represents the most severe lighting condition. In this case, to be clearly visible, the maximum brightness of the image must be significantly greater than the brightness of the ambient light reflected back from the phosphor. It is precisely here that the new LCDs have perhaps their greatest value. Demonstrations have clearly shown that when a CRT display and an LCD placed side by side are illuminated by simulated high-altitude, full-brightness sunlight, the CRT image washes out but the adjacent LCD image suffers a negligible effect.

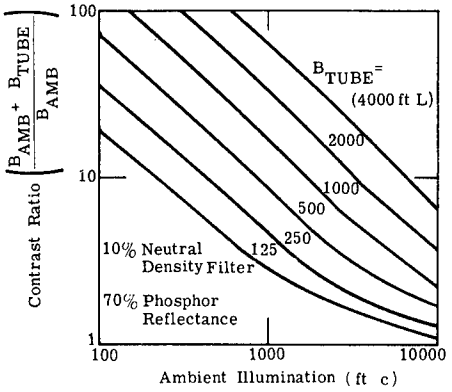


Fig. 7.2 Contrast ratio as a function of ambient illumination and CRT display highlight brightness.[10]

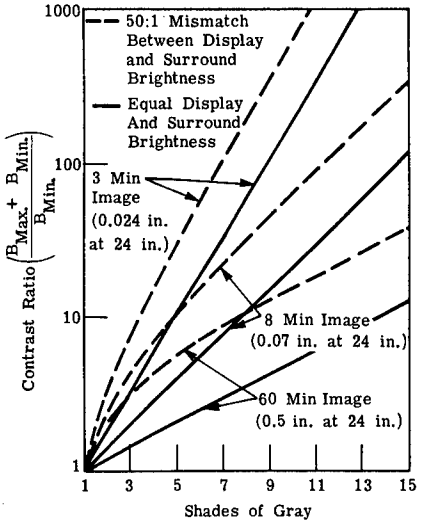


Fig. 7.3 Contrast ratio versus shades of gray.[10]

The maximum brightness that can be obtained in most of the better high-resolution CRT displays is generally about 500 to 2000 fL. If a neutral density filter of density 1 (10% transmission, which is considered about optimum) is placed in front of a CRT display, the filter reduces the tube's apparent brightness by a factor of 10, and it reduces the brightness of incident ambient illumination by a factor of 10 on the way to the tube face and by another factor of 10 when the ambient illumination is reflected by the tube face back through the filter. Thus, if that CRT display has a highlight brightness of 500 fL and a surface reflectance of 70%, and if we define the contrast ratio as

$$(B_{amb} + B_{disp})/B_{amb} , \tag{7.1}$$

then ambient illumination of 2000 fc, after reduction by 0.1 by the filter on the way to the tube face, by 0.7 on reflection, and by another 0.1 on passing through the filter again, yields a contrast of about 4.5.

The CRT contrast ratio can be read easily from the curves in Fig. 7.2, which shows contrast ratio as a function of ambient illumination and CRT display highlight brightness. The contrast ratio required by the observer's eye for any given number of successive gray shades when the observer is scanning the display is shown in Fig. 7.3. The number of gray shades seen by the eye with a given contrast ratio depends markedly on the image size as well as the brightness. Three different target sizes are plotted: 3, 8, and 60 min of arc. These curves are based on data from Blackwell (see Bibliography under "Observer: Eye, Performance, and Performance Measures") and include the "field factors" that transform laboratory threshold data into a form more appropriate for operational use.

Using the curves in Figs. 7.2 and 7.3, one can calculate the CRT brightness required for a given number of successive gray shades under a variety of

conditions. As noted above, the maximum image contrast ratio is about 4.5 to 1 when a 10% neutral density filter is used on a 500-fL display in a brightly sunlit environment. If an image with gray shades generated on the display subtended 8 min of arc at an operator's eye, the operator could discriminate about five shades of gray, according to Fig. 7.3. Under the same conditions, about seven shades of gray could be seen on a 1000-fL display (according to Figs. 7.2 and 7.3). To facilitate target recognition by an operator, at least seven shades of gray are desired in order to see the internal structure of low-contrast targets. "Seven shades of gray" has been a useful rule of thumb, since that is what has been available on many good-quality CRTs. Recent work by Silverstein et al.[11,12] has demonstrated the relationship between the available shades of gray and the ability to discern targets on two-dimensionally sampled double-twisted nematic LCDs. This work is probably the most up-to-date material available to the LCD system designer.

The recently developed active matrix addressed twisted nematic liquid crystal digital flat panel displays use polarizing filters and make relatively low-brightness displays very readable even when the display face is illuminated by strong simulated sunlight. Under such circumstances, even the best anti-reflective-coated CRTs with filters in front of them cannot match the performance of these LCDs, as demonstrated by Silverstein and his associates in mock-ups of aircraft cockpits with simulated high-altitude solar illumination coming in at all angles to the pilot—through the windshield, over the shoulder, etc.

In the case of HMDs, the demands for wide fields of view, high resolution, freedom from distortion, light weight, and acceptable helmet balance have led to a major series of engineering design and trade-off studies by Kocian and associates at the U.S. Air Force Armstrong Aerospace Medical Research Laboratory (AAMRL). Those studies are summarized in Ref. 13. Many other researchers in various government, academic, and industrial laboratories are also considering the visual psychophysical optimization of such a display. For example, Tsou and colleagues at AAMRL report in Ref. 14 that

> field-of-view and resolution requirements have been studied and analyzed by many HMD designers. Naturally, pilots prefer the same helmet display field of view, which is defined as the angular extent of information that can be presented instantaneously to the pilot by the display, as [that of] their unaided vision or, failing this, [they want a field of view] as large as possible. However, aircraft sensor performance, helmet display image source bandwidth, or other visual presentation subsystem capabilities place restrictions on scene content or resolution of scene detail to severely limit the instantaneous helmet display field of view that can be made available to the pilot. Application-specific design problems of this type have guided efforts to enlarge sensor field of view, to improve image source resolution, etc., but the problem encompasses more than just hardware. Preliminary studies using different visual psychophysical techniques to measure the visual field, a measurement of where one can see, with helmet-mounted displays of varying fields of view suggest that many factors such as eye/head dynamics, spatial awareness, and maybe even mental attentional processes may affect the perceived apparent field of view so as to impact pilot performance.

Many experiments designed to answer the field of view and resolution trade-off question have been performed in full-mission simulators, trainers, etc. At this time, there is no clear consensus among researchers on what the optimum balance is, except perhaps that the answer seems to depend on the mission

scenario. The answer to this question and many other trade-off issues may be properly obtained only through field tests. Greene,[15] of CCNVEO, using simulated night vision goggles during operational helicopter maneuvers, has obtained data that suggest that a horizontal field of view between 40 and 60 deg with at least 20/60 resolution[a] is sufficient to support contour and nap-of-the-earth (NOE) flying. It has been reported, on the basis of the General Dynamics F-16 Falcon Eye program, that a subtense of 30 deg is adequate to support the air-to-ground mission with a resolution comparable to that of the Texas Instruments head-slewable FLIR.[16] Many air-to-air flight tests to evaluate HMDs have also been completed, and reports on those tests are forthcoming. For the coming years, many such flight evaluations with actual HMD hardware are planned. They include the Army's LH, the Navy's I-NIGHTS, and the Air Force's AFTI/F-16 CAS and A-16 Night Attack programs. Undoubtedly, these flight tests will generate a wealth of data for use by HMD engineers in developing guidelines.

We should not leave the topic of display/observer interaction without mention of the very controversial article by Roscoe,[17] in which he points to the use of HUDs as the probable cause of a significantly large number of serious pilot errors, allegedly from HUD-induced misaccommodation.[b] Since not all the supporting evidence is unequivocal, we shall reserve judgment on this issue. Past research related to this issue is thoroughly reviewed and documented in an Army Human Engineering Laboratory technical report.[18] Additional pertinent sources of information are listed in the Bibliography under "HUDs and Disorientation."

7.2.3 Forms of Displayed Imagery

Many excellent treatments of the information content of an image are available, such as the excellent review by Linfoot[19] and, more recently, those by Silverstein et al.,[20,21] which discuss the role of color in displays of symbolic imagery. (See also the Bibliography under "Observer: Eye, Performance, and Performance Measures.") Here we simply point out that the information in a display is determined by the product of the number of picture elements in a frame and the number of discernible steps in brightness (the gray scale) of each element. The information flow rate is thus the product of the number of picture elements per frame, the number of brightness levels per element, and the number of frames per second. This product governs the bandwidth requirements of the system.

As noted in Sec. 7.2.4 on the signal-to-noise ratio (SNR) in the displayed image, the SNR produced by a piece of equipment and the SNR needed by the eye in an image vary with spatial frequency. Usually the SNR falls with increasing spatial frequency, though often some form of "compensation" is introduced to offset or diminish this problem. Shannon and Weaver[22] have defined the flow through a "channel." Since the SNR in an image is a function of spatial frequency, the information flow is the integral of SNR at each fre-

[a]As compared with the 20/20 resolution of normal healthy young eyes.

[b]Between the time this chapter was submitted and when it went to press, an IDA report specifically addressed this and related issues of HUD and HMD perceptual problems. (L. M. Biberman and E. A. Allvisi, "Pilot errors involving HUDs, HMDs, and NVGs," IDA P-2638, AD-A250719, Institute for Defense Analyses, Alexandria, VA, Jan. 1992.)

quency over the range of spatial frequencies of the display. The corresponding evaluation of the electrical signal involves the temporal equivalent of the spatial frequencies mentioned above and is the limiting minimum for determining bandwidth.

In halftone or pictorial representations, the number of levels or gray scales per picture element tends to be a number such as 6, 8, 10, or possibly 12 levels or "shades of gray." Two of these are used for white (maximum-brightness or "full on") and black ("off") levels. Because of this, the problem of symbolic data display, which often needs to show only a black level or a white level, becomes highly simplified compared to a picture of a scene with features represented by many more small variations in gray scale from picture element to picture element.

In digital systems, such elements are often referred to in terms of pixels (or picture elements) and levels (or shades of gray). Generally, analog systems are poorly described in terms of the pixel concept, since the very number of separable or resolved elements is a direct consequence of the dependence of the SNR on frequency. If the SNR is low, the variation in level between adjacent elements may not exceed the random variation in level, and thus adjacent picture elements may not be separable, or, in more common terms, may not be "resolvable." Because of these effects, both the bandwidth and the dynamic range (and thus the SNR) must be much greater in a pictorial display than in a device that only needs to reproduce alphanumeric symbols or the like.

7.2.4 Signal-to-Noise Ratio in a Displayed Image

The concept of SNR in a displayed image is one that is basic, important, and necessary in the design and/or specification of a display system. This concept, although developed from the work of several researchers, including the extensive lifelong work of Otto Schade, Sr., has since the late 1960s been extended and put into a more easily applied engineering form by Rosell et al.[23–27]

Most of the work in Refs. 23 through 27 was performed in the U.S. Air Force 698DF program in an attempt to establish the SNR requirements for an electro-optical imaging system. Earlier tests and experiments by many psychophysicists are reviewed in Chapter 2 of Ref. 8.

The earliest psychophysical experiments performed by Rosell and Willson employed simple rectangular images on a uniform background. These images were electronically generated, mixed with additive white noise, and displayed on a television monitor. The same amount of noise was added to both the rectangular image and its background. The purpose of the experiments was to determine the probability that an observer will detect a displayed image as a function of the image's SNR. These experiments proved easy to perform and, over the years, reruns of the experiments to establish equipment calibration have produced highly consistent results.

Although the SNRs for most equipment fall off first slowly and then rapidly with increments in spatial frequency, the sensitivity of the eye is low at both low and high spatial frequencies. The combination of the SNR presented and the SNR required determines the transferable information.

7.2.5 Effects of Vibration on Perception of Displayed Information

The environment strongly affects the choice of display parameters. For example, we resort to the analogy that it is more difficult to read a newspaper while riding a bus on a cobblestone street than on a smooth highway. The result of such relative vibratory motion between observer and display can be reduced materially through the use of a much larger display at a much greater viewing distance. The same amplitude of vibration is then small compared to the display imagery, while the overall angular resolution of the observer is less affected, but the vibration still produces enough retinal image blur to degrade visual perception. In addition, important physiological effects cause further degradations of observer performance, such as *biodynamic interference.*[28,29] Many and varied trials have been made and many simulations have been proposed and/or carried out in studies of the effects of vibration. Data for such trials and simulations are presented with more than adequate references in a broad collection by Boff and Lincoln of human perception and performance reviews.[30] Some recent representative studies are reviewed briefly below.

7.2.5.1 Fixed-Wing Aircraft Effects. Early work on vibration effects in fixed-wing aircraft was carried out and reported in 1976 by Rosell et al.[26] At the request of the Air Force Avionics Laboratory, as it was then known, Rosell planned and executed a flight experiment in which four subjects were exposed to a series of tests in the laboratory, in parked aircraft, and in aircraft under various flight conditions. In all of the tests, a recorded video was shown to the observers, and measurements were recorded and then compared. The experiment found that the effects of flight in turbulent air were strong indeed on the subjects' ability to perceive detail.

Four flights were flown during the first half of the spring 1975 Air Force 698DF program, and a total of 1928 data points were taken for all conditions. Figure 7.4 illustrates the placement of observers and displays.

In Figs. 7.5 and 7.6 the data from the four flights and the laboratory test are plotted. Straight and level data from flight 1 have been averaged with the laboratory data and are represented by the solid line in Fig. 7.5. The data from the two flights in turbulent air have also been averaged and are represented by the dashed line. Note the drastic increase in required SNR at the eye as the roughness of flight conditions grows. Note also that these effects are, as expected, much worse for imagery of small features and high spatial frequency than for imagery of large features and low spatial frequency—that is, these effects do not degrade reading newspaper headlines as much as they degrade reading the fine print of stock prices in the financial columns. The results of somewhat analogous experiments in a helicopter, highly condensed, are presented in Sec. 7.2.5.2.

7.2.5.2 Helicopter Data

Resolution Experiments on a Fixed Panel-Mounted Display. In the spring of 1990 R. Vollmerhausen[31] of CCNVEO planned a series of helicopter flight tests to evaluate the reduction in visual acuity. The tests were to be carried out in a helicopter whose condition was less than factory fresh, i.e., the rotors were

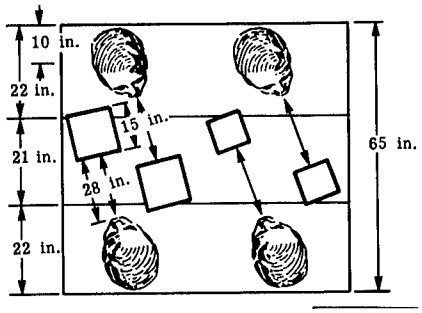


Fig. 7.4 Layout of observers' compartment.[26]

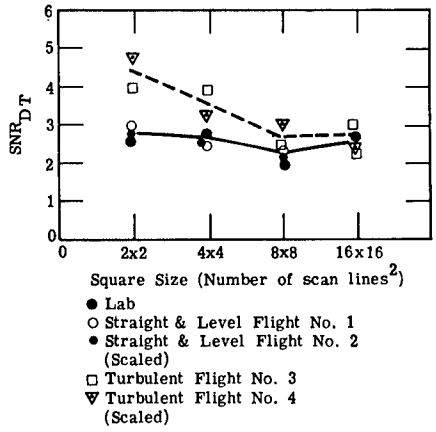


Fig. 7.5 Threshold SNR at the display (SNR$_{DT}$) as a function of square size for the four flights and laboratory comparison data.[26]

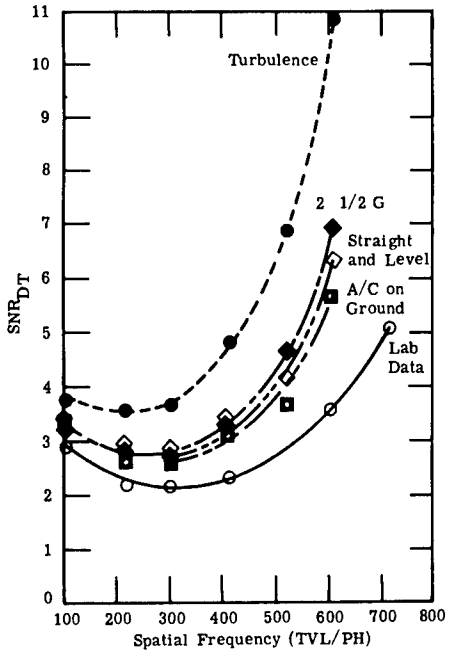


Fig. 7.6 Composite average threshold SNR$_{DT}$ for bar pattern recognition.[26]

not in a very good state of trim and balance. Factory maintenance representatives said the helicopter should get maintenance for rotor balancing. The aircraft was an OH-58D, a light scout helicopter with a four-blade rotor.

Although Vollmerhausen desired to do a scientifically meaningful experiment to obtain contrast sensitivity data much as one might in a well-controlled laboratory, the need to conduct the tests with rigorously correct sinusoidally modulated test charts at various source contrasts on a flying, vibrating platform at night made such an experiment infeasible because of its complexity, the experimental space limitations, the large number of test flight hours required, etc. It was necessary first of all to gain an insight into the problem at a reasonable cost in resources and money. Thus the flights were planned to determine visual acuity, a second choice but worthwhile as a source of data to shed light on the magnitude of the vibration problem.

In the flights observers were asked to view an eye chart and Air Force bar patterns that were on either backlit transparencies or frontlit paper prints.

The general findings were that at relatively low speeds the observers' acuity in flight was about 10% less than on the ground with no vibration effects. At higher speeds the acuity of the five observers could be seen to drop—by 40% at top operational speeds (Figs. 7.7 and 7.8; Tables 7.1 and 7.2).

Figures 7.7 and 7.8 show the calculated vibration-associated displacements and acceleration spectra, respectively; data were taken at hover and 50, 70, and 120 knots (VH max). For the test aircraft, the vibration spectra are dominated by the 6.6-Hz rotor blade rotation frequency (labeled "once per rev" in the figures) and by the fourth harmonic (labeled "four per rev"). Table 7.1 shows the acuity the observers were able to achieve for high-contrast bar patterns as a function of aircraft speed, i.e., hover through 120 knots. In the table, "Diagonal A" and "Diagonal B" refer to the axial positions of the bar charts being rotated 45 deg from the vertical in two directions, A and B. Table 7.2 gives results for low-contrast bar patterns and for alphanumerics.

Reference 30 documents the effects of vibration frequency and amplitude on display perception. The data also show that there are several distinct frequencies that have significant effects on vision. These include the rotational frequency of the helicopter rotor shaft (sometimes called the single-blade frequency), primarily due to the imbalance of the blades as a whole; the multiblade frequency, primarily due to the downwash of each of the blades in turn on the

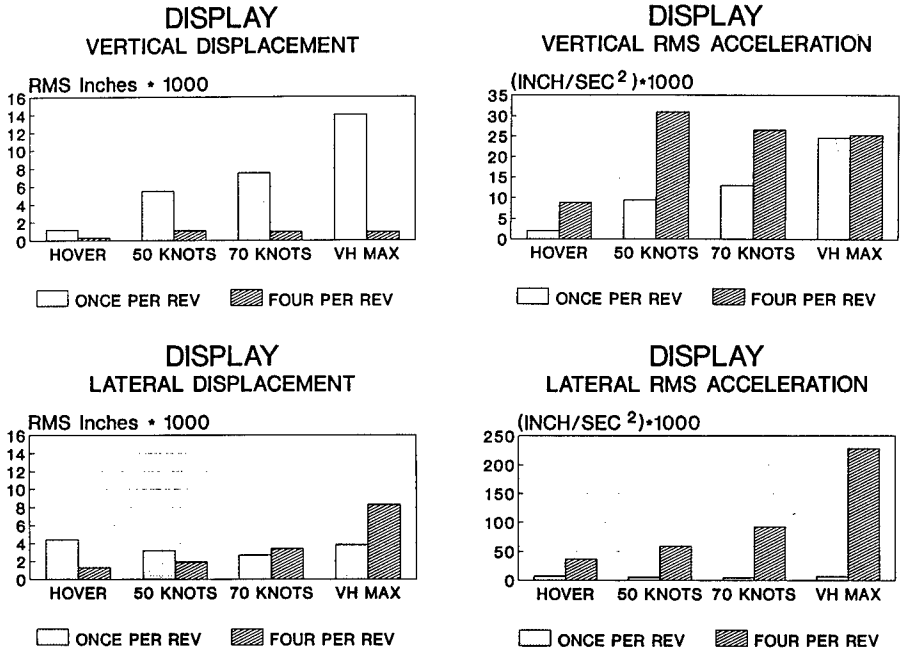


Fig. 7.7 Calculated vibration-associated displacement spectra for display in tests of visual acuity in helicopter flight.[31]

Fig. 7.8 Calculated vibration-associated acceleration spectra for display in tests of visual acuity in helicopter flight.[31]

Table 7.1 Visual Acuity Versus Helicopter Speed for High-Contrast Bar Patterns (from Ref. 31)

	Horizontal Resolution (cy/mrad)					Vertical Resolution (cy/mrad)				
	Obs 1	Obs 2	Obs 3	Obs 4	Obs 5	Obs 1	Obs 2	Obs 3	Obs 4	Obs 5
Baseline	1.25	1.65	1.4	1.8	1.3	1.25	1.65	1.4	1.8	1.3
Hover	1.0	1.45	1.35	1.35	0.9	1.0	1.4	1.35	1.35	0.9
50 knots	1.0	1.45	1.4	1.5	1.0	1.0	1.45	1.4	1.5	1.0
70 knots	1.0	1.4	1.25	1.4	0.95	1.0	1.4	1.25	1.4	0.9
120 knots	0.8	1.05	0.95	1.1	0.8	0.8	0.95	0.95	1.1	0.8
	Diagonal A (cy/mrad)					Diagonal B (cy/mrad)				
	Obs 1	Obs 2	Obs 3	Obs 4	Obs 5	Obs 1	Obs 2	Obs 3	Obs 4	Obs 5
Baseline	1.25	1.65	1.4	1.6	1.3	1.25	1.45	1.25	1.6	1.15
Hover	1.0	1.35	1.25	1.3	0.95	1.0	1.35	1.1	1.3	0.85
50 knots	1.0	1.35	1.25	1.35	1.0	1.0	1.45	1.25	1.35	0.9
70 knots	1.0	1.35	1.25	1.2	0.95	1.0	1.35	1.1	1.2	0.9
120 knots	0.8	1.0	0.85	1.1	0.8	0.8	1.05	0.75	1.1	0.7

Table 7.2 Visual Acuity Versus Helicopter Speed for Low-Contrast Bar Patterns (from Ref. 31)

Low-Contrast Bars										
	Horizontal Resolution (cy/mrad)					Vertical Resolution (cy/mrad)				
	Obs 1	Obs 2	Obs 3	Obs 4	Obs 5	Obs 1	Obs 2	Obs 3	Obs 4	Obs 5
Baseline	0.75	0.85	0.85	0.75	0.75	0.75	0.85	0.85	0.75	0.75
Hover	0.65	0.85	0.85	0.75	0.65	0.65	0.85	0.85	0.75	0.65
50 knots	0.65	0.85	0.85	0.75	0.65	0.65	0.8	0.85	0.75	0.65
70 knots	0.65	0.8	0.75	0.75	0.65	0.65	0.8	0.75	0.75	0.65
120 knots	0.6	0.7	0.6	0.7	0.6	0.6	0.7	0.6	0.7	0.6
Alphanumerics										
	Horizontal (lines/mrad)					Vertical (lines/mrad)				
	Obs 1	Obs 2	Obs 3	Obs 4	Obs 5	Obs 1	Obs 2	Obs 3	Obs 4	Obs 5
Baseline	1.0	1.05	1.4	1.35	1.4	1.0	0.85	1.4	1.5	1.15
Hover	1.0	1.15	1.25	1.1	1.05	1.0	1.15	1.15	1.2	0.95
50 knots	1.0	1.3	1.3	1.2	1.15	1.0	1.3	1.25	1.2	1.05
70 knots	1.0	1.3	1.25	1.1	1.1	1.0	1.15	1.15	1.1	1.05
120 knots	0.8	0.9	0.9	0.85	0.85	0.8	0.9	0.9	0.85	0.85

fuselage; the engine spectrum; and aerodynamic effects of flight. Some of the data from Ref. 30 might well be combined with the Vollmerhausen data for a variety of useful estimates by display designers.

Biodynamic Interference in Helmet-Mounted Displays. The common opinion among many crewspace designers seems to be that the use of an HMD would

eliminate the deteriorative effects of vibration on human visual perception, since the image is at optical infinity in an HMD. However, discussions with U.S. Army Apache pilots and Israeli Air Force pilots indicate that under conditions of high vibration (e.g., high-speed cruise and maneuvering) significant biodynamic interference effects have been experienced with HMDs. Head-coupled resonance of the Apache Hellfire laser ranging system has also been described. However, without accurate measurements these field reports remain anecdotal.

The effects on reading performance of whole-body vertical vibration in the range from 2.5 to 25 Hz, which includes the region from 3 to 5 Hz where the biodynamic feedthrough from seat to head is the largest, as well as a method to minimize those effects, have been investigated and reported by Wells and Griffin,[32,33] of the Institute of Sound and Vibration Research of the University of Southampton. Data obtained in simulators show that vibrations cause 130% increases in mean reading time per unit acceleration and 30% increases in percentage reading error per unit acceleration. With vertical and horizontal image stabilization, these decrements in performance were reduced to a less than 40% increase in reading time per unit acceleration and a less than 10% increase in reading error per unit acceleration. Furthermore, flight trials show that stabilizing the image significantly reduced the mean in-flight reading time to approximately 25 s (about 20% longer than the control condition—stationary on the ground) with a 4% reading error (0.4% for the control condition).

The Southampton image stabilization system employs a low-pass filtering approach, and it has generally been successful in improving pilots' ability to read symbology on a display even under vibration. It is less successful in assisting a pilot who is to acquire and track off-boresight targets using a helmet-mounted sight in the vibratory and turbulent environment of an operational helicopter, mainly because of phase lag introduced by the low-pass filter. Under this condition, the head tends to oscillate involuntarily, primarily in elevation, at common rotor-blade pass frequencies (about 5 to 6 Hz). This involuntary head motion causes the sighting device (reticle) also to oscillate with respect to the target. The presence of significant involuntary head motion, and in turn uncommanded reticle motion, elevates pilot workload and fatigue and degrades weapon system performance. The vibratory environment of current tactical helicopters, especially during aggressive maneuvering, or for aircraft that are operating out of track and balance, makes these biodynamic interference effects especially troublesome for helmet-tracking tasks. Alternatively, a more "adaptive" filtering is needed to eliminate biodynamic interference effects.

Lifshitz et al.[34] have investigated an adaptive filtering technique for tracking precision under vibration. Such filtering can be used to estimate the component of total head motion that results from involuntary vibration feedthrough. This estimated component is subtracted from the total command to yield only the "voluntary component" for driving the weapon system. Also, the involuntary component is used to stabilize the projected image, thereby reducing image blurring and pilot fatigue. Such adaptive filters have the key advantages of introducing no phase lag (unlike conventional filters) and becoming active only when biodynamic interference is detected. Lifshitz et al.[34] further report:

> . . . The results indicate that, for tracking tasks involving continuously moving targets, improvements of up to 70% can be achieved in percent on-target dwelling time and of up to 35% in rms tracking error, with the adaptive plus low-pass filter configuration. The results with the same filter configuration for the task of capturing randomly positioned stationary targets show an increase of up to 340% in the number of targets captured and an improvement of up to 24% in the average capture time. The adaptive plus low-pass filter combination was considered to exhibit the best overall display dynamics by each of the subjects.

However, in their target tracking experiment, the target dwell time is still 21% less than the stationary condition, and the radial tracking error is 48% larger. In their target acquisition experiment, the time per target capture is 19% more compared to the stationary condition. This level of performance, especially the radial tracking error, may be unacceptable during actual combat. Current work by Tischler, one of the investigators, is being done in simulators with the cooperation of the U.S. Air Force Armstrong Aerospace Medical Research Laboratory (AAMRL) at Wright-Patterson Air Force Base, Ohio, and the National Aeronautics and Space Administration Human Factors Division at Ames Research Center, Moffett Field, California. Both laboratories are planning on evaluating and possibly improving the various image stabilizing techniques in actual flight trials. AAMRL researchers are also working closely with the University of Southampton.

7.2.6 Sampling Effects in Displays

7.2.6.1 The Transfer Function of a Finite Sampling Process. In this section, Silk[35] analyzes the effects of multidimensional sampling as follows:

> The conventional definition of modulation transfer function (MTF) does not apply to the discrete sampling process because discrete sampling is inherently nonlinear and inhomogeneous. The concept can be redefined and reapplied in a useful way, but great care must be exercised in interpreting this new MTF. In particular, the fast rolloff is emphatically not a filtration process but an artifact of the inability of the lattice to support frequencies above the Nyquist limit. In fact, input frequencies above this limit appear as "alias" signals at lower output frequencies. So the usual relation between input and output frequency components in terms of the MTF will not hold unless the input scene is prefiltered so that all frequency components above the Nyquist limit are removed.
>
> Some confusion has arisen in the discussion of transfer functions associated with finite sampling of images. The usual construction of an optical transfer function begins with the assumption that an optical system "smears" a visual image according to

$$i(x) = \int h(x-x')\, o(x')\, dx' \ , \tag{7.2}$$

> where i is the image, o is the object, and h represents the action of the optical system. For simplicity we are considering one-dimensional functions. Since this is a convolution, the Fourier transforms have a simple relationship:

$$\tilde{i}(k) = \tilde{h}(k)\, \tilde{o}(k) \ . \tag{7.3}$$

> The transform of the smearing function h is called the *optical transfer function.* It is in general complex. Its modulus is called the *modulation transfer function,*

or MTF. The utility of the concept lies partly in the fact that successive smearings of the above form

$$i(x) = \int h_1(x-x_1) \int h_2(x_1-x_2) \ldots \times \int h_n(x_{n-1}-x_n)\, o(x_n)\, dx_n \ldots dx_2\, dx_1 \tag{7.4}$$

yield a composite transfer function, which is simply the product of the component transfer functions,

$$\tilde{i}(k) = \tilde{h}_1(k)\tilde{h}_2(k) \ldots \tilde{h}_n(k)\tilde{o}(k) \ . \tag{7.5}$$

In writing these equations we have made two implicit assumptions about the smearing process. First, it is linear. Thus, for example, saturation effects cannot be represented in the above model. Second, it is homogeneous. That is, all points in the object field are smeared equally.

In the case of finite sampling, these assumptions are no longer valid. The image function $i(x)$, which we obtain from a given object, depends on the details of how we orient our sampling lattice; as a consequence, linearity holds only approximately and over a limited range of frequencies. The result of this non-standard smearing is that the Fourier transform of the sampling function is not, strictly speaking, a transfer function, and it does not obey the above rules. We shall see that it is possible to fix things up if we are willing to observe a few caveats in the interpretation of our results.

Let us first construct the analog of the MTF for the sampling process, which we represent by

$$i(x) = \sum_{m=0}^{n-1} \varepsilon\, \delta(x-m\varepsilon) o(m\varepsilon) \ , \tag{7.6}$$

where n is the number of points and ε is the sampling period. The function $o(x)$ incorporates the effects of the system up to sampling, and $i(x)$ is understood to be subject to subsequent reconstruction. This expression resembles the usual form of a smearing process except that the integral has been replaced by a discrete sum over a finite set of lattice points. We take the Fourier transform of both sides and replace the object function by its inverse Fourier representation to obtain

$$\tilde{i}(k) = \int \tilde{f}(k-k')\tilde{o}(k')\, dk' \ , \tag{7.7}$$

where

$$\tilde{f}(k) = \frac{\varepsilon}{2\pi} \sum_{m=0}^{n-1} \exp(-im\varepsilon k) \ , \tag{7.8}$$

so the relationship between the Fourier transforms of the image and the object is no longer a simple proportionality but rather a convolution. This demonstrates that the concept of MTF is not well defined for the finite sampling process.

We now set about recovering a usable MTF. Our first task is to evaluate the sum in the previous equation. Using the identity

$$\sum_{m=0}^{n-1} x^m = \frac{x^n - 1}{x - 1} \ , \tag{7.9}$$

we obtain

$$\tilde{f}(k) = \frac{1}{2\pi} \exp\left[-ikL\left(1 - \frac{1}{n}\right)/2 \right] \frac{\sin(kL/2)}{\sin(kL/2n)} \,. \tag{7.10}$$

The phase factor arises from our choice of origin in x space and will not concern us here. The rest of the function is sharply peaked about $k = 0$; the width of the peak is $2\pi/n\varepsilon$, and its strength is unity. Therefore, for relatively low frequencies, the effect of the convolution is to map all of the object strength at a given frequency to a small range of nearby frequencies in the image. So if we restate the definition to something like, "the MTF is the response to an input frequency impulse at k within the image frequency range $k \pm 2\pi/L$," we find that, in fact, the MTF is unity up to the Nyquist limit and then cuts off sharply, simply because higher spatial frequencies cannot be accommodated on the lattice.

But there is a second, more fundamental, problem with this function: There is not a single peak, but an infinite number. The interval between peaks is equal to the sampling frequency, as shown in Fig. 7.9. This replication, which arises from the sampling, has two distinct effects. The first is that input spatial frequencies whose magnitudes are below the Nyquist limit are replicated an infinite number of times at higher frequencies. These replicants can be removed by reconstruction filters before the image is displayed, as discussed in Sec. 7.2.6.2. For the remainder of the present section, we consider the second effect: Any input frequencies above the Nyquist limit will be replicated below it. For a real sinusoidal input (which contains equal positive and negative frequency components) the in-band response is illustrated in Fig. 7.10. It is precisely the phenomenon known as *aliasing*—high input frequencies are masquerading as lower ones in the output.

Aliasing (Figs. 7.10 and 7.11) is the cause of the confusion in the discussion of the MTF. Our careful redefinition of MTF for the case of finitely sampled images has the unfortunate effect that it is no longer true that the MTF is equal to the ratio of the image to the object at a given frequency, simply because the image may contain contributions from remote object frequencies. So equality only holds if the object contains no frequency components above the Nyquist limit.

To avoid aliasing, the scene must be filtered *before* the finite sampling process occurs. In physical terms, once the light hits the sensor element it is too late to worry about aliasing. Postfilters can change the overall shape of the response and eliminate the replicants above the Nyquist limit, but at a given frequency the ratio of alias to faithful response will be unchanged; separation of the two would require some model of the visual scene.

While it is important to recognize that the alias response is not noise in the usual sense of the word (it is, after all, part of the signal itself), its effect is to

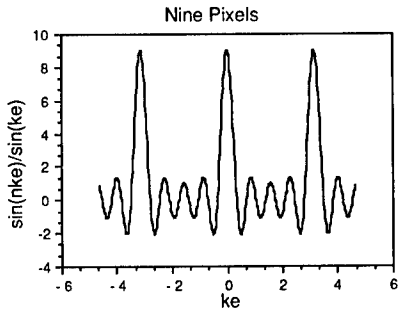


Fig. 7.9 The ratio of sine functions for $n = 9$ samples.[35]

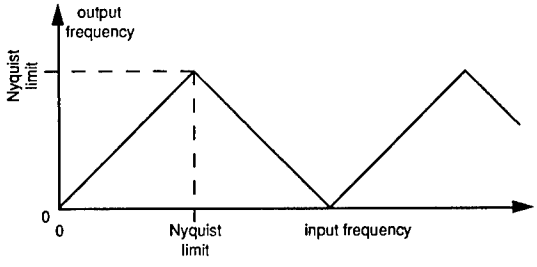


Fig. 7.10 Spectrum replication.[35]

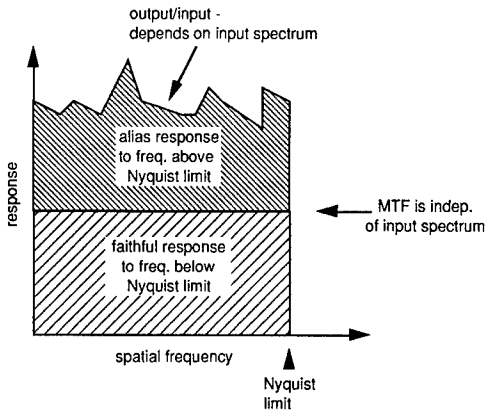


Fig. 7.11 Aliasing effects fold back those amplitudes at frequencies above Nyquist to add to those below Nyquist.[35]

mask the part of the signal that can be faithfully represented and therefore may be equivalent to noise from the point of view of a human observer.

Consider the combined influence of an aperture slit followed by sampling with a staring array. As before, we work in one dimension. Suppose we are observing a scene that contains equal contributions from all spatial frequencies. In the incoherent limit, the MTF of the diffractive blur spot from a slit is

$$\text{MTF} = \begin{cases} 1 - k/k_c \ , & 0 < k < k_c \\ 0 \ , & \text{otherwise} \ , \end{cases} \tag{7.11}$$

where

$$k_c = \frac{k_\gamma}{f/\#} \ . \tag{7.12}$$

Since there is a cutoff, there will be no aliasing if the cutoff frequency is lower than the Nyquist limit. If it is higher, though, the situation is illustrated in Fig. 7.12. The portion of the slit response that lies above the Nyquist limit has folded back across to lower frequencies.

We can use the definition of $f/\#$ to obtain a relationship between cutoff frequency and sampling frequency,

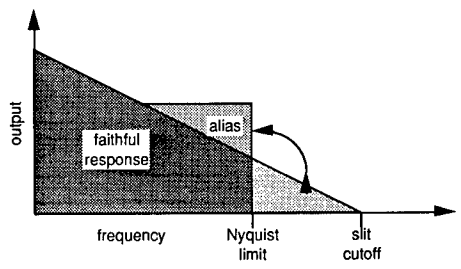


Fig. 7.12 Transfer of aliased signal into the passband.[35]

$$k_c = 4\frac{\varepsilon}{s}k_N \ , \tag{7.13}$$

so in the case for which the blur spot size is equal to the sampling interval, the cutoff is four times larger than the Nyquist limit, with the implication that for all frequencies at least half of the signal is aliasing.

This situation is ameliorated considerably by finite sensor size. The above analysis applies to a staring array of point sensors; if we consider the other extreme, where the sensor size is equal to the spacing (i.e., no gaps), then this is a prefiltration process whose MTF is proportional to $|\text{sinc}(k\varepsilon/2)|$.

7.2.6.2 Sampling and Display Processing. Vollmerhausen[36,37] has provided some conclusions drawn from his own experimental and analytical work on sampling and display processing. The Nyquist sampling theorem suggests that an image can be reconstructed from a finite number of samples; it does not suggest that using the sensor samples as discrete display intensities will constitute the image reconstruction. Reconstruction techniques can be used to calculate the value of image intensity at points *other than those sampled;* the calculated points along with the original samples can then be displayed. In Ref. 36, Vollmerhausen discusses the sampling theorem and how sampled imagery should be processed and displayed. The main points are:

- The frequency spectrum resulting from sampling is the original spectrum repeated at intervals of the sample frequency, weighted by the spectrum of the display pixel or reconstruction function.
- If the original spectrum was band limited and sampled at a rate above twice the highest frequency present, then the replicated patterns at multiples of the sample frequency will not overlap, and a filter can be used to recreate the original image; the impulse response of the display can be used to remove the unwanted replicas.
- A filter in the frequency domain is accomplished by convolving the filter transform over the image samples. In the case of the sampling theorem, a perfect low-pass filter is assumed, which involves convolving a $\sin(x)/(x)$ with the sample data.
- If sensor imagery is sampled just sufficiently to avoid significant aliasing, but is not excessively sampled, then the replicated spectrums above and below the baseband (original) spectrum will be closely spaced to the original but not overlapping. In this case, good display reconstruction techniques will be needed to discriminate the original spectrum from the replicated spectrums.
- In many cases, the eye and normal display blur spots do not provide a sharp differentiation between the desired original spectrum and the replicas at each sample frequency. Without reconstruction techniques, we are faced with accepting sample artifacts or accepting loss in perceived resolution of the original image (Fig. 7.13).
- Many of the artifacts labeled as "aliasing" result in fact from the presence of the higher frequency replicated spectrums and not from spectrum overlap. These artifacts can be corrected.
- Reconstruction techniques can be used to remove the high-frequency replicas and permit clean image information to be displayed or processed.

Vollmerhausen reconstructed a sampled image with a series of algorithms.

The original and the reconstructions are shown in Fig. 7.13.

Implementation of display processing need not be electronically complex but does require that we use more display pixels than sensor samples. Any ratio of display pixels to sensor samples can be used, but the simplest and most effective are two display pixels for each sensor sample and three display pixels for each two sensor samples. See Ref. 37 for a thorough if brief discussion of reconstruction algorithms.

Vollmerhausen pointed out that obvious and easy conclusions, based on the sampling theorem, do not apply to all sampling and reconstruction processes. *Sampled systems will not accurately replicate image frequencies up to half the sample rate unless adequate samples are provided and proper image reconstruction techniques are employed.* If these conditions are not met, sampling artifacts will occur when sensor limiting resolution is near the Nyquist frequency. In predicting sensor system performance, an analyst should evaluate the sensing techniques and the signal processing actually employed in the hardware. It is not sufficient to invoke the "two samples per cycle" rule and forego consideration of sample interval and reconstruction technique.

On the basis of simple computer simulations and limited testing of solid-state cameras, it appears that the sample rate should exceed 2.5 samples per resolved cycle unless sample function reconstruction techniques are used.

Further, while the task performance implications of using sampled sensors and displays are not clear, it seems unwise to ignore sampling artifacts that

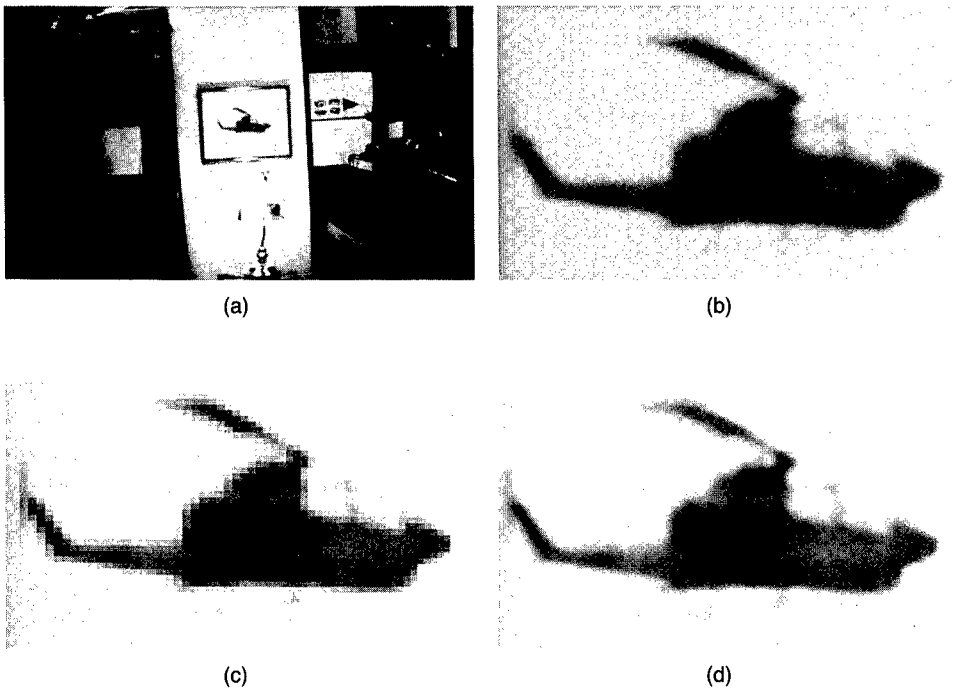


Fig. 7.13 Reconstruction of a sampled image: (a) original, (b) bilinear reconstruction, (c) pixel replicate, and (d) sampling theorem reconstruction.[37]

occur during the evaluation and testing of hardware when those artifacts lead to the periodic disappearance of the test pattern.

Vollmerhausen pointed out the care needed not to assume that the results first observed are valid! He strongly recommended that during the conduct of minimum resolvable temperature (MRT), minimum resolvable contrast (MRC), and similar tests, the test conductor should ensure that the limiting resolution performance credited to a sensor is not phase dependent. For example, in one test discussed by Vollmerhausen,[38] the highest frequency credited to a particular camera under test would be 0.48 cycles/mr, although the highest frequency increased to 0.54 cycles/mr when the camera was slightly reoriented in angle. Clearly the phase dependence must be checked by slightly moving (angling) the camera or sensor relative to the target pattern and checking to see that the pattern does not partially or totally disappear. Otherwise false impressions of performance can be created.

7.2.6.3 Effect of Sampling Grid on Flat Panel Displays. Most displays exhibit lines or picture elements (pixels) with some appreciable inactive, usually dark space between the lines or between the pixels. Often the entire intelligence to be conveyed is carried in the small amount of brightness variation along a line or between a group of pixels. However, the variation between the dark spaces and the brighter lines is usually far greater than along a line or between a group of pixels. Thus the observer is usually acutely aware of the line structure to the detriment of the observer's ability to notice the much smaller variations conveying intelligence. The late Otto Schade, Sr., created photographic analogs showing a high-resolution picture being "drowned out" by the line structure in both line-scanned and two-dimensionally sampled displays (Figs. 7.14 through 7.18). Figures 7.14 and 7.15 are reproductions from an Electronic Industries Association (EIA) test chart and a map legend sampled by a one-dimensional raster, while Figs. 7.16 and 7.17 are reproductions from the same test chart and map legend sampled by a two-dimensional point raster. Figure 7.18 is a bar chart of increasing frequency in one dimension.

Often this defect in CRT design is intentionally built in by the tube manufacturer, who makes a "high-resolution tube with a very small-diameter beam." This allows the tube to be used for low- to high-resolution applications, with more dark space in lower resolution display systems. One way to beat this problem is to shape the spot or vibrate the line so that it partially and controllably fills the dark spaces. This condition is called a *flat field.*

The high SNR between the lines and spaces produces an eye-brain problem that draws attention from the information contained in the line, just as noise inhibits listening to speech or music. The benefit of the narrow lines on a TV display tube is that such a tube can accommodate low-resolution imagery with large spaces between lines, or high-resolution imagery with small spaces between lines—that is, one size fits all! What is needed is a spot that somewhat overlaps the adjacent lines. Although such flat-field imagery can be obtained by defocusing the spot and thus losing a considerable amount of MTF in the process, it is clear that more information is readable by the observer in spite of the loss in modulation and thus the loss of some information.

The problem of reconstructing an image with a more or less rectangular distribution of pixels leads to a further problem. Silverstein et al.[11] have done

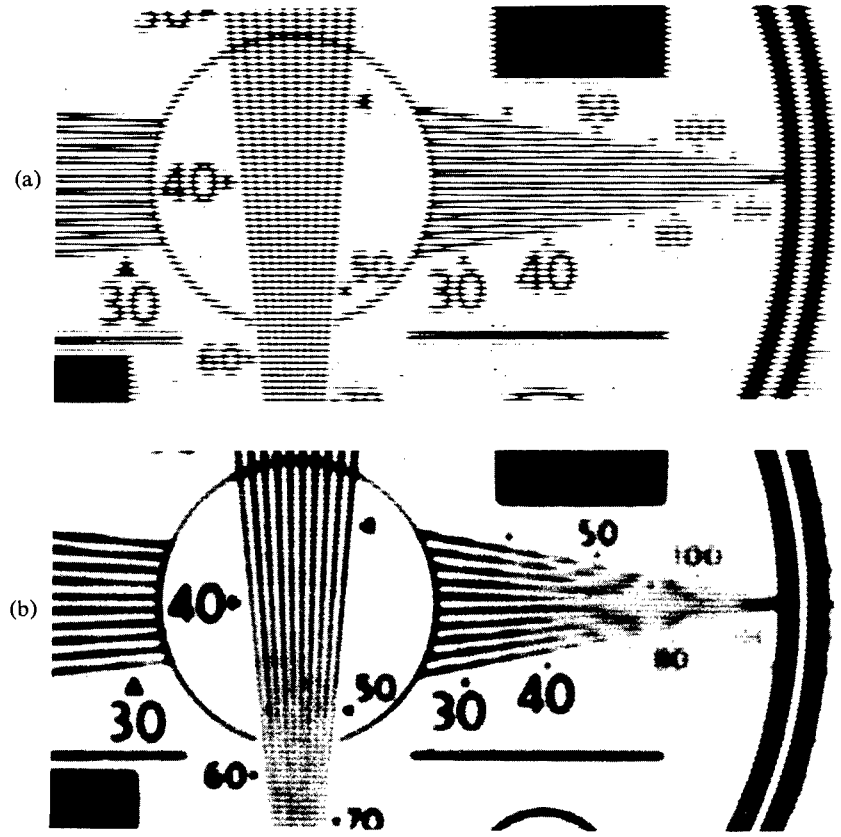


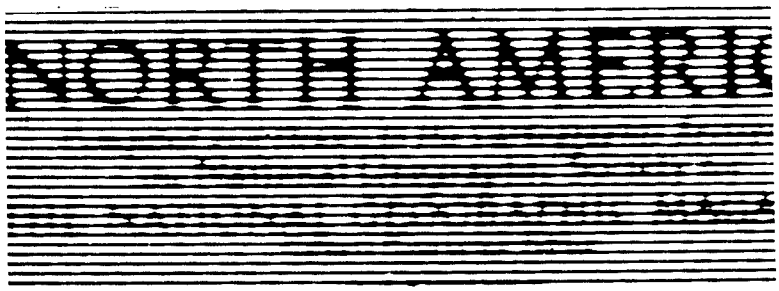
Fig. 7.14 Line raster process using sine-wave response factor of camera; $\bar{r}_c = 0.4$ at modulation frequency $f_m = 0.5$ raster frequency f_r, where f_r is for 70-in. pattern: (a) High MTF at the display (MTF_d) generates interfering line structure and (b) MTF_d reduced to obtain "flat field." See Ref. 39 for details.

appreciable work in attempting to understand this problem and how to alleviate it, as discussed later.

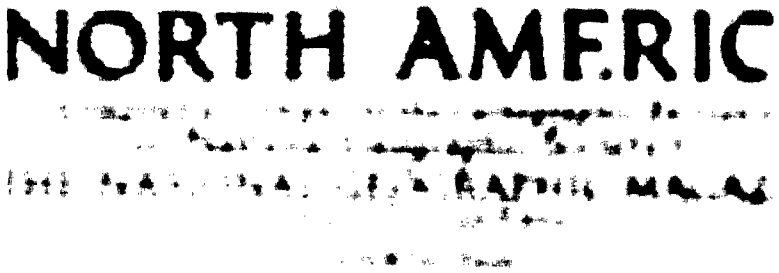
The arrangements of the physical pixels in the various flat panel display formats give rise to a quantization of the image, a form of granularity, that acts as still another form of sampling. Consider how you would write a line, and what the line would look like with the various pixel arrangements and with the line drawn at various angles. If the grid is rectangular and the color is one of the three primaries, a horizontal or vertical line is reproduced simply and well. What about lines at other orientations?

Silverstein et al.[11,20] have examined many of the problems of image sampling by CRTs and flat panel digital displays. Handling of a variety of line shapes and directions and colors requires the design of algorithms that optimize for each type of pixel arrangement, each color to be used, and the object to be drawn.

Thus the MTFs of such multicolor digital displays are not simple single functions, and generally are not representative of a simple line whose width



(a)



(b)

Fig. 7.15 Line raster process as in Fig. 7.14: (a) high MTF_d and (b) flat-field MTF_d.[39]

is that of a single pixel. Figure 7.19(a) shows the three commonly used arrangements of color pixels—RGB delta triad, RGB diagonal, and RGBG quad. Figures 7.19(b) through (e) show the pixel configurations in the three formats for horizontal and vertical lines seen as yellow, i.e., having only red and green pixels activated. For example, in the triad format, a horizontal yellow line would involve writing alternating pixels of red and green and blank (where the blue would come), thus giving a line of pixels like that in Fig. 7.19(b). A horizontal yellow line in the diagonal format would use the same pixels and have much the same form, while in the RGBG quad format it would be arranged as in Fig. 7.19(c), with no blanks. Displaying vertical yellow lines, the pixels in the three formats would be arranged as in Fig. 7.19(d). Displaying slanted lines, the three formats would give pixel arrangements as in Fig. 7.19(e). Clearly the imagery is dependent on a multiplicity of factors! Note that color balance is not achieved simply. For example, in the RGBG formats shown in Figs. 7.19(d) and (e), two greens and one red are required for yellow.

Now imagine that proposed sensors are two-dimensionally sampled with, say, 640 samples horizontally, and these are mapped onto a display with 80 pixels per inch with 7.75 useful inches, giving 620 horizontal pixels. How does

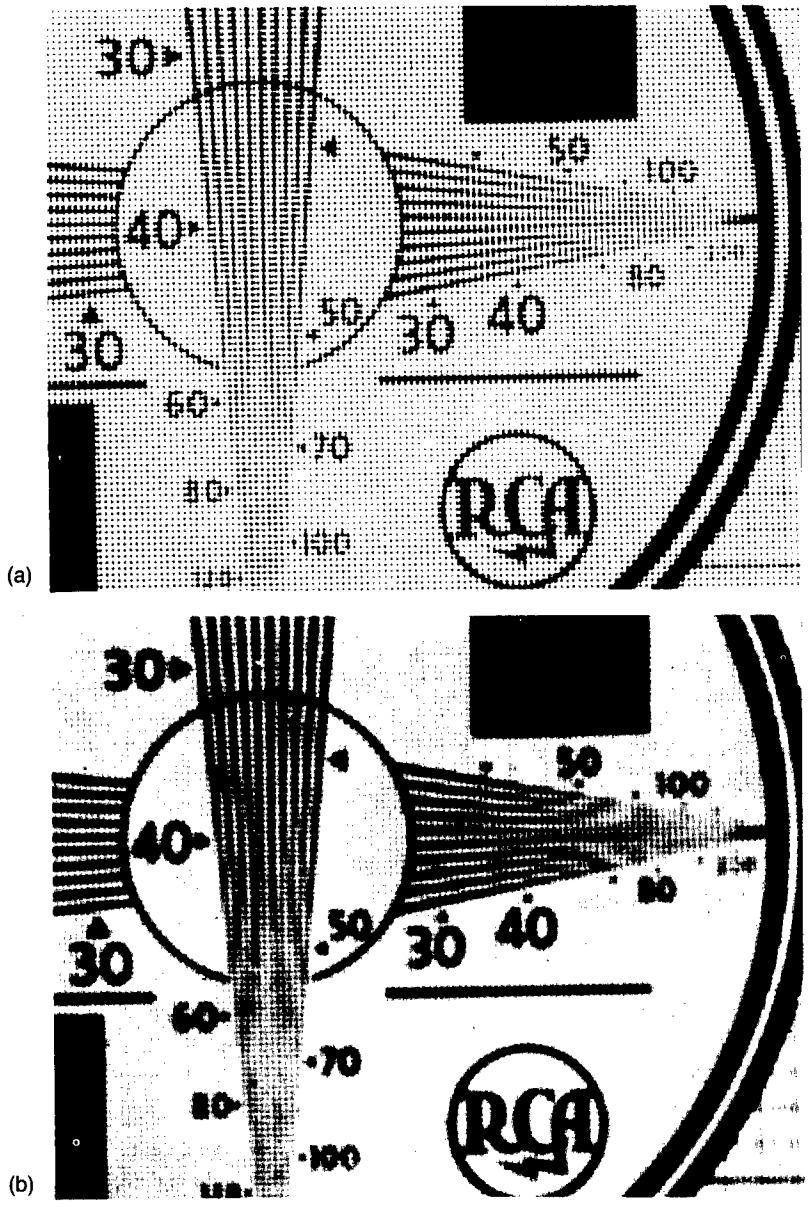


Fig. 7.16 Point raster process, raster frequency in the x direction f_{rx} = raster frequency in the y direction $f_{ry} = 70$; sine-wave frequency response of camera $\tilde{r}_c = 0.4$ at modulation frequency $f_m = 35$: (a) high MTF_d shows aperture diameter δ of structure and (b) MTF_d reduced to produce a "flat field."[39]

one map 640 samples onto 620 display pixels? Clearly one must either (1) use a reconstruction algorithm to produce a smooth function from the 640 samples and sample that to fit onto the 620 pixels, *and* do so in both dimensions, not just along one line, or (2) rethink the system formats more intelligently.

One finds that sampling onto a display gives rise to its own sampled-data-related imagery with strong aliasing, i.e., the folding of frequencies and their

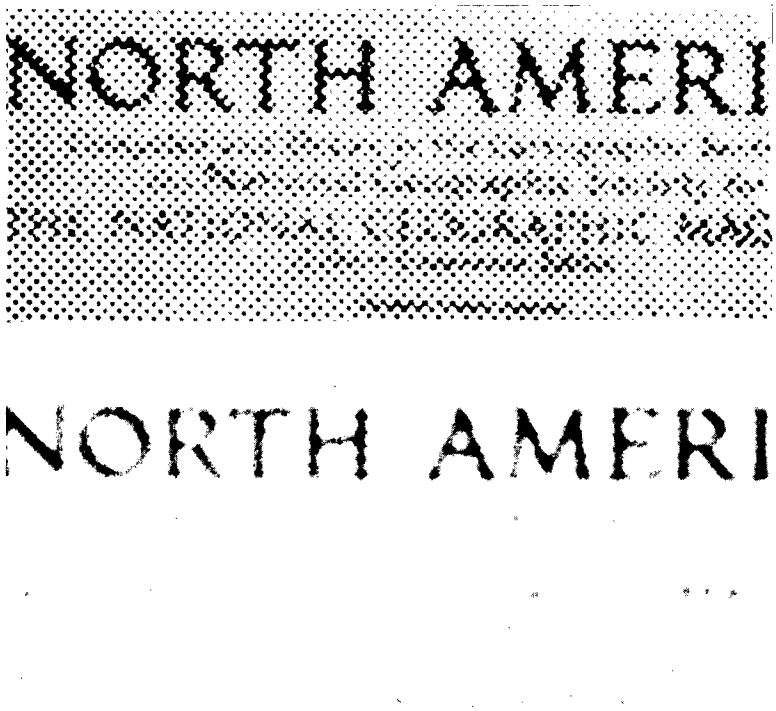


Fig. 7.17 (a) Point structure caused by high MTF_d interferes with detection of fine detail and (b) flat-field MTF_d improves detection of detail.[39]

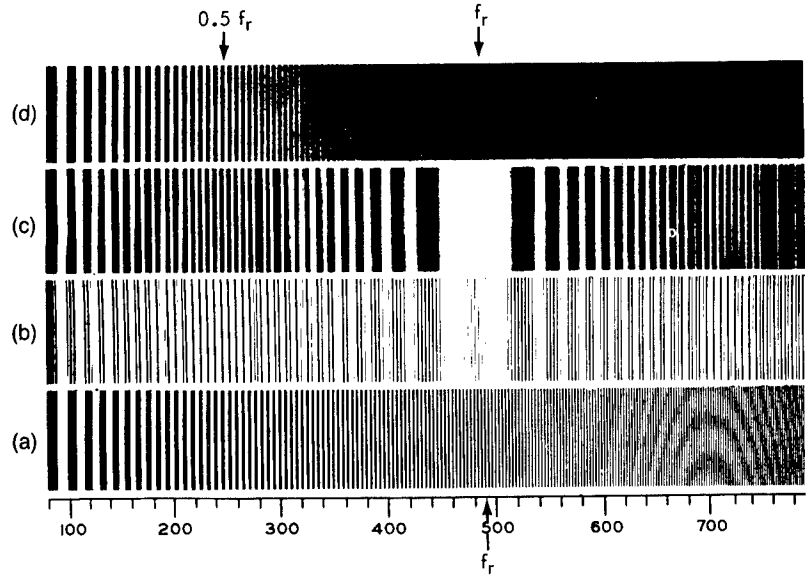


Fig. 7.18 Photographic proof of repeating line-number spectra ("sidebands") obtained by a line raster process (raster frequency $f_r = 490$).[39]

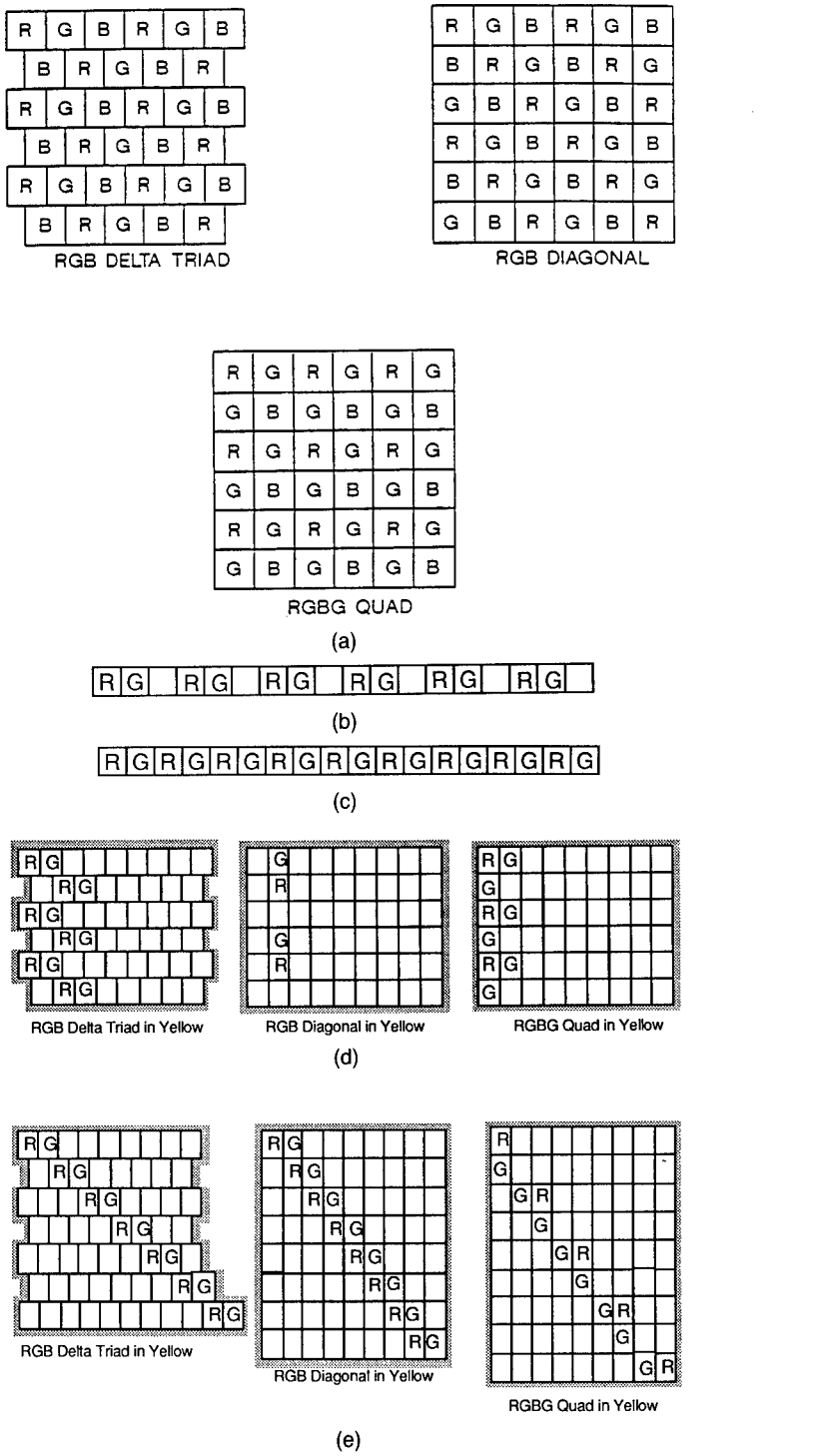


Fig. 7.19 (a) The three commonly used color pixel formats,[11] (b) pixel structure of horizontal yellow line in RGB triad and diagonal formats, (c) pixel structure of horizontal yellow line in RGBG quad format, (d) pixel structures in vertical yellow lines in each of three formats, and (e) pixel structures of diagonal yellow lines in each of three formats.

associated amplitudes back into the bandpass of the system (recall Fig. 7.11). This gives rise to the folding of low frequencies of significant amplitude into an otherwise low-gain, high-frequency region. Clearly this is undesirable.

7.2.6.4 Sampling Caused by Frame or Field Rates Relative to Motion in Scene. Many displays such as movies and television rely on sampling in the time domain. In Ref. 2, Watson et al. derived the spatiotemporal frequency spectra for some simple moving images and illustrated how these spectra are altered by sampling in the time domain.

They construct a simple model of the human perceiver that predicts the critical sample rate required to render sampled and continuous moving images indistinguishable. The rate is shown to depend on the spatial and the temporal acuities of the observer and on the velocity and spatial frequency content of the image. Several predictions of this model are tested and confirmed. The model is offered as an explanation of many of the phenomena known as *apparent motion*. Finally, the implications of the model for computer-generated imagery are discussed.

The general notions Watson et al. present regarding sampled displays and visual filtering can be extended to an arbitrary spatial image undergoing an arbitrary transformation over time, and the sampling process can be extended to the two spatial dimensions as well as time. They provide answers to some long-standing puzzles in perceptual psychology and to some modern problems in advanced visual displays.

The adverse effects of a frame rate or field rate that is too low on time-sampled imagery can, in some airborne applications, exceed those caused by almost all other sampled imagery problems. This factor must be critically examined in any design process.

7.2.7 Display Storage and Other Requirements

The storage, or phosphor persistence, for the short-time storage requirements of a display system or a display tube vary from less than one-sixtieth of a second for 60-fields-per-second refreshed TV-type displays to as much as a few minutes for a "rolling map" display. Such storage characteristics may be preset properties of the phosphor, programmed storage characteristics of a display subsystem, or entirely under the control of the operator who chooses to "freeze" the image.

Storage may be designed into a system through a number of elections: a direct-view storage tube (DVST), a scan converter of any of various forms, or straightforward memory devices with suitable data access such as tape, disk, or metal oxide semiconductor (MOS) or related solid-state devices.

Straightforward memory, or a simple direct-view storage tube, serves the simple storage problem fairly well, costing the designer a moderate amount of deterioration in both resolution and dynamic range over that achieved either with a simple CRT or with more advanced storage techniques that feed simple CRTs as the final output devices.

The use of a more sophisticated memory allows for selective write features. This makes possible continuous writing into a memory bank, fresh data being written into the oldest memory space that was just erased. Programmed reading-beam routines then read out the newest data followed by older data,

writing this on a CRT in such a manner that the newest data appear as the top line while the last line on the bottom disappears.

Most sensor displays will have requirements for symbology such as target marker symbols, artificial horizon and steering symbology, or cursors for target designation. It is obviously desirable for these symbols to create minimum interference and smearing when positioned and moved across storage images.

7.3 DISPLAY TECHNOLOGIES

This section summarizes and then compares some of the more completely developed display technologies before briefly examining them individually.

The use of the personal computer has placed demands on the display technologist as no single factor has done before. As a result, almost every form of display technology has risen to the market's demands, and almost unthinkable progress has been made. This situation is summarized well, albeit briefly, in a special issue of *Information Display*[40] consisting of seven brief but concise reviews of the display technologies.

The following technological descriptions have been compiled from material by Reingold,[41] Sherr,[42] Heilmeier et al.,[43] and Kocian.[13]

Cathode-Ray Tubes (CRTs). CRTs represent a mature technology of high reliability in widespread use for black and white or color. Associated circuits and hardware are readily available. Requires moderately high voltages for bright displays and substantial depth behind display surfaces. A variety of types exist for black and white, for color, for long persistence, for short persistence, for storage, and in miniature for the so-called head-mounted displays.

The miniature CRT[13] has been and remains the image source of choice for helmet display applications, especially if the image source is also located on the helmet. The cathodoluminescent and faceplate materials used in current CRTs can still be significantly improved to attain desired resolution, luminance, and contrast goals. The small gun apertures and the demanding CRT drive conditions associated with high-resolution and high-luminance performance often demand cathode current load levels of 5 to 10 A/cm^2. This is well above the 2-A limit that permits reasonable cathode (and therefore CRT) life to be accommodated by conventional oxide cathodes.

- *Video-driven image reproducers:* Basically television-like reproducers of pictorial or symbolic imagery, which may be either halftone or *only* black *or* white, i.e., text-like or symbolic. Image usually refreshed at 30 frame, 60 field rate.
- *Extruded-beam signal generators:* Symbols are limited to those built into the beam for alphanumeric symbols. Any shapes are possible if built into the tube or formed by alternate superposition of existing characters. Refresh necessary.
- *Storage tubes:* Large variety of methods for achieving storage. One method is by electrical charge stored on dielectrics within the tube in a manner to modulate spatially electron flow to the screen.
- *Digitally addressed flat panel CRT:* This product of Northrup Research in the early 1970s was an early effort to get a nearly flat panel display with many of the CRT attributes. It took the form of an unconventional

CRT about 2 in. thick. It employs an area cathode and dynode aperture plates. A multiplicity of electron beams is formed by the plates, one for each resolution element. A particular beam is selected by applying proper voltages to each plate in a binary selection scheme. The beam passing through the final plate impinges on a phosphor screen as in a conventional CRT. Resolution up to 80 lines per inch achieved. Viewing areas up to 7 × 7 in. available. This clever solution to a problem of the early 1970s has been overtaken by the various other flat panel display technologies.

Plasma Panels. Plasma panels are transparent panels often containing large arrays of discharge electrodes, usually in a common gas cavity. Both ac and dc versions exist. Plasma display technology is being developed in many sizes and for many applications. For large graphic displays it is the only technology seriously challenging the CRT. A brief description follows. A more complete discussion appears in Sec. 7.3.4.

- *dc plasma panels:* In one type of dc structure, two sets of parallel electrodes oppose each other in the gas, one set being directed orthogonally to the other. An aperture plate, placed between the electrode sets, confines the discharges. Appropriate addressing voltages on two intersecting electrodes cause an electrical breakdown and emission of light at the intersection; an addressing voltage on only one electrode is too small to ignite the discharge. In this matrix arrangement, the gas discharge cell, which is sufficiently nonlinear for the purpose, functions as a two-input AND circuit. This device is usually operated one row (or column) at a time, with signals on the opposing electrodes determining which cells in the row (or column) will be on. Since the duty cycle in these devices becomes smaller as the device becomes larger, the peak currents limit the array size.
- *ac plasma panels:* In the ac plasma panel, as in the dc panel, two sets of electrodes oppose each other across a discharge gap and are directed orthogonally to one another. However, at each discharge site, the dielectric surfaces that isolate the electrodes from the gas define two capacitances, which are in series with the discharge. An alternating voltage applied across the two electrode sets is too small by itself to ignite discharges. However, a pair of write voltages applied across two selected electrodes will ignite a pulsed discharge that extinguishes as ions, and electrons flow to the dielectric surfaces and charge the capacitors. This charge augments the applied voltage on the next half-cycle to ignite a second discharge, which then charges the capacitors in preparation for the third discharge. This sequence of pulsed discharges, which characterizes the "on" state of a cell, terminates when an erase signal on the two intersecting electrodes produces a controlled discharge that reduces the charge on the series capacitance below the minimum required for ignition.

Electroluminescent (EL) Panels. Electroluminescent displays consist of an EL powder or evaporated film between two electrodes, one of which is transparent. EL displays can be made in many colors. However, most displays are of a single color, usually green or orange because of the higher efficiency achieved with copper-activated and manganese-activated materials. When a potential is ap-

plied across the EL material, visible light is emitted. The potential may be ac or dc, depending on the specific structure, but EL displays usually operate in the ac mode. The resolution or pattern is defined by the electrodes. Luminance is typically 5 to 30 fL, although luminance in the thousands of foot-lamberts was achieved and demonstrated in 1974.

Liquid Crystals. A thin, clear layer of a cholesteric material placed between transparent electrically conducting covers becomes turbulent when excited by an electric field and scatters ambient light in a manner that yields an apparent brightness related to the applied field. When an aggregate of such cells forms a two-dimensional array, a digitally addressed display results. These displays can be small, light, and relatively inexpensive. The driving circuitry for large arrays is the more costly part, not the liquid crystalline materials, although the present-day cost of microelectronics makes the circuitry rather inexpensive.

Light-Emitting Diodes (LEDs). LED displays are now a mature technology for small-scale displays such as those in pocket calculators, small-area indicators, and related applications. Their utility for larger area or ambient brightness applications depends on improved luminous efficiency, lower power dissipation in driving circuits, and costs to challenge other technologies.

Projection Displays. High-luminosity CRTs, including storage tubes such as the Tonotron, liquid-crystal-controlled reflectors, and light valves of the oil-film type, are available and useful for various different levels of projected image brightness and size. The projection CRTs fill the needs of small-screen systems. The liquid crystal projection units are relatively small, light, and inexpensive, making them desirable for use in small meetings. The oil-film systems fill the need for small-to-large theater screen displays in black and white or color. A typical device is the Eidophor. Projection CRT displays are finding increasing application in displays for tactical systems in sizes from 3 to 6 ft on a side. Current tubes, with typical *f*/0.9 optics, can develop 200 to 300 lm output after optical surface losses. Resolutions of 1000 TV lines have been achieved on 5-in. projection CRTs. New longer life and more efficient phosphors are necessary to expand the application of projection CRTs.

Oil-film light valves are useful for a wide variety of command and control display applications of the fixed-site type. Devices can typically provide 525-line TV images with light outputs of 5000 lm. They are in general large, complex, and expensive systems. Small sealed-off light valves are available, but light output and resolution are limited by light source and cooling.

The more recent versions of smaller projection displays utilize projection lamps, mirrors, and LCD elements to switch the various pixels in color and/or black and white. Because the power comes from large projection bulbs, the image size and brightness fill a large number of projection needs for moderate-sized, moderate-resolution, lightweight image projectors.

The many display technologies cannot be ranked without a carefully considered weighting of needs such as performance, cost, size, and weight. The display suitable as a watch face does not necessarily do well as an aircraft head-up display or a road sign.

Miller[44] has reviewed and compared the CRT with flat panel display technology. His summary follows:

Wide usage of CRTs in television, radar, and oscilloscopes for more than half a century has provided the basis for tremendous investments in this technology, resulting in highly developed capabilities at low cost and continued improvements in luminance, resolution, size, and color, even a half century later. The CRT far exceeds any other technology in terms of image performance per unit cost. Flat panel displays, therefore, are finding application only where the CRT is too big, is too fragile, or consumes too much power for the intended use.[c] These conditions frequently exist in military systems, leading to active participation by all branches of the military in the development of flat panels.

The basic phenomena for flat panel displays were developed in the same time period as the CRT. The gas plasma grew out of studies done with the Crookes tube. Liquid crystals were discovered about a century ago. Electroluminescence was first observed more than 50 years ago. Development of practical display devices using these phenomena was delayed, not by a lack of understanding of the phenomena themselves, but by the need to develop practical matrix addressing techniques to make the display devices flat. The electron beam (cathode ray) that scans the face of the CRT is both its strength and its weakness. It greatly simplifies the task of addressing thousands of picture points spread across the two-dimensional faceplate, but it also fixes the demand for the large volume of the tube, the power consumption, and the hard vacuum needed inside the tube. Matrix addressable displays had to await the development of modern thin-film fabrication techniques to produce the panels and integrated driver circuits to operate economically the hundreds of matrix row and column lines.

The CRT is too well developed to be easily displaced by flat panel technologies, but flat panels are finding acceptance in new applications that are unsuitable for CRTs. Three technologies—plasma panels, liquid crystals, and electroluminescence—are currently in contention for most of these applications. Several other approaches, including electrochromics, electrophoretics, magneto-optics, and solid ceramic devices, appear to have fallen by the wayside. Vacuum fluorescent displays and light-emitting diodes are in wide use in low-information-density displays but are not being seriously proposed for high-resolution matrices.

After the CRT, the next electronic display device to see widespread use was the "Nixie" tube. This was a neon bulb with multiple electrodes in the shape of the numerals. Energizing the appropriate electrode provided a convenient and effective numeric readout. Through development of the gas mixture and electrode structures that solved lifetime problems, this approach was extended into the matrix-addressed plasma display. Plasma panels are in general use in commercial and military applications. These are monochrome devices with resolutions up to more than 100 lines per inch. They are currently being produced by a number of companies in Japan and the United States and have been built in sizes up to more than 1 m^2. Although these devices offer reduced volume compared to the CRT, they do not significantly reduce either the weight or power consumption. Color plasma panels have been fabricated by using argon as the active gas and applying patterned phosphors to the inside front surface of the panel. Ultraviolet light emitted by the argon excites photoluminescence in the phosphors. Several companies claim to have overcome problems with crosstalk between picture elements and degradation of the phorphors due to exposure to the plasma, but prototype devices have not yet been made available for evaluation.

Liquid crystal displays have a significant history of successful application at low information density in watches and calculators. They have been extended into matrix-addressed monochrome displays, which are in wide use in laptop computers and similar applications and are mainly produced in Japan. Active matrix color liquid crystal displays are being produced and sold in Japan for consumer television applications in sizes up to 5 in. on the diagonal, and more recently they have been demonstrated in sizes of 14 in. on the diagonal. The

[c]Flat panel displays are now also recognized as necessary where ambient illumination is high.

reputation for low cost and low power consumption that liquid crystal displays have acquired from their use in watches and calculators and early laptop computer applications is not being preserved as designers work toward higher performance displays. The higher contrast supertwist design, which sometimes uses a compensating layer, requires a backlight for operation, as do the active matrix color displays. The supertwist structure and compensator and the active matrix for multiplexing color displays add cost and complexity. The backlight adds to power consumption. At the same time, the operating temperature range and the acceptable viewing angle of these displays are limited. [Recent commercial work not yet formally announced has controlled temperature by heaters, raising display temperature above normal ambient, for example, by an amount acceptable for commercial passenger-carrying aircraft.]

Thin-film electroluminescent (TFEL) displays are produced by companies in Japan, Finland, and the United States, Planar Systems in the United States being the leading producer. These devices have applications in military, industrial, and medical equipment and are finding their way into computer applications. TFEL displays are being applied to a large number of tactical military systems because of inherent characteristics with regard to reliability, compactness, weight, gray shades, viewability, and power consumption. The displays can be operated under the full range of military environmental conditions and can be made to meet the requirements of sunlight legibility and compatibility with night vision devices. Because the development of this technology has been so recent, worldwide manufacturing sources are still somewhat limited. The manufacturing experience thus far has produced significant reductions in cost along with increases in performance, yield, size, and resolution. The major research efforts in conjunction with TFEL are directed toward development of full-color displays. Currently, the limitation in color display development is the lack of an adequate blue thin-film phosphor. The brightest, most efficient blue phosphor is SrS:CeF, which can produce several foot-lamberts but emits an unsaturated pale blue green. A promising candidate is ZnS:Tm, which emits a saturated blue at a fraction of a foot-lambert. Ongoing research is following approaches that are expected to improve both of these materials as well as to uncover others.

Military applications of displays have a long but somewhat limited history, being mainly radar, sonar, and a few airborne applications. The radar and sonar applications tend to be in large systems where the size, weight, and power of a CRT display device are not inconsistent with the rest of the equipment. In recent years, however, the microelectronics revolution has caused a major growth of the amount of electronic information being acquired, processed, and distributed in the tactical battlefield environment. As the interfaces between all of this information and the human operators, displays are becoming a requirement for systems involved in target acquisition, intelligence, command and control, maintenance, logistics, and training. In general, this means lightweight, low-power portable devices with the ruggedness to operate under the full range of environmental conditions. This has led to the application of plasma panels, for example, in the AN/UYQ-10, liquid crystal devices in the Remotely Monitored Battlefield Area Sensor System (REMBASS) monitor, and a TFEL display, which replaced a plasma panel in the Digital Message Device AN/PSG-5.

The ability of all of the previously discussed display technologies to produce color displays has been demonstrated to varying degrees. The use of color displays in military systems is much discussed, and some have been successfully used for displaying symbolic information, but very few, if any, requirements for color in systems designed to view terrain actually exist. The most extensive requirements for color displays, ranging across a wide variety of equipment, are expected to result from the efforts of the Defense Mapping Agency and the Engineering Topographic Laboratory to provide digital map data bases for use in military systems. Displays of these map data bases will be practically useless without color to distinguish among the various kinds of information on the screen at the same time. The need for color also exists in sensor displays to enhance target recognition capabilities.

CRT display technology is so highly developed that we merely tabulate its strengths and weaknesses and concentrate on the less-well-known versions such as the Digisplay®,[45] the Multimode Tonotron®,[46] single- and double-ended scan converters, and storage tubes.[44]

For a most convenient and quite up-to-date review of CRT displays and a separate review of flat panel displays, see the manuals and programs by Beta Review, Inc.[47] System designers could well benefit from these two programs available on disk, each with an excellent manual in adequate detail for almost all purposes except research into or development of new display technologies, but even there it might be of some future help to compare what can now be done with what is possible in a subsequent year.

Table 7.3 is a summary of CRT effectiveness factors according to Reingold.[41] Although his review is more than 15 years old, it still warrants attention.

It is clear that there are but two serious objections to the CRT for all its many good features. The CRT is a big, empty (vacuum) bottle that can and often does take a lot of valuable space for the display area it offers. Also, in some specific modes of beam addressing at high speeds, it can require significant amounts of power in the beam deflection circuits. Its technology, on the other hand, is very well developed and the associated driving circuits are so completely debugged that Anderson[48] was moved to write an excellent paper asking the question "Why consider anything else?" More recently, articles by Wurtz[49] and by Iki and Werner[50] look at the same question 16 years later.

Table 7.3 Summary of CRT Display Effectiveness Factors (from Ref. 41)

Factor	Required	Achieved
Brightness	Average—minimum of 50 fL Maximum—3000 fL	Yes (far exceeded) Yes (far exceeded)
Contrast	Viewable in shade for stationary displays Viewable in direct sunlight	Yes Only with DVSTs
Halftones	Radar displays—two-tone acceptable Television—five or more required	Yes Yes
Resolution	Minimum size commensurate with eye acuity Size constant with brightness and position	Yes No, but adequate for most purposes
Flicker	None present	Yes, for most applications
Distortion	Size constant with brightness and position	See "Resolution"
Accuracy	Position linear with input voltage	System rather than device limited
Blemishes	Radar displays—minimum loss of resolution elements Television—indiscernible loss of picture detail	0.04% maximum blemished area 0.005 to 0.01% maximum lost resolution elements
Volume-to-area ratio	Overall volume small for desired viewing area	Poor; display device volume and shape may dictate equipment volume and shape
Power consumption	Negligible fraction of total equipment power	Yes, for TV No, for random access

Why consider anything else indeed? For small or otherwise limited demands, both LEDs and LCDs offer cheaper, more compact displays. As size increases, complexity increases at a severe exponential rate. The problem shifts from the display to the interconnections and to the driver circuits that until recently were a tour de force in microelectronics. Microelectronics in the 1990s permit much more flexibility. Tuttle[51] points out that display interfaces are often misunderstood and are therefore likely to be designed in a less than optimal fashion. By careful attention to emerging display technologies and input devices it is possible to fabricate electronics that can take advantage of these advances without system redesign. Scan conversion techniques have been known for many years, but cost, weight, and size have precluded them from all but the largest, most expensive systems. Advances in memory density and large-scale integrated (LSI) sync strippers, analog-to-digital (A/D) converters, etc., have changed this completely. Now a scan converter occupies less than 50 in.^{2} of printed circuit board and consumes less than 10 W. The cost is now well under $1000 for commercial converters and is still dropping. Tuttle describes some of the constraints and techniques used to optimize a design and allow for future expansion. By incorporating this technology, not only can one keep pace with the newer displays but also old problems such as flicker can be eliminated, imparting more flexibility and higher quality to systems.

A comparison of flat panels and CRTs is in order. The conventional home television set has three driving circuits: one to brighten the beam to the level required for each picture element, and one each to deflect the beam in the directions of the x and y coordinates. One of these three drive circuits is a video driver, while the x driver operates in the upper audio region and the y driver is usually about a 60-cycle circuit. As an interesting comparison, note that the largest non-CRT video-compatible display in operation in 1978 was the Hughes 100 $\times$ 100 element LCD. Since then great strides have been made in flat panel displays, which are discussed later in this chapter.

On a smaller scale, at slower display rates, it is clear that alphanumeric displays for pocket calculators and cash registers have reduced the production costs of drivers and displays. In 1975 these devices were offered for prices at which "a $5 bill would supply display array, driver circuits, and change." In 1990 a hobbyist can buy such an LCD for less than a dollar.

The liquid crystal array has grown only recently from the Hughes 100 $\times$ 100 LCD of the very early 1970s to the early 1990 prototype of the double-twisted liquid crystal assembly in the form of 480 $\times$ 640 pixels over a 6- $\times$ 8-in display face. Even though the liquid crystal cells are but "printed circuits" of silicon and sapphire with only drops of liquid crystalline material per display, the development has been an expensive multidecade effort leading to a clear, highly developed new technology.

On the other hand, the development of a new CRT display, including the design of a new cathode-ray tube, is merely a matter of a few weeks for an electronic designer, a machinist, and a glass blower. This should explain the fact that, although flat panel displays could feasibly have replaced CRTs for some purposes, they have not done so where costs are quite low for new limited developments using CRTs. Unit costs of small flat panels in very-high-volume production are, of course, very low for inexpensive watches, calculators, and other small displays in inexpensive large-production applications.

As flat panels increased in size and pixel density, the addressing complexity grew rapidly, and the low power requirements for the color LCDs now had to include power for the backside-located illuminators, usually fluorescent lamps.

7.3.1 Summary of Conventional CRT Display Techniques

Although a complete treatise on CRTs is not feasible here, a good survey of CRTs is available.[47] The following short summary attempts to cover some of the most important points.

7.3.1.1 Available Phosphors for CRTs. The many varied phosphor screens available for CRTs give rise to families of subtechnologies for color and storage effects in display tubes. The depth of penetration achieved by variable beam velocities can yield multicolor displays with the proper beam controls. The factors governing penetration and luminous efficiency of such phosphors are covered in Refs. 52 through 64.

Table 7.4 shows the results of Kingsley and Ludwig's measurements[52] of the cathodoluminescence efficiency of a variety of phosphors. The compositions of the more common phosphors and their "P" designators are shown in Table 7.5. This list changes as the technology progresses. The Electronic Industries Association usually maintains the best current listings.

Table 7.4 Cathode-Ray (CR) Efficiencies of Various Phosphors (from Ref. 52)

Phosphor	Efficiency (%)	Phosphor	Efficiency (%)
Zn_2SiO_4:Mn	4.7	Y_2O_3:Eu	6.5
ZnS:Cu	12.4	Gd_2O_3:Eu	9.6–11.7
(Zn,Cd)S:Ag	18.7	Y_2O_2S:Eu	13.1
(Zn,Cd)S:Cu	9.7	La_2O_2S:Eu	10.6
$Zn_3(PO_4)_2$:Mn	3.1	Gd_2O_2S:Eu	10.2
$CaWO_4$:Pb	3.4	YOCI:Eu	12.9
$MgWO_4$	2.9	La_2O_2S:Tb	11.8
Zn_2SiO_4:Mn	6.8		

7.3.1.2 Short-Persistence CRTs. A short-persistence phosphor CRT with high brightness is ideally suited for television imagery. If a short-persistence CRT is to be used for a low-frame-rate sensor display, a scan converter must be used for intermediate storage of the sensor imagery, which is read out electrically at TV rates and displayed on the CRT. At present, a number of CRTs can provide brightness levels of 800 to 2000 fL at resolutions of 150 TV_{50} lines per inch or greater.[d]

7.3.1.3 Long-Persistence CRTs. For many years CRTs with either P7 or P14 long-persistence phosphors (Table 7.5) were widely used for air-to-air and air-to-ground radar and IR displays. Although electronic processing has now

[d]See Fig. 7.35, Table 7.12, and the accompanying text on sensor resolution for explanations of TV_{50} and other expressions of sensor resolution.

Table 7.5 Compositions of the More Common Phosphors and Their "P" Designators

P Designator	Material	P Designator	Material
P1	Zinc silicate:manganese	P22 *(continued)*	Zinc sulfide:silver, zinc; cadmium sulfide:copper; yttrium oxysulfide: europium
P2	Zinc sulfide:copper		
P3	Zinc beryllium silicate:manganese	P23	Similar to P4
P4	Zinc sulfide:silver and zinc cadmium sulfide:silver	P24	Zinc oxide
P5	Calcium tungstate	P25	Calcium silicate: lead:manganese
P6	Similar to P4	P26	Same as P19
P7	Zinc sulfide:silver and zinc cadmium sulfide:copper	P27	Zinc phosphate:manganese
P8	No information	P28	Zinc cadmium sulfide:copper
P9	Calcium pyrophosphate	P29	Similar to P2 and P25
P10	Potassium chloride (dark trace—nonluminescent—called a *scotophor*)	P30	This phosphor is no longer available
P11	Zinc sulfide:silver	P31	Zinc sulfide:copper
P12	Zinc magnesium fluoride:manganese	P32	Calcium magnesium silicate:titanium, zinc; cadmium sulfide:copper
P13	Magnesium silicate:manganese	P33	Magnesium fluoride: manganese
P14	Similar to P7	P34	Zinc sulfide:lead:copper
P15	Zinc oxide	P35	Zinc sulfide selenide:silver
P16	Calcium magnesium silicate:cerium	P36	Zinc cadmium sulfide:silver:nickel
P17	Zinc oxide and zinc cadmium sulfide:copper	P37	Zinc sulfide:silver:nickel
P18	Calcium magnesium silicate:titanium and calcium beryllium silicate:manganese	P38	Zinc magnesium fluoride:manganese
		P39	Zinc sulfide:silver, zinc; cadmium sulfide:copper
P19	Potassium magnesium fluoride:manganese	P40	Zinc sulfide:silver, zinc; cadmium sulfide:copper
P20	Zinc cadmium sulfide:silver	P41	Zinc magnesium fluoride:manganese, calcium; magnesium silicate:cerium
P21	Magnesium fluoride:manganese		
P22	Zinc sulfide:silver; zinc silicate:manganese; zinc phosphate:manganese	P42	Zinc sulfide:copper; zinc silicate:manganese:arsenic
		P43	Gadolinium oxysulfide:terbium
	Zinc sulfide:silver, zinc; cadmium sulfide:silver, zinc; cadmium sulfide:silver	P44	Lanthanum oxysulfide:terbium
	Zinc sulfide:silver, zinc; cadmium sulfide:silver; yttrium oxysulfide: europium	P45	Yttrium oxysulfide:terbium
		P46	Yttrium aluminate:cerium
		P47	Yttrium silicate:cerium
	Zinc sulfide:silver, zinc; cadmium sulfide:silver; yttrium oxysulfide: europium	P48	70:30 mix P46-P47
	Zinc sulfide:silver, zinc; cadmium sulfide:copper; yttrium oxide:europium		

largely supplanted these CRTs, as well as storage tubes and analog scan converters, they are treated here mostly for their historical interest.

Long-persistence CRTs have three disadvantages: (1) their brightness is low, (2) persistence is fixed and nonlinear, and (3) selective erasure is not possible. The low brightness makes it impossible, for example, to use such a display in an aircraft cockpit environment without ambient light shielding by means of a visor or hood. Even then, the operator is forced to adjust quickly from exterior ambient light levels as high as 10,000 fL to the low brightness of the CRT long-persistence phosphor.

Where ambient light can be controlled and cost is an overriding consideration, a long-persistence CRT can provide an acceptable display for low-resolution sensors. Since both P7 and P14 phosphors have short- and long-persistence components of different colors, it is possible to filter out the long-persistence component and present TV-frame-rate imagery (see Tables 7.6 and 7.7). For longer persistence, storage devices such as DVSTs are indicated.

Fast-erase storage tubes (Table 7.6) and conventional DVSTs are fade-erased continuously (in a manner analogous to phosphor decay) across the whole display area. Fade erasure results in a loss of resolution and gray shades and in scan-to-scan interference. Symbology presented on fast-erase storage tubes and conventional DVSTs causes smearing when moved across the stored image.

Table 7.6 Typical Characteristics of Fast-Erase Storage Tube (from Ref. 10)

Characteristics	Fast-Erase Storage Tube
Resolution (shrinking raster) (stored lines/in.)	100
Brightness (fL)	800 (at maximum resolution; brighter if smearing occurs or if resolution is compromised)
Writing speed (in./s)	100×10^3
Storage time (s)	10
Shades of gray	6

Table 7.7 Typical Characteristics of Multimode Tonotron® (from Ref. 10)

Characteristics	Multimode Tonotron®
Resolution (shrinking raster) (lines/in.)	
Stored writing gun	120
TV gun	150
Symbology gun	100
Erase gun	70
Brightness (fL)	1000
Writing speed (in./s)	60×10^3
Erase speed (in./s)	10×10^3
Storage time (min)	2
Shades of gray	7
Deflection	Electrostatic (3 focused guns)

7.3.2 Specific Specialized CRT Technologies

In 1973 the Northrup Corporation combined the microchannel plate amplifier with the older, conventional CRT principles to achieve a relatively thin CRT of high sustained brightness. This tube was assigned the Digisplay® trademark. Jeffries[45] describes the Digisplay® essentially as follows:

> The display, known as the Digisplay®, utilizes an area electron source followed by a series of thin aperture control plates, which are aligned and act collectively to generate a group of scanning electron beams. The position of the group of beams is determined by the digital addressing signals applied to decoding electrodes that are deposited on control plates.
>
> The display envelope consists of two standard rectangular CRT-type faceplates, sealed to the center glass mounting plate. The area cathode is attached to one side of the mounting plate, and the fused stack of six control plates is attached to the opposite side of the mounting plate. The leads for addressing the display are deposited on the mounting plate and come out through the solder glass seal between the front faceplate and the mounting plate. The external dimensions of the tube are 4 × 5 × 1.5 in. The phosphor connection is brought through the front of the faceplate in order to eliminate voltage breakdown problems between the high voltage on the phosphor and the relatively low voltage on all other parts of the tube.
>
> Figure 7.20 is an exploded view of the display tube. The display has 150 × 150 resolution elements over a 2.7- × 2.7-in. format. The specific functions of the six aperture plates that form the heart of the device are as follows:

1. *First plate:* Forms electrons from cathode into 150 × 150 beams.
2. *Second plate:* Reduces the number of beams to a group of 10 beams high by 150 beams wide.
3. *Third plate:* Reduces the number of beams to a 10 × 10 beam group.
4. *Fourth plate:* Reduces the number of beams to 10 colinear beams.

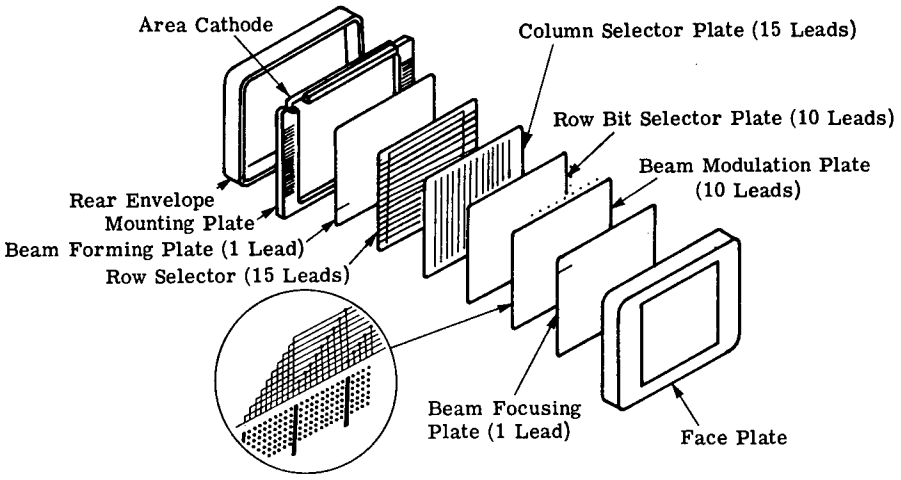


Fig. 7.20 Exploded view of 150×150 element Digisplay®.[45]

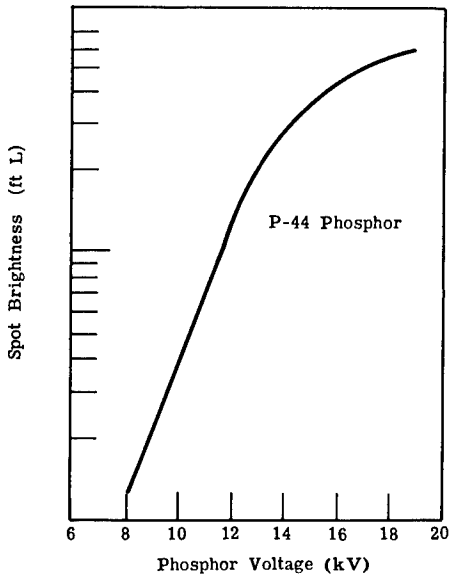


Fig. 7.21 Brightness versus phosphor voltage.[45]

5. *Fifth plate:* Modulates the 10 beams individually.
6. *Sixth plate:* Controls and adjusts the focus of the electron beams en route to the phosphor faceplate.

The six plates are fused together prior to tube assembly to maintain permanent hole alignment and to increase their ruggedness.

The tube utilizes a type P44 phosphor target. A plot of spot brightness as a function of phosphor voltage is given in Fig. 7.21. The maximum brightness measured was 6000 fL at a phosphor potential of 19 kV. Total input power to the tube was approximately 10 W at maximum brightness.

7.3.3 Electrical and Visual Output CRT Storage Tubes[65]

Often, a single display is needed to receive data from a variety of input sources. Some, such as low-light-level TV or forward-looking infrared (FLIR) sensors, produce data at rates of many megabits per second. Others, such as side-looking synthetic aperture radars, produce an equivalent number of data bits, but over a period that is a small fraction of an hour. Thus the data rate and the storage requirements imposed by these two sensors can differ by a ratio of perhaps a thousand to one.

Electrical storage tubes permit economical buffering of low-data-rate information for presentation at conventional TV rates. In alternate system modes, other sensors that operate at TV rates may then bypass the storage tube buffer, and the video data directly drive a conventional CRT.

Electrical storage tubes are sometimes used to store the more complex data in displays and, by video mixing at the monitor, these data may then be combined with information such as a nonstored alphanumeric overlay. An economical, high-quality presentation with desirable man/machine interactive properties results.

As digital solid-state memory costs decrease and as more information is in digital form, digital scan-conversion memory systems are becoming more attractive. When digital systems are used, analog information is converted to digital format by an analog-to-digital (A/D) converter. A digital memory is used for storage, following which the digital information is converted back to analog form for display. These techniques have been used primarily in the lower information density displays and where real-time interaction between the operator and the machine is critical and cost is secondary.

The economy and performance of analog scan converters and analog DVSTs once assured them a multitude of high-density, high-resolution applications. DVSTs were once the only means of displaying the rolling, continuous imagery of sensors on aircraft, for example. Now, however, electronic processing—especially that by image conversion boards—can provide the same functions more economically.

In some CR storage tubes, storage takes place directly on the viewing surface. The cathodochromic or dark trace storage CRT uses a viewing surface whose optical properties are changed following electron beam excitation, and this surface is viewed by either transmitted or reflected light.

The substrate in these 12- to 16-in.-diameter tubes is green backlighted for viewing. Writing speed is of the order of 10,000 in./s, and erasure is accomplished by an internal heater that raises the substrate temperature for 5 to 10 s. A high-resolution display with good gray-scale rendition may be presented.

In other devices, such as those used in some computer graphics terminals, storage takes place at the light-emitting phosphor.[66] The display is bistable, e.g., no halftones. A writing electron beam establishes a high-resolution electron charge pattern on the phosphor, which then modulates a lower velocity electron beam that is continually flooding the entire viewing surface. Luminance is adequate for a controlled office or laboratory environment. Writing speed is about 5000 in./s in these displays, which can present about 10 characters per inch.

When storage with higher light output and gray-scale capability is required, another class of DVST is available[67] in which storage is accomplished near the viewing surface.

These tubes utilize a low-voltage electrostatic charge or an insulator supported by a metal mesh, located several tenths of an inch away from the light-emitting phosphor or viewing screen, which is at a high voltage. An electrostatic charge, representing the image to be sorted, is deposited on the insulator by a focused high-velocity electron beam. This charge, located at the individual mesh webs, then modulates a lower velocity electron beam that is flooding the entire mesh surface. The more positively charged mesh elements allow the approaching low-velocity flood electrons to pass through those mesh openings. The electrons are then accelerated to strike the viewing phosphor at high energy. Less positively charged elemental areas restrict the quantity of electrons, resulting in lower light output. This design results in a transfer characteristic that permits a moderate gray-scale capability, depending on the local charge pattern.

The integrating characteristics of the storage tube surface are especially important in bringing signals or a picture up out of noise. For this function, successive lines or frames of information can be cumulatively added to each

other. The signal is spatially repetitive, while the random noise is not; thus the signal-to-noise ratio builds rapidly and a clearer picture is obtained. To erase, or to return the charge image to the unwritten or more negative state, a flood of electrons is caused to land at a low velocity. This is accomplished by the metal mesh driving positive; the insulator follows, and the flood electrons are able to land at low velocity, charging the insulator in a negative direction until landing stops. When the metal mesh is returned to its normal potential, the more negative charge creates a black or erased condition.

One specialized tube, the projection Tonotron® designed for projection applications, provides a 4-in.-diameter image with a luminance of 10,000 fL.

The high luminance in the output of this tube results primarily from continual phosphor excitation, rather than the usual repetitive, but short, high peak loading of more common projection CRTs. Because of this, the operating life of this projection storage tube is not adversely affected by operating at such high-output luminance levels, which tend to cause short lifetimes in conventional CRTs.

A further advancement in DVSTs allows a more versatile usage. An example is the Multimode Tonotron®, which utilizes a storage surface that enables writing and erasing by two different mechanisms. This is achieved with multiple electron guns operating at various beam energies.

The high-performance, economical analog scan converter may be used as a link in a display system. Electrical-output scan-converter storage tubes are used in display terminals as buffer-storage elements and for video storage, scan conversion, and integration. A display of one format is often converted to television scan rates because of the availability of such display monitors in many packaging configurations and the ease with which the converted TV signal may be transmitted to one or many display stations.[46,68]

In double-ended storage tubes, the memory target is at the tube center, with writing and reading guns on either end. This arrangement permits simultaneous writing and reading. Several types of memory targets are available:

- A transmission grid-modulation type exists that is similar to the DVSTs described previously. With it, the read beam electrons are modulated as they scan and pass through the central mesh. Since both the write and read beams strike the signal output mesh, the rf intensity of the read beam and the consequent demodulation at the tube signal output electrode are used to separate the currents coming from the write and read beams.
- The EBIC, or electron bombardment-induced conductivity type, uses a very-high-energy writing beam to discharge or write on a thin, continuous insulator with a metallized backing.
- The membrane scan-converter target[46] utilizes a thin membrane target so that a charge deposited on the write side transfers to the read side. The writing electron-beam energy is low; thus, writing electrons do not get into the read output, and special signal separation techniques are not necessary. Limiting resolution is 2000 TV lines per target diameter and writing speeds are up to 8 μs per target diameter.

In double-ended scan converters, the read collector is used for the output electrode. The transfer characteristic is shown in Ref. 46. Selective erasure is

readily achieved by adjusting the beam energy of one of the guns, permitting displays similar to those on the Multimode Tonotron® as well as "passing scene" and gradual fade types of display. These tubes are generally used in a non-destructive readout mode where the reading action does not remove written information. For some applications it is desirable that the reading process also remove the written charge; this is called *destructive read.*

Another destructive, erase, double-ended scan converter uses a silicon diode array target.[68] This is similar to a target used in a sensitive TV camera tube, and the target exhibits electron gain.

Writing speeds of about 1.3 ns per target diameter are achieved and permit applicability to transient recording signals up to 1 GHz in bandwidth. The target can retain information on a signal for as long as 100 ms. Readout is an orthogonal TV scan with the transient intersections for each scan line recorded and processed. Alternatively, the trace can be read out and displayed on a TV monitor.

Another variation of scan-converter technology is the single-ended scan converter. A single-ended scan converter is a tube having only one electron gun that is time-shared between accepting information for writing and the presentation of it during reading. Most use a solid target with an insulating island. Its written charge image controls the proportion of read beam allowed to land on the adjacent conducting surface, which then constitutes the signal current. A very-high-resolution storage tube,[69] of which only a few have been built, is based on RCA's 4.5-in. return-beam vidicon. The target is similar to other silicon-target storage tubes, except that 7000 silicon dioxide stripes are contained in the 2- × 2-in. target.

The limiting resolution of this storage tube is about 4 times that of the more popular silicon storage tubes, providing about 16 times the information capacity. Storage of 50 to 100 pages of printed text is feasible. Resolution, shading, gray scale, write time, erase time, repeatability, linearity, and storage time are the trade-off parameters in the selection of a video memory.

7.3.4 Plasma Panels

Gas discharge tube displays in the form of banks of neon or argon indicator lamps have long been used. The main problem in terms of their large-scale use in displays has been associated with the fabrication, wiring, and driving of such lamps. Plasma panels are a logical outgrowth to achieve the functions of large arrays of lamps without the conventional difficulties. Plasma panels can be divided into two basic forms: the ac type, typified by the early Owens-Illinois Digivue® panel, and the dc type, represented by the early Burroughs Self Scan® panel.

The two technologies are first briefly described and compared, and then each is examined in some detail. Slottow[70] compares the two technologies essentially as follows[e]:

> The term *plasma display* was first used to describe a gas discharge device in which the discharges are insulated from the exciting electrodes by dielectric surfaces. Although these surfaces prevent the development of continuous discharges, this

[e]See also "Plasma Technology for Displays" in the Bibliography.

structure can support a sequence of pulsed ac discharges in response to an alternating voltage. In recent years the term plasma display has also been used to describe devices in which the electrodes are immersed in the gas. This kind of structure can support continuous or pulsed discharges, but in most applications the discharges are unidirectional. To distinguish between the two kinds of structure we refer to the first class as ac plasma displays and to the second class as dc plasma displays. Because they are simpler conceptually, we begin our discussion with the dc plasma displays.

dc Plasma Panels. In one type of dc structure two sets of parallel electrodes oppose each other in the gas, one set being directed orthogonally to the other. An aperture plate, placed between the electrode sets, confines the discharges. Appropriate addressing voltages on two intersecting electrodes cause an electrical breakdown and emission of light at the intersection; an addressing voltage on only one electrode is too small to ignite the discharge. In this matrix arrangement, the gas discharge cell, which is sufficiently nonlinear for the purpose, functions as a two-input AND circuit. This device is usually operated one row (or column) at a time, with signals on the opposing electrodes determining which cells in the row (or column) will be on. Since the duty cycle in these devices becomes smaller as the device becomes larger, the peak currents limit the array size.

Except for a difference in geometry, segmented gas discharge displays are also operated and connected in the same way. Corresponding segments are connected to define a "row," while the segment electrode that opposes all the segments of a character defines a "column."

Information storage can be added to this structure in two ways. One technique depends on exciting the panel with voltage pulses that exceed the normal ignition voltage but are too narrow to allow new discharges to mature. A cell that is already "on," however, retains sufficient ionization products during the interval between pulses to ensure a discharge at the next pulse. A second technique requires a current-limiting resistance at each cell in series with the discharge. The combination of resistance and gas cell provides a bistable luminous element in which a voltage that is too small to ignite a discharge is more than large enough to maintain one. In both cases, the memory is associated with the volume properties of the discharge.

Although the matrix structure reduces the required number of connections to the panel, it would be desirable economically to reduce the number further. An important nonstorage dc discharge device, the Self Scan® panel, accomplishes this objective at the cost of increased panel complexity. In this device, a set of scanning discharges, only slightly visible to the viewer, is provided by a multiphase (usually three-phase) driver. The ignition of a scanning discharge requires both adequate voltage and the volume priming from the preceding discharge. A separate set of electrodes transfers the scanning discharges to the front of the panel according to the information content. These devices do not have internal memory.

Because the intensity of these nonstorage panels can be controlled by current or pulse width, provision of gray scale is not difficult, and their application to new kinds of television displays is being widely studied. Multicolor, important for television as well as for some information display applications, can be provided by phosphors excited by the ultraviolet components of gas discharges.

ac Plasma Panels. In the ac plasma panel, as in the dc panel, two sets of electrodes oppose each other across a discharge gap and are directed orthogonally to one another. However, at each discharge site, the dielectric surfaces that isolate the electrodes from the gas define two capacitances, which are in series with discharge. An alternating voltage applied across the two electrode sets is too small by itself to ignite discharges. However, a pair of write voltages applied across two selected electrodes will ignite a pulsed discharge that extinguishes as ions and electrons flow to the dielectric surfaces and charge the capacitors. This charge augments the applied voltage on the next half cycle to ignite a second

discharge, which then charges the capacitors in preparation for the third discharge. This sequence of pulsed discharges, which characterizes the "on" state of a cell, terminates when an erase signal on the two intersecting electrodes produces a controlled discharge that reduces the charge on the series capacitance below the minimum required for ignition.

In the Digivue® version of the ac plasma panel, no aperture plate is required in the gas. The discharge is confined instead by the electric field and by the pressure of the gas. This panel is self-registering, with discharges occurring at the electrode intersection. The simplicity of the structure has made possible the development of commercial devices with more than 250,000 discharge sites at a density of 3600 per square inch. Panels with over 1 million cells at a density of 6889 per square inch have been made experimentally.

The basic structure of the ac panel has also been realized in segmented form for numeric displays.

This ac plasma panel device is also a matrix array, but here each discharge site functions not only as a two-input AND circuit, but also as a bistable storage element. The memory depends, in this case, on the storage of electric charge on dielectric surfaces. Volume memory effects are present, however, and these are now being exploited in extensions of the technique.

The ac plasma display is also being studied for application to television and it, too, can provide multicolor through excitation of phosphors.

The plasma display technology is being developed in many sizes and for many applications. For large graphic displays it is the only technology seriously challenging the cathode-ray tube.

7.3.5 Liquid Crystals

For many years, a class of cholesteric organic material has generated research interest because of the crystal-like properties of the liquid materials. The chemistry and properties of liquid crystals are treated extensively and thoroughly by Creagh.[71] They have been applied to a number of problems with useful, but not dramatic, results.

In 1968 and 1970, however, Heilmeier et al.[43,72] showed in considerable detail that, by electrically exciting the materials, a controlled amount of optical scattering could be produced that made the liquid crystals seem brighter and darker than surrounding material. From this early work, a variety of technologies have emerged, from wristwatch and pocket computer displays to TV-compatible displays and projection systems.

The work of Kobayashi[73] is singled out to illustrate just one class of data on but one class of liquid crystals operated in but one mode. The amount of data needed to make a choice depends on temperature, power efficiency, color, dynamic range needed, form of crystalline material, and form of excitation. While Kobayashi considered dynamic scattering that is a result of the turbulence caused by passing a current through the liquid crystals, a related paper by Jones and Lu[74] considered field-effect liquid crystalline devices, which employed twisted nematic crystals. They observed the resultant changes in polarization.

Although this work was an advance, the slow response made liquid crystals suitable only for low-information-rate displays. Lipton and Koda[75] reviewed this problem at the 1973 meeting of the Society for Information Display (SID) and indicated that if the technology was to be useful for such high-rate applications as television or other dynamic imaging applications, the problems

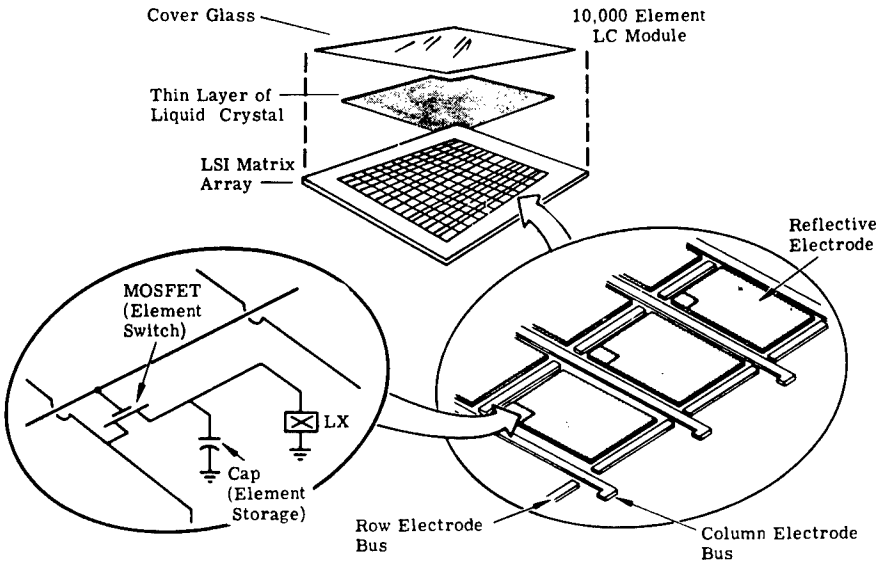


Fig. 7.22 Construction of liquid crystal display module, early 1970s.[78]

of crosstalk between elements and the slow response time of the elements must be overcome. For an excellent review, see Kmetz and Von Willisen.[76]

Lipton et al. presented another paper,[77] on a liquid crystal television display using a silicon-on-sapphire switching array, at the 1975 SID meeting. The amount of progress made in the two years since the appearance of Ref. 75 can be judged from the later paper:

> Live television has been demonstrated on a liquid crystal display. The display uses a transistor switch at each matrix point in order to solve the traditional liquid crystal problems of crosstalk between elements, slow response time, and limited multiplexing capability.[f]
>
> The use of a transistor at each matrix point of the display represents the most general approach towards solving these problems. When combined with a suitable storage capacitor, such a transistor-capacitor combination can act as a sample-and-hold circuit to allow rapid addressing of a selected element. This approach has previously been shown capable of allowing TV rate operation of liquid crystal displays.

Figure 7.22 shows the construction of a liquid crystal display module. An array of thin-film transistors (TFTs) is used to control the optical properties of a liquid crystal layer that is sandwiched between the TFT substrate and a second substrate with a transparent conductor layer. The applied voltage at each element of the display is used to control the orientation of the liquid crystal. Figure 7.23 shows, for its historical interest, some of the first liquid crystal imagery obtained.

[f]This was accomplished in the early 1970s by Ernstoff[78] in the Hughes display laboratories. Progress in LCDs accelerated thereafter.

(a) (b)

Fig. 7.23 Enlarged copy of 100 × 100 liquid crystal imagery (a) with picture imperfections and (b) defect free.[78]

In recent years the drawbacks of liquid crystals have been greatly reduced in severity. Liquid crystals are now used in projection devices, as discussed in later sections.

7.3.6 Comparison of Selected Currently Available Flat Panel Displays

Table 7.8 compares the parameters of selected liquid crystal, electroluminescent (EL), and plasma displays reported by Beta Review[47] to be available as of September 1989 and to have a minimum diagonal size of 250 mm and at least a 640 × 480 pixel matrix within minimum screen dimensions of about 8 × 6 in. The displays are listed by increasing diagonal size, contrast ratio, and luminance within three categories: (1) contrast unreported or less than 10; (2) contrast at least 10 and luminance unreported or less than 15 fL; and (3) contrast at least 10 and luminance at least 15 fL. Only 8 displays of the 23 listed meet the criteria of category 3, and only 2 of the 8 are LCDs.

Conversations with manufacturers indicate that the usable LCD screen dimensions had grown from 6 × 8 in. in September 1989 to 10 × 10 in. in June 1990 while maintaining pixel density. We understand that Beta Review, Inc., plans periodic updates of the program *How to Select a Flat Panel Display,*[47] which we used to create Table 7.8.

7.3.7 The Relative Merits of LEDs and LCDs

The strong potentials of LCDs and LEDs are thoroughly reviewed by Nuese et al.[79] See also Goodman[80] and "Light-Emitting Diodes" in the Bibliography. Loebner[81] has reviewed the properties of electroluminescent materials. His review particularly treats the colors available from various materials rather thoroughly.

The strong potential of both LCD and LED technologies for display purposes needs a careful comparison when making a choice for display design. Such an analysis has been made in Ref. 80 by Goodman, whose paper has been modified for use here in the rest of this section in severely abridged form. Note that Goodman's paper was written in 1974 and that parameters such as the speed

Table 7.8 Selected Flat Panel Displays by Screen Diagonal Size, Contrast Ratio, and Luminance (from Ref. 47)

(Minimum diagonal 250 mm; pixel matrix at least 640 × 480; minimum screen dimensions about 8 × 6 in.)

Category[a]	Screen Diagonal, mm	Pixel Matrix, dots		Horizontal Pixel Density (pixels/cm)	Contrast Ratio	Luminance, (fL)	Technology	Subtechnology[b]
		Horizontal	Vertical					
1	272.8	640	480	29	5	—[c]	LCD	STN
1	272.8	640	480	29	5	—	LCD	STN
1	280.0	640	480	29	8	—	LCD	STN
1	288.5	1120	780	47	—[c]	—	LCD	DST
1	324.9	1024	800	40	—	30	EL	ACTF
1	386.3	1024	768	33	—	20	Plasma	AC memory
2	250.4	640	480	32	20	—	LCD	TFT-TN
2	264.0	640	480	30	10	—	LCD	DST
2	264.4	640	480	30	10	3	LCD	DST
2	272.8	640	480	29	12	—	LCD	DST
2	272.8	640	480	29	20	—	LCD	DST
2	279.8	640	480	29	10	—	LCD	MST
2	280.2	640	480	29	10	—	LCD	MST
2	288.0	640	480	28	12	—	LCD	DST
2	288.3	640	480	28	12	—	LCD	DST
3	263.6	640	480	30	10	21	Plasma	DC
3	263.6	640	480	30	10	21	Plasma	DC
3	263.6	640	480	30	20	32	Plasma	AC memory
3	263.6	640	480	30	150	21	Plasma	DC
3	263.9	640	480	30	15	15	LCD	DST
3	264.0	640	480	30	15	17	LCD	DST
3	441.8	1024	1024	33	25	75	Plasma	AC refresh
3	452.0	1024	800	30	60	18	EL	ACTF

[a]Category 1: Contrast unreported or less than 10.
Category 2: Contrast at least 10; luminance unreported or less than 15 fL.
Category 3: Contrast at least 10; luminance at least 15 fL.

[b]AC = alternating current
ACTF = ac thin film
DC = direct current
DST = double-layer supertwist
MST = monochrome supertwist
STN = supertwisted nematic
TFT-TN = thin-film transistor/twisted nematic [active (extrinsic) matrix]

[c]— = unreported

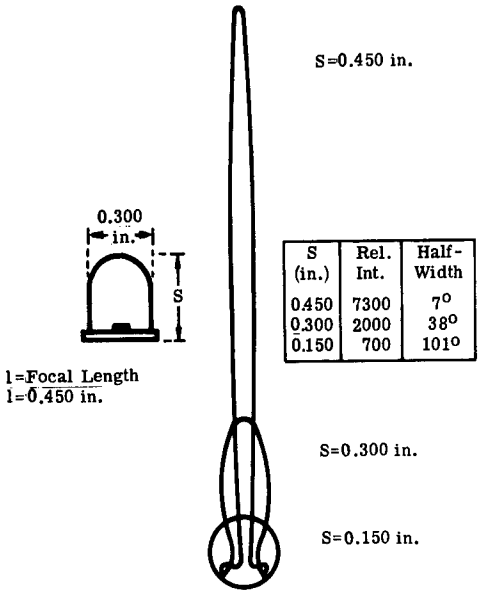
of response have increased tremendously, so that operation of large liquid crystal displays is now feasible at normal TV rates:

> Since LEDs emit radiation, they are more visible in dim ambient light and less visible in bright ambient illumination . . . the luminescence intensity is proportional to the current passing through the diode. The visibility of an LED for a given power or current input is dependent primarily on two factors, the external quantum efficiency of the human eye to the wavelength of the emitted radiation (the luminous efficiency in lumens per watt). The luminous power efficiency of the LED (the ratio of the light output in lumens to the input power in watts) is proportional to the product of the external quantum efficiency and the luminous efficiency.
>
> Table 7.9 lists performance data for $GaAs_{1-x}P_x$ diodes grown either by the vapor-phase epitaxy or liquid-phase techniques. The $GaAs_{0.6}P_{0.4}$ diodes are more efficient at high currents than they are at low currents because of the occurrence of nonradiative recombination, which is less important at high current levels than it is at low current levels. Due to this fact, and the integration capability of the eye, diodes operated in a pulsed or strobed mode require less average current than diodes operated in a nonpulsed manner to produce the same time-averaged luminous output. Whether operated in a pulsed or nonpulsed mode, the upper limit on the light output is set by the luminous power efficiency and the maximum current that can be passed through the diode.
>
> One of the important factors limiting the efficiency of LEDs is the mismatch in index of refraction between air ($n=1$) and the luminescent material ($n\cong 3.6$). As a result of this mismatch, the critical angle for total internal reflection is quite small (~ 16 deg), so that almost all of the randomly emitted junction radiation is reflected back into the semiconductor. This is particularly harmful with the common LEDs made from $GaAs_{0.6}P_{0.4}$ material because the energy of the radiation is close to the bandgap of the material, and the totally internally reflected radiation is strongly absorbed by the semiconductor.
>
> When a diode is encapsulated in a plastic lens, the total internal reflection is reduced because of the lens' index of refraction ($n\cong 1.8$ for epoxy) and the shape of the lens. A simple hemispherical dome lens improves[82] the external efficiency by a factor of approximately 2 to 3. The plastic encapsulation can also be used to shape the angular distribution of the emitted radiation, as shown in Fig. 7.24.
>
> When GaP substrate is used instead of the standard GaAs substrate, it is possible to improve the external quantum efficiency, not only by utilization of plastic encapsulation, but also by making the undersurface of the substrate optically reflecting.[83] Plastic encapsulation and usage of a GaP substrate improve the efficiencies of $GaAs_{1-x}P_x$ LEDs above those listed in Table 7.9.
>
> Because LCDs are passive devices that modulate the passage of light, they are at least as legible in bright ambient illumination as they are in dim ambient lighting. Furthermore, their legibility is strongly influenced by the spatial distribution of the lighting.
>
> Typical transmitted light curves are shown in Fig. 7.25 for a dynamic scattering device. As the data demonstrate, high contrast ratios can be obtained with a collimated source. The curves also indicate that the contrast ratio, as a function of voltage, is not monotonically increasing for all angles. Indeed, for some angles, the inverse is true. For backlit dynamic scattering displays, the strong angular dependence of the contrast ratio curves is not a serious problem, because it is relatively easy to use a quasicollimated source so situated that the contrast ratio for normal viewing angles increases with voltage.
>
> However, the proper orientation of the illuminating source, the display, and the observer is not as easy to arrange with reflective dynamic scattering displays in ambient lighting. Because dynamic scattering is a forward-scattering phenomenon, it is common to use a specular mirror back electrode to reflect the scattered light back to the observer when the illumination source is on the same side of the display as the observer (Fig. 7.26). The observer should be at an appropriate position so that only the reflected, scattered light is within his field of view and none of the specularly reflected, unscattered light is observable by him.

Table 7.9 Performance Characteristics of Different Types of LEDs* (from Ref. 80)

Type	Color	Peak Emission Wavelength (Å)	Quantum Efficiency		Luminous Power Efficiency (lm W^{-1})	
			Research	Commercial Performance	Research Result	Commercial Performance
$GaAs_{0.6}P_{0.4}$	Red	6490	5×10^{-3}	2×10^{-3}	0.33	0.15
$GaAs_{0.35}P_{0.65}$:N	Red-orange	6320	4×10^{-3}		0.76	—
LPE:GaP:Zn, O	Red	6925	15×10^{-2}	$0.5–2 \times 10^{-2}$	3.0	0.1–0.4
VPE:GaP:N (Zn diffusion)	Green	5700	1×10^{-3}	$0.3–0.8 \times 10^{-3}$	0.6	0.18–0.48
PE:GaP:N (grown junction)	Green	5700	3×10^{-3}	$0.5–2.0 \times 10^{-4}$	1.8	0.03–0.12
$GaAs_{0.15}P_{0.85}$:N	Yellow	5890	0.8×10^{-3}	$0.5–0.8 \times 10^{-3}$	0.36	0.23–0.36
VPE:GaP:N ($N > 10^{20}$ cm^{-2})	Yellow	5900	1×10^{-3}	—	0.45	—

*The values given in this table are for dc current densities of less than 20 Å cm^{-2}.



S (in.)	Rel. Int.	Half-Width
0.450	7300	7°
0.300	2000	38°
0.150	700	101°

Fig. 7.24 Radiation patterns for a $GaAs_{1-x}P_x$ diode encapsulated with a plastic lens.[80]

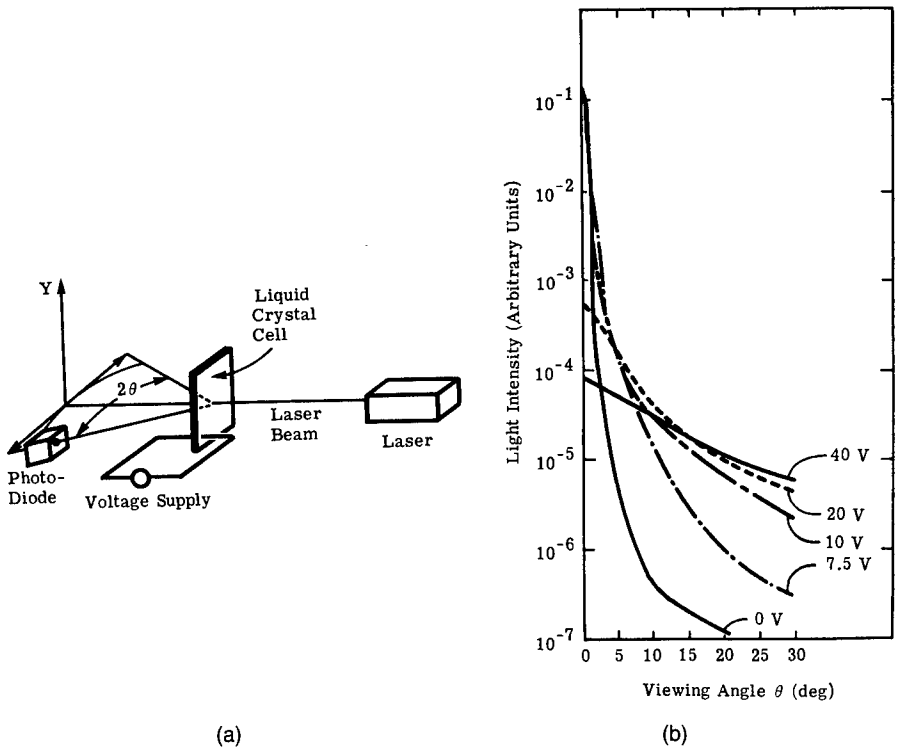


Fig. 7.25 (a) Diagram of measuring apparatus for transmissive dynamic scattering cells and (b) typical scattered light intensity as a function of viewing angle and voltage.[80]

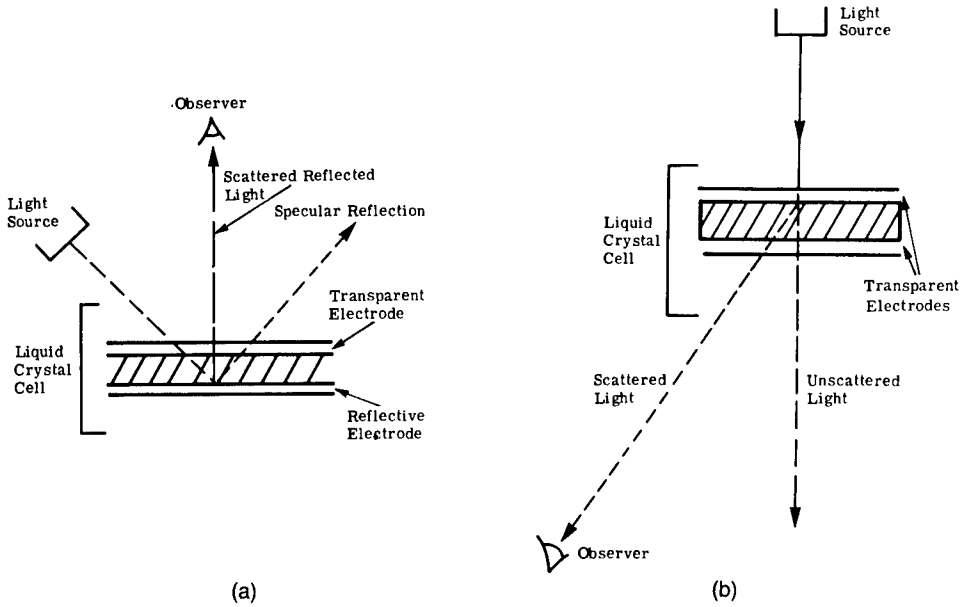


Fig. 7.26 Diagrams of the viewing conditions for dynamic scattering cells for twisted nematic liquid crystal operation (a) reflective device with a specular mirror and (b) transmissive device.[80]

The twisted nematic effect, in which the liquid crystal material is oriented parallel to the electrodes, but with a 90-deg twist between the orientation directions at the two electrodes, is shown schematically in Fig. 7.26. Since the liquid crystal material is birefringent, it rotates the polarization direction of the incidental polarized light by 90 deg. Thus, with no applied field, the incident light is absorbed by the second polarizer, which is oriented parallel to the first one. When an electric field is applied, however, the twist of the material is removed, and the liquid crystal does not rotate the polarization of the light. This results in the transmission of light through the (nominally) parallel polarizers. A plot of transmission versus applied voltage for a typical display cell is shown in Fig. 7.27. Gray levels can be obtained by applying intermediate values of voltage to the pixel. Goodman continues:

> Although the contrast ratio versus voltage curves for twisted nematic displays are strongly dependent on the viewing angle, as shown in Fig. 7.27, the glare problem in reflective applications is minimal. Because the twisted nematic effect does not induce the scattering of light, a diffuse reflector can be used behind the display instead of a specular reflector. Consequently, the observer on the same side of the display as the illumination can see the voltage-induced change in the light transmission of the device without the presence of reflective glare from the back reflector. It should be noted from the curves in Fig. 7.27 that, with a sufficiently high voltage, the transmitted light curves are fairly isotopic with viewing angle and that the viewing cone can be as much as ± 50 deg.
>
> There is no intrinsic wavelength dependence in the visible region for either of the effects; however, color can be introduced into both effects by external means. With dynamic scattering, either dichroic mirrors or colored filters can be used if so desired. For twisted nematic displays, colored filters over each pixel can be utilized.

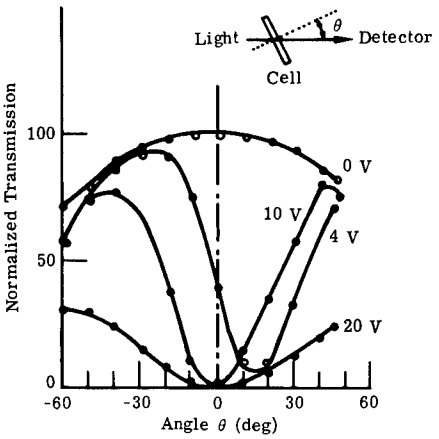


Fig. 7.27 Light transmission of twisted nematic cell with crossed polarizers as a function of turning angle for various values of applied field at 5 kHz (cell thickness is 30 μm).[80]

Power Dissipation. The power dissipation in an LED is the product of the voltage drop across the diode and the current passing through it.

It is not possible to give a precise value for the power dissipation of LEDs because of the wide range of possible uses, but typical power densities range between 0.1 and 10 W/cm^2 for 2-V dc across the diode. For example,[84] a typical power dissipation for a segment in a calculator display with an integral lens is 0.6 to 0.7 mW, whereas the power dissipation of a segment in an LED watch display without a lens is about 3 mW.

With LCDs, the power dissipated by the liquid crystal display itself is very small. For devices using the dynamic scattering effect, typically 0.1 to 1 mW/cm^2 is used at a drive voltage of 15 V and at 30 Hz. The power dissipation is due to the fact that, for satisfactory operation, dynamic scattering devices require that the resistivity of the fluid be in the 10^8 to 10^9 Ω-cm range. The per-segment power dissipation of an eight-digit, seven-segment calculator using a dynamic scattering display is about 0.02 mW.

By the nature of the effect, twisted nematic displays do not need any resistive current flow in the material when the voltage is applied. Therefore, the material resistivity can be as high as obtainable by chemical purification and is usually above 10^{10} to 10^{11} Ω-cm. Hence, at the normal operating frequency of 30 Hz, the power dissipation is entirely capacitive. With an applied voltage of 5 V, the capacitive power dissipation is about 5 μW/cm^2.

Response Times. The rise and decay times for LEDs are primarily controlled by electronic processes and are very fast. Typically, the response times are in the 10- to 1000-ns range.

The same is not true for LCDs. The transient response of the electro-optic effects is determined by the motion of the bulk fluid; consequently, the speeds are in the millisecond-to-second regime. The rise is proportional to the fluid viscosity and the square of the thickness, and is approximately inversely proportional to the difference between the square of the applied voltage and the threshold voltage squared. The decay time is proportional to the viscosity and the square of the liquid crystal layer thickness. Because of the thickness dependence of the speed, it is highly desirable to keep the fluid thickness as small as technically feasible. In practice, typical thickness values are of the order of 10 to 20 μm.

At 20°C with an applied voltage of 10 V_{rms} and for a 12-μm-thick layer, rise times on the order of 10 ms have been observed with twisted nematic devices and

200 to 300 ms for dynamic scattering displays. For the same value of applied voltage, the much lower threshold voltage for the twisted nematic effect compared to the dynamic scattering mode results in a significantly faster rise time. This difference is not true for the decay times, where, for the same thickness and temperature given above, the decay times for both effects are of the order of 100 to 500 ms.

It is difficult to give exact numbers for the response times not only because of the expected variations from one material to another in the threshold voltage and other material-dependent parameters, but also because the response times are angle-dependent and can vary by as much as a factor of three, depending on the angular relationship between the display, the light source, and the observer.[85,86]

Temperature Dependence. The device parameters are all temperature dependent, but not enough to degrade device performance significantly.

The temperature dependencies of the response times are not noticeable because the variation in transient response is not very large with temperature[87] and, more importantly, the response times are so short that the temperature dependence is not detectable by the eye.

At constant current, the light output of a $GaAs_{0.6}P_{0.4}$ LED decreases at a rate of approximately 1%/°C between 0 and 50°C, which is unimportant in most display applications.[88]

The last relevant parameter is the forward voltage temperature coefficient of the current-voltage characteristic. The change with temperature is about minus 2 mV/°C, but this is also unimportant because LEDs are usually driven by a constant current source.

Unlike LEDs, whose operation is basically unaffected by ± 25°C excursions about 20°C, many of the electro-optic properties of the LCDs are distinctly affected by the same temperature variations. Indeed, until about five years ago, there were very few materials that exhibited the mesophases at 20°C. However, today many liquid crystal materials possess a mesophase range of at least 70°C with the low end of the range at or below 0°C.

Of more significance is the fact that the viscosity, conductivity, and response times are exponentially dependent on temperature with an activation energy of 0.5 to 1.0 eV. In dynamic scattering devices, this means that both the power dissipation and the cutoff frequency are strongly temperature dependent. Furthermore, the speed of operation of devices must be carefully considered not only at 20°C but also at lower temperatures where the speed of response becomes somewhat slow.[89,90]

The threshold voltage for dynamic scattering is quite constant with temperature, but for the twisted nematic effect, the threshold voltage decreases at a rate of about 0.5%/°C as long as the operating temperature is not too close (within 10°C) to the critical temperature separating the mesophase and the isotopic phase.

Circuit Compatibility. LEDs are high-current, low-voltage devices that are very compatible with discrete and integrated bipolar transistors and discrete metal oxide semiconductor (MOS) transistors. However, they are not readily driven by MOS integrated circuits (ICs), particularly complementary metal oxide semiconductor (CMOS) ICs, because of the current required by the LEDs. It is easy to matrix-address LED displays.

The electrical properties of LCDs and LEDs are very different. LCDs are low- to medium-voltage devices with low current requirements. Consequently, they can be easily driven by CMOS ICs. To obtain long operating life for LCD displays, it is necessary to excite them with bipolar waveforms with the dc component being less than 0.2 V.

LCDs do not possess a strongly asymmetric current-voltage characteristic like LEDs. Rather, in the useful range of operation, they can be modeled as leaky capacitors.

The electro-optic transfer function of an LCD changes rapidly for applied voltages greater than the threshold voltage and is symmetric about 0 V. The non-

linearity in the electro-optic transfer function does permit some multiplexing of LCDs, but to nowhere the degree that is feasible with LEDs. Two prime factors affect the multiplying capability of LCDs. One condition is that the applied voltage signals to the matrix coincide so that only certain specified matrix elements turn on and no others. In addition, it has been established that, unlike LEDs, both field effect[91] and dynamic scattering[92] devices respond to the driving signal in an rms fashion. This statement is true as long as the response times are much longer than the duration of the scanning pulse, which is typically a few milliseconds or less. Because of the coincidence requirement, the rms value of the scanned waveform on a selected element is less than three times the threshold voltage and decreases rapidly as the number of the digits to be multiplexed increases.[91,92] Under these driving conditions the contrast ratio is very viewing angle dependent, as indicated by the data in Figs. 7.26 and 7.27.

Multiplexing of both dynamic scattering and field effect devices is further complicated by the temperature dependence of some of their properties. In the field effect case, the threshold voltage varies at a rate of about 0.5%/°C, which can be significant, since it is desirable to maintain a small constant voltage difference between the threshold voltage and the bias level across the nonselected elements. The threshold voltage for the dynamic scattering effect is not as temperature sensitive as for the twisted nematic effect, but the cutoff frequency varies quite rapidly with temperature. Since the dynamic scattering effect only occurs for driving signals whose frequencies are less than the cutoff frequency, the temperature dependence of the cutoff frequency must be taken into account in selecting the rate at which the multielement display is scanned.

The so-called two-frequency approaches can be used to multiplex both dynamic scattering[93,94] and twisted nematic devices,[95] but they require larger voltage amplitudes and the dissipation of far more power than the single-frequency method described previously.

For displays using dynamic scattering or the twisted nematic effect, anywhere from four to seven digits have been reported as the maximum addressing capability.[91,96]

Packaging. The rapid development of LEDs has been greatly assisted by the utilization of integrated circuit processing techniques. The most common LED technology today uses vapor-phase epitaxity of *n*-type $GaAs_{1-x}P_x$ on either GaAs or GaP substrates. Zinc is diffused onto the *n*-type layer to form the *p-n* junction. This technology is reproducible and provides relatively high-quality devices at high yields. Furthermore, the surface of the device through which the Zn diffusion is done is usually passivated with a double layer of SiO_2 and Si_3N_4, which helps to provide long-term device stability. Liquid-phase epitaxy can also be used for the creation of the diodes, but it is primarily restricted to the manufacturing of red GaP diodes doped with zinc and oxygen.

Given a specific technology for making the doped *p-n* junction, a number of approaches can be used to package the LEDs. This variety is prompted by the desire to maximize the display's contrast ratio and to minimize the power consumption and the cost of the device.

The packaging approaches for segmented numeric displays can be divided into three basic categories: the monolithic, the hybrid or silver, and the light-pipe or stretched segment structures. Figure 7.28 is a schematic illustration of all three types.

In the monolithic approach, the whole seven-segment numeric consists of a single chip of $GaAs_{1-x}P_x$. The zinc diffusions are done so as to define each of the segments. Since this approach uses a lot of the luminescent material, it is usually restricted to small (0.1-in.-high) numerics with lenses to provide magnification.

In hybrid or silver displays, each segment consists of a discrete bar of the luminescent material with the diffused *p*-type regions. The bars are mounted in the seven-segment layout on some low-cost substrate such as a circuit board. This approach is utilized in displays ranging in numeric height from 0.1 to 0.3 in.

There are different variations of the light-pipe method.[82] One of these is illustrated in Figs. 7.28(c) and (d), but they all involve the same concept. A single light-emitting chip is placed at the bottom of a cavity whose top surface is in the shape of a rectangular segment. By means of the appropriate passive optics, the radiation emitted by the chip irradiates the top surface of the cavity so that it appears to an observer that the whole segment is uniformly emitting. This technique is favored for large character heights (≥ 0.3 in.) because it requires the least raw semiconductor material of the three methods.

LCD packaging technology is particularly flexible with respect to size variations. The basic LCD is a sandwich structure with the liquid crystal material enclosed between two glass plates that are partially covered with conductive coating. The conductive coating is easily patterned so that the character height, font, and layout can be readily changed. Commercial liquid crystal displays vary in size from several tenths of an inch in height to more than 10 in. on a side.

Moisture, oxygen, and ultraviolet light can seriously degrade many liquid crystal materials, thereby impairing device performance. Appropriate sealing

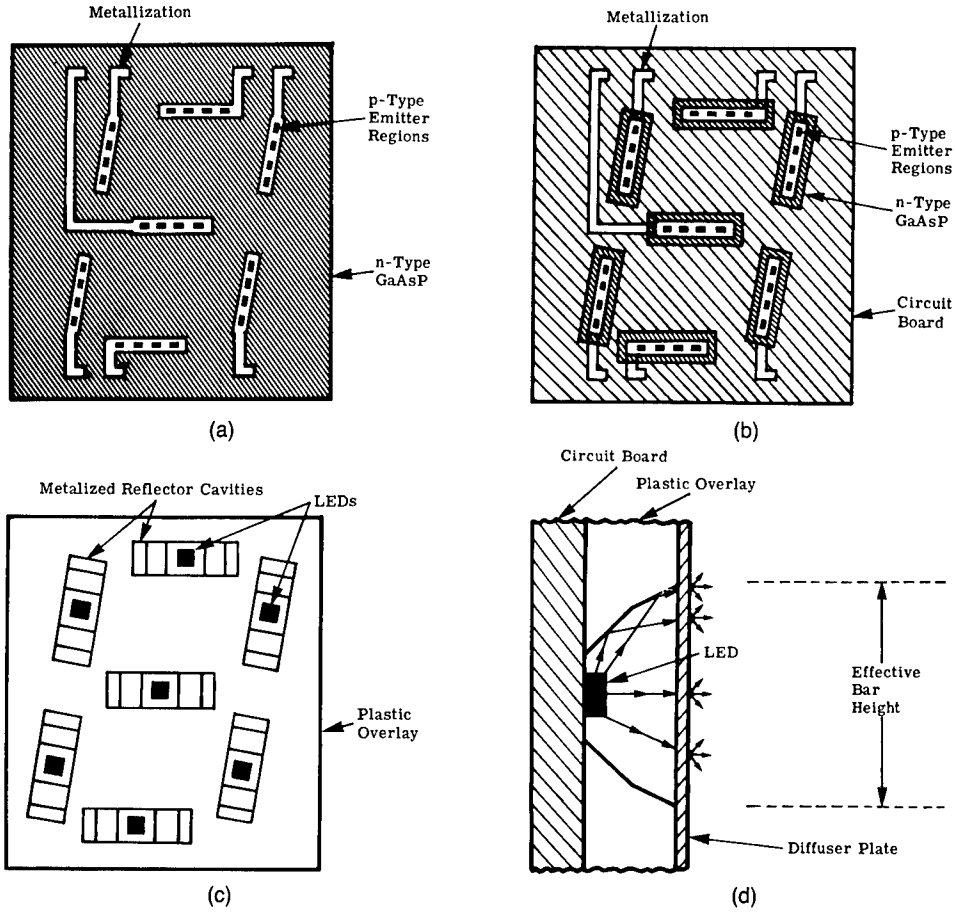


Fig. 7.28 Schematic drawings of different types of LED numeric packages[80]: (a) monolithic structure in which the entire device is a single chip of $GaAs_{1-x}P_x$; (b) bar segment device in which a bar of $GaAs_{1-x}P_x$, containing a series of diffusd regions, is used to define each segment of the numeric; (c) light-pipe structure, top view; and (d) cross section of a light-pipe structure.

techniques must be used to diminish the effects of moisture and gases on the properties of the liquid crystal material.[97]

Reliability. The radiative efficiency of LEDs can decrease with time for several reasons, including surface leakage, the diffusion of contaminating impurities such as copper into the *p-n* junction region, and the formation of intrinsic nonradiative recombination centers. However, the decrease in LED performance associated with these problems can be made fairly small. Test results for properly prepared diodes indicate that diodes can operate continuously for more than 50,000 hours at room temperature.[87]

The life of LCDs depends on the stability of different properties of the display, which include the conductivity, the temperature range of operation, the display's cosmetic appearance, and the response times. Changes in these parameters may occur either with or without the application of a voltage. As described in the previous section, proper packaging is necessary to minimize the effect of deleterious contaminants on the fluid's chemical stability. Also, the elimination of the dc components of the excitation is important to diminish the likelihood of electrochemically induced failure. Consequently, virtually all commercial LCDs are excited by an ac signal whose driving frequency is greater than 30 Hz. When the proper precautions are taken, the operating life for LCDs using dynamic scattering is cited[97] as being greater than 15,000 hours at 20°C. Adequate data are not present in the literature giving the life of displays with the twisted nematic effect; however, it is expected that field effect devices should have operating lives at least as long as LCDs using the dynamic scattering mode. An exact prediction of the display life is difficult to make because many of the degradation mechanisms are still not well characterized.

Economic Factors. The cost of LED displays has dropped by factors of 10. This sharp decrease in price has occurred because of improved manufacturing methods, high-volume production, better packaging techniques that allow the use of less LED material, more efficient LEDs, and a decrease in the cost of the semiconductor wafers. The price reduction for LED devices has occurred at approximately the same rate as did the price decrease for silicon integrated circuits at the same time after their introduction. This trend is expected to continue.

LCDs have also decreased in price to the extent that a typical cheap digital watch sells for less than a dollar. Approximately one-third of the cost of an LCD is for the raw materials, the biggest fraction of this being due to the transparent conductive coating and to the polarizers used in field effect displays.

A few more processing steps, whose cost scales with size, are needed to manufacture twisted nematic displays than are necessary for dynamic scattering devices. Also, the polarizer cost is not negligible. For many applications that use displays up to only 2 or 3 in. in the lateral dimension, the extra cost of the additional steps is unimportant in the choice between twisted nematic and dynamic scattering displays.

The direct expense of the display itself is only one part of the total display system cost, which is the final economic factor that the user must consider. For example, the cost of the driving circuit is important. LEDs can be easily multiplexed and require about 2 V for excitation, whereas LCDs are not easily multiplexed, and most need more than 3 or 4 V for excitation. This advantage for LEDs is offset by the fact that they require much more current than LCDs and, as a result, extra bipolar transistors are normally needed to interface the LED with the MOS driving circuit, while most LCDs are directly MOS compatible.

Summary. A summary of the properties of LEDs and LCDs is presented in Table 7.10.

LEDs have numerous positive features compared to LCDs. These include having (1) good viewability in low ambient light levels; (2) lower cost at present for small displays; (3) good multiplexing properties; (4) a wide range of operating temperatures; (5) high speed; and (6) proven reliability.

Table 7.10 Summary of LED and LCD Characteristics (from Ref. 80)

Category	Comments	
	LEDs	LCDs
Visual appearance	Medium to wide viewing angle. Visible in dim ambient illumination but not as visible in bright ambient. All colors available except blue.	Medium viewing angle. Viewability insensitive to intensity of ambient illumination. All colors available.
Power dissipation	0.1 to 10 W cm^{-2} at 2 V	5 to 10 μW cm^{-2} at 3 to 15 V
Response times	10 to 1000 ns	10 to 500 ms
Temperature dependence	Unimportant. Operating range of -40 to 100°C	The temperature dependence of the operating temperatures can be significant. Operating range about 0 to 70°C.
Circuit compatibility	High-current, low-voltage devices. Bipolar transistors usually required. Unipolar wave forms adequate for excitation. Easily multiplexed.	Low-current, low-to-medium voltage devices. CMOS IC compatible bipolar wave forms necessary.
Packaging	Semiconductor processing techniques. Different structures are used to maximize light output with a minimum of LED material.	Glass and organic fluid technology. Need for hermeticity. Flexible with respect to size variations.
Reliability	> 50,000 h	> 10 to 20,000 h
Economics	Prices have dropped sharply in last few years. Well along learning curve. Cost is area sensitive.	New relatively immature technology. Relatively low-cost raw materials.

On the other hand, LCDs possess the following advantages: (1) enhanced visibility in high ambient lighting; (2) much lower power consumption; (3) direct compatibility with MOS-integrated circuitry; (4) good size and format flexibility; and (5) lower cost for larger displays. Since neither device is ideal, the predominance of either technology will depend on which of the above criteria are more important in a specific application.

One further comment in favor of LCDs must be made. At present there is no display technology, including CRT technology, that offers as complete and as true a color spectrum as the newest of the color LCDs.

7.3.8 Special Projector Display Technology

7.3.8.1 General Electric Single-Gun Color Display Projector. The need for a ground-based display with characteristics suitable for color images after extensive and sophisticated data processing may be met by an interesting, innovative approach described by Good.[98] His paper is reproduced below, extensively edited and abridged:

The single-gun color TV light-valve projector was introduced and reported by True.[99] It uses a separate xenon light source, a fluid control layer, and a projection

lens. Optically it is much like a slide or movie projector. Miniature grooves are created on the deformable surface of the control layer by the electrostatic forces from the charge deposited by the scanning electron beam, which is modulated by video information. These groove patterns are made visible by use of a "dark field" or Schlieren optical system consisting of a set of input slots and output bars. The resulting television picture is imaged on the screen by the projection lens.

Figure 7.29 shows the xenon lamp, the sealed light valve, and the Schlieren projection lens. The cross sections of the light body, the color filters, and the input slots and output bars are shown below the light valve. Green light is passed through the horizontal slots and is controlled by modulating the width of the raster lines themselves. This is done by means of a high-frequency carrier applied to the vertical deflection plates and modulated by the green video signal. Magenta (red and blue) light is passed through the vertical slots and is modulated by diffraction gratings created at right angles to the raster lines by velocity modulating the electron spot in the horizontal direction. This is done by applying a 16-MHz (12-MHz for blue) signal to the horizontal deflection plates and modulating it with the red video signal. The grooves created have the proper spacing to diffract the red portion of the spectrum through the output slots while the blue portion is blocked. For the 12-MHz carrier the blue light is passed and the red is blocked. Thus, three simultaneous and superimposed primary color pictures are written with the same electron beam and projected to the screen as a completely registered full-color picture.

The resolution or sharpness of the projected image depends on the quality of the projection optics, the diffraction limit of the output slots, and the definition of the actual patterns created on the fluid surface itself. True and Bates[100] have been able to improve this latter situation by developing modulation techniques that consider the filtering action of the output bars and slots, in the frequency plane, to the sidebands that exist in the first- and second-order optical spectra because of the video modulation. By a combination of the above technique and the use of higher performance optics, the horizontal resolution of the green image of the single-gun color unit has been increased to over 900 TV lines. The horizontal resolution of the red and blue pictures has been increased to over 500 TV lines, being restricted primarily by the diffraction limit of the output slots. The com-

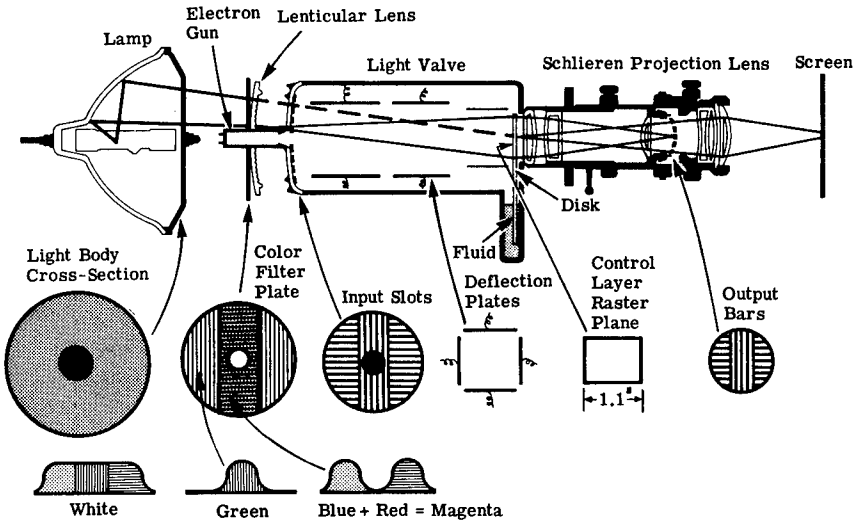


Fig. 7.29 Diagram of GE early single-gun color light-valve projector.[98]

bined horizontal resolution of a white picture tends to be more of an average and may be as high as 800 TV lines because of the luminance contribution of the green picture. These improvements are particularly helpful in reproducing the output of high-quality television sources such as a red-green-blue (RGB) color TV camera, a computer-generated image, or computer-generated graphics and alphanumerics. The quality of an off-the-air broadcast is somewhat better, but in this case the bandwidth has already been limited to 270 or 280 TV lines by the encoding and decoding of the National Television Standard Code (NTSC) color signal.

Since the original projector was introduced, the light output of the color projector has more than doubled through increased lamp power (650 W versus 500 W), the use of low-reflection optical surfaces, and optimized fluid writing conditions. Recently, experiments have shown that light output can be increased several-fold by a further increase in lamp power. Current light output for the single-gun color unit is typically around 300 lm on white or equivalent to 800 open gate lumens as measured for slide and movie projectors. In fact, this value is higher than the output of most 35-mm slide projectors. The black-and-white projector produces about 1000 lm on white or 1800 open gate lumens. The 1000-lm figure would provide 13 fc of incident light on a 7.5- × 10-ft screen or a viewable brightness of over 30 fL for a screen gain of 2.5.

These light-valve projectors have found many applications in the large-screen field due to their light output and inherent registration. The fluid layer has proven to be extremely stable under motion platform accelerations, so that the light-valve projector is serving in the flight simulation field at a number of locations, both with camera and probe pickup and with computer-generated images. Units are fed from a variety of signal sources such as computer-generated graphics, TV camera pickup, and videotape recorders. The screens are typically 5 to 10 ft wide and are usually used in the rear-projection mode. A "wide-screen" color TV demonstration showed the feasibility of using this projection system in a mode that is analogous to Cinemascope or Panavision in the movies. A standard Cinemascope anamorphic lens was added to the projector, which increased the picture width by a factor of 2. The projector was driven from a videotape that had been recorded from a "squeezed"-format Cinemascope movie film. No changes were made in the NTSC signal or the original raster format. The resulting 4- by 11-ft picture appeared quite acceptable in spite of the limited bandwidth of the NTSC-encoded video signal.

7.3.8.2 Liquid Crystal High-Power Displays. Since the advent of LCDs with a significantly large number of elements to make imagery of good quality, a new form of large-image projector has come into prominence. These units (see, for example, Fig. 7.30) are relatively small, light, and inexpensive, making

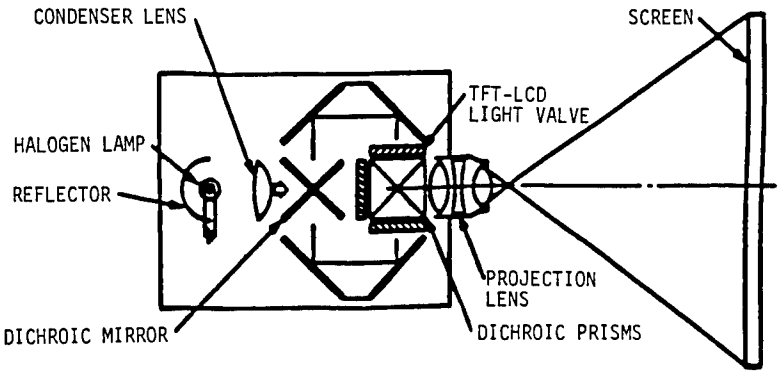


Fig. 7.30 Schematic of generic LCD projector.[101]

them suitable for small meetings and conference rooms. Basically these units are one of three generic designs[102]:

1. A design with three projector lamps, three liquid crystals, three different primary color filters—one for each projection subunit, and three projection lenses set to combine the three colored images on some sort of a screen for either front or back illumination.
2. A somewhat similar design but using a combiner before the projection lens.
3. A design using the Hughes scanned light valve technology.

The following material discusses two of the most recent approaches developed by Hughes. The first involves an addressed liquid crystal technology, and the second involves a scanned light valve technology.

Hughes Highbright™ Color Display. For years pilots have endured the experience of sunlight washing out their displays. Various solutions have been suggested and implemented in an attempt to reduce, if not eliminate, this problem. The situation is most severe in the bubble-like canopy cockpits of fighter aircraft, where the glare of the noonday sun can almost eliminate a display. Attempted solutions involve placing filters over the display screens. These filters include, but are not limited to, contrast enhancement filters, liquid crystal shutters, neutral density filters, and antireflective coatings.

The problem becomes even more severe when one is dealing with full-color displays, since most of the filters attenuate a broad spectrum of visible wavelengths. Even when they are "tuned" to pass the three primary colors, they lose a significant amount of brightness.

The Hughes Aircraft Company has developed a technology that was conceived to solve the sunlight readability problem for color displays. This technology provides the brightness, contrast ratio, sunlight rejection, color saturation, and color stability necessary to overcome sunlight and night vision goggle (NVG) compatibility problems in aircraft of the 1990s. The technology is liquid crystal projection using active-matrix liquid crystal modules.

This technology consists of an illumination source, a light transport system, a tristimulus filter system, a light polarization and modulation system, a lens, and a rear-projection screen. A high-intensity white light from the illumination source (xenon, metal halide, or other source) is coupled to the display unit via a fiber optic cable (Fig. 7.31). This cable plus a lens system defines the light transport system. This white light is sent through special dichroic filters of the tristimulus filter system to achieve the three primary colors of blue, green, and red [the wavelengths are dependent only on the Commission Internationale de l'Éclairage[g] (CIE) coordinates desired]. Each of these primary colors is passed through a light modulation system, which, in this case, consists of transmissive active-matrix liquid crystal modules. Here, each primary light channel becomes linearly polarized and the electronic display information modulates the polarized light (see discussions of liquid crystal display modules and the twisted nematic effect in Secs. 7.3.5 and 7.3.7). The three modulated channels are recombined into a full-color image at the X-prism and projected onto a rear-projection screen through a low-distortion lens.

[g]International Commission on Illumination.

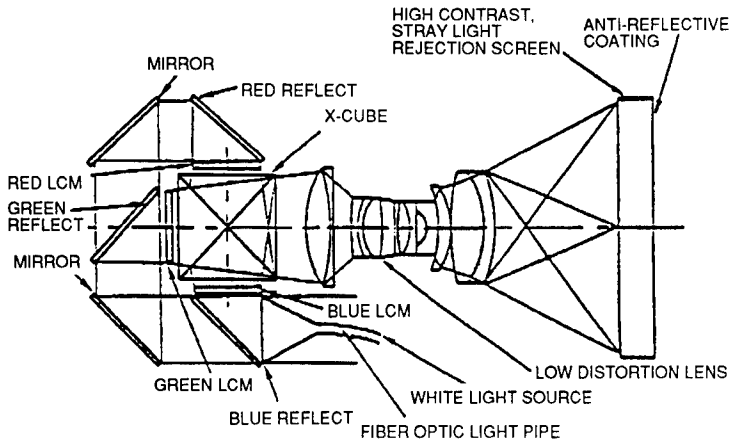


Fig. 7.31 Liquid crystal projection.[103]

The development work in the 1970s, discussed in Sec. 7.3.5, coupled with recent technological advances in optics, materials, liquid crystal modules, and electronics, provides real-time video display capability that allows evolution to the next-generation unit called the Highbright™ Color Display (HCD).

Several technological breakthroughs have been achieved in this display prototype, including the ability to use 10% of the available xenon arc lamp light; the best percentage achieved prior to this was 1%. Other features include a 180-MHz fiber optic data link, an 80% efficient fiber optic light cable, and a 90% efficient polarization system. Nearly perfect color registration is maintained in a rugged, flightworthy design that can withstand repeated landings on aircraft carrier decks.

Figure 7.32 shows a representative HCD aircraft display system. The overall display dimensions are 7 in. wide × 7 in. high × 12 in. long at a weight of 18 lb. This provides a 6- × 6-in. display area with a 5.8- × 5.8-in. image size. The resolution is 512 × 512 pixels from three transmissive polysilicon active-matrix liquid crystal modules developed by Xerox Palo Alto Research Center for Hughes Industrial Products Division. Each pixel on the display screen is a full-color pixel. This is in contrast to other technologies requiring a pixel "group" for a full-color pixel. For aircraft applications display sizes up to 12 × 20 in. are achievable with this approach. The single full-color pixel feature of the HCD is less subject to aliasing than displays in which a pixel "group" is needed to show color. Table 7.11 gives the specifications of the Hughes HCD Model H66D.

To achieve sunlight rejection, the HCD incorporates Hughes' patented circularly polarized screen system, which achieves better than 70% transmission of the available display light through the screen. Randomly polarized light, such as that from the sun, cannot enter this system and is eliminated from being added to the display background and reducing the contrast. Fresnel reflections are held to a minimum by using broadband antireflection coatings.

An alternative to the design configuration shown in Fig. 7.32 is to incorporate the illumination source into the display unit. This approach is useful

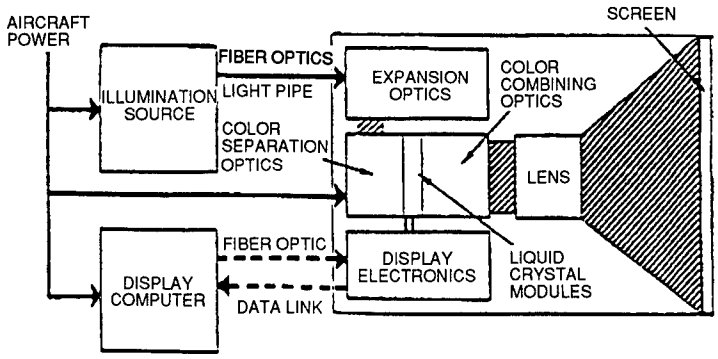


Fig. 7.32 Hughes HCD aircraft display system.[103]

Table 7.11 Hughes HCD Model H66D Specifications (from Ref. 103)

	Daytime Operation (10,000 fL ambient)		Nighttime Operation (0 fL ambient)	
Brightness, fL	10–1000		0.1–10	
Contrast ratio:				
White	35:1		>100:1	
Green	20:1		>100:1	
Red	12:1		>100:1	
Blue	11:1		>100:1	
CIE color coordinates (standard; any coordinates may be chosen):	U′	V′	U′	V′
White	0.209	0.474	0.209	0.474
Green	0.094	0.528	0.094	0.528
Red	0.479	0.528	0.479	0.528[a]
Blue	0.120	0.245	0.120	0.245
Resolution, full-color pixels for a 6- × 6-in. display	88/in.		88/in.	
Distortion	<0.5%		<0.5%	
Positional accuracy	<±0.5%		<±0.5%	
Mean time between failures (h)	>8000		>8000	
Power consumption (W)				
Display unit	25		26	
Illumination unit	210		40	
Interfaces				
Standard	Fiber optic, digital, serial		Fiber optic, digital, serial	
Optional	Digital, byte parallel		Digital, byte parallel	
Screen size (in.)				
Standard	6 × 6		6 × 6	
Optional	5 × 5 to 10 × 10		5 × 5 to 10 × 10	
Exit pupil				
Angle (deg)	0–45 and custom		0–45 and custom	
Dimensions (in.)	4 × 4 to 10 × 10		4 × 4 to 10 × 10	

[a]Compatible with night vision goggles.

for cockpits that can handle an HCD display unit power of 130 W rather than the 25 W from the display unit design using the separate illumination unit.

Hughes Liquid Crystal Light Valve. The Hughes photosensor-addressed liquid crystal light valve (LCLV) is for the generation of large-screen images. Projectors based on this device are in use at both military and civilian command and control centers.

The LCLV uses an input image to generate a replica output image by means of light from an outside source. In many ways it is an optical analog of a field-effect transistor, in that a small amount of light is used to control a large amount of light. It achieves high resolution and good light immunity, allowing display systems of high performance.

A cross section of a typical LCLV is shown in Fig. 7.33. It consists of a series of thin-film layers, sandwiched with a layer of liquid crystal between two glass substrates. These thin-film layers include transparent conductive coatings on either substrate, a photosensor layer, a light-blocking layer, and a dielectric mirror.

In operation an audiofrequency ac voltage is applied between the two transparent electrodes. This voltage is divided by the relative impedances of the various layers; its total is chosen to bias the liquid crystal material just below its electro-optical threshold, when there is no input image. When there is an image input into the LCLV, it generates carriers in the photoconductor and lowers the impedance of that layer. This spatially varying impedance pattern causes a variation in the applied field across the liquid crystal layer, which results in a corresponding variation in its birefringence. This effect is used to generate an image in the light that is incident on the LCLV from the output side, thus replicating the original input image, but with greatly increased luminance.

A simple display system for use with an LCLV is shown schematically in Fig. 7.34. The input image is generated by a CRT. Since both the CRT and the LCLV have fiber optic faceplates, this image is coupled directly into the photosensor, modulating the liquid crystal birefringence as explained above. Typically (for a display application) the output light is generated by a xenon arc lamp. A polarizing beamsplitter is used both to polarize the light from the arc lamp, which illuminates the output side of the LCLV, and to analyze the light reflected from the LCLV. Any light that is rotated by the birefringence pattern in the liquid crystal passes through the beamsplitter and is projected by a projection lens onto the screen.

The LCLV can also be used in applications other than large-screen displays. In particular, a number of groups have used the LCLV for optical data processing. In these applications the arc lamp is replaced by a laser, and the LCLV is used to convert an incoherent image (from a CRT, or relayed from the real world by a lens) into a coherent one. This coherent image is then used by the optical data processing system.

The resolution of the LCLV is limited primarily by the characteristics of the thin films used in its fabrication, and not by any arbitrary mask pattern such as those in active-matrix displays. Its modulation transfer function (MTF) is gradual, like that of a CRT, with a 50% point near 15 to 20 line pairs/mm and with a limiting resolution beyond 30 line pairs. In addition, the combi-

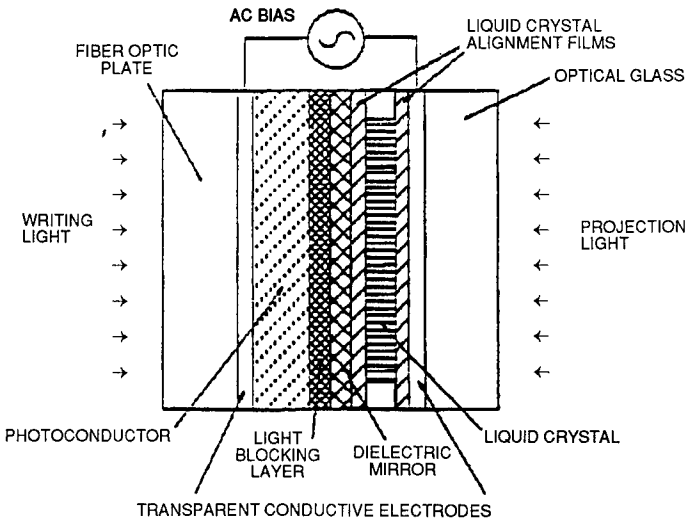


Fig. 7.33 Cross section of Hughes liquid crystal light valve.[103]

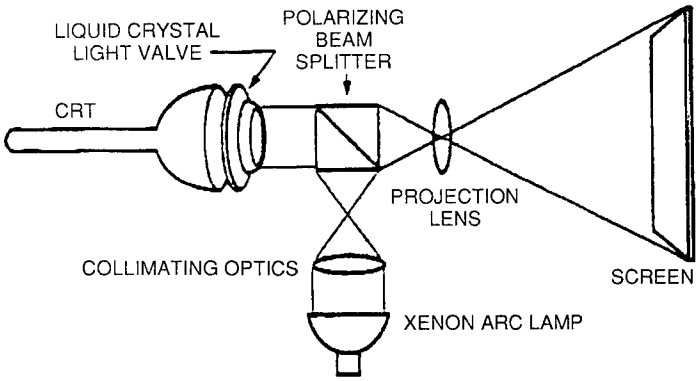


Fig. 7.34 Schematic of LCLV display system.[103]

nation of a light-blocking layer and a dielectric mirror allows output light levels in excess of 2000 lm per light valve to be used without either excessive heating due to absorption or activation of the photosensor by the output light. This combination of properties permits the use of low input light levels (about 75 $\mu W/cm^2$) to generate bright, high-resolution images.

7.4 DISPLAY SPECIFICATION AND CALIBRATION

Several techniques exist for measuring and specifying display performance; most specify "resolution," which can result in widely different resolution numbers for the same device. The cascading of several devices such as a scan

converter tube, a video amplifier, and a CRT in series creates additional complexities in specifying or predicting a total system resolution, especially when the resolution of each individual device is specified differently. A standard for comparing and combining the respective resolution of several devices would thus be useful. Such a resolution standard may be arbitrarily selected but should be meaningful in application to sensor displays and should be capable of convenient and consistent measurement (see Ref. 7).

There are no "official" standards or procedures for defining or measuring display device performance. Although the EIA undertook a study to set up a new military specification on display and display storage tubes, the EIA report stopped after outlining definitions and methods for the storage and reproduction of the electrical signals. It did not address the topic of the optical outputs of such tubes, stating that the task was too difficult.[104] That EIA report is an otherwise useful guide and should be in the hands of every display designer.

A 1989 article by Keller and Zavada[105] lists the many committees in the many organizations attempting to create an acceptable and accepted set of standards for displays. *Different committees often are concerned with standards for a particular technology used in displays and not for displays generally. Though this is not always the case, it is common enough that the reader is cautioned to understand fully whether the standards he or she wishes to use apply to the proposed technology.*

The most frequently used specifications for display device resolution are shrinking raster, limiting television response, and spatial frequency response or MTF. In general, all but the last do not allow one to predict operator performance when viewing a display. Nevertheless, the following definitions are common throughout the industry worldwide and are thus restated here.[10]

7.4.1 Shrinking Raster Resolution

Shrinking raster resolution is determined by writing a raster of equally spaced lines on the display and reducing or "shrinking" the raster line spacing until the lines are just on the verge of blending together to form an indistinguishable blur. An experienced observer normally determines this flat field condition at about 2 to 5% peak-to-peak light intensity variation. Since the energy distribution in a CRT spot is very nearly Gaussian, the flat field response factor occurs at a line spacing of approximately 2σ, where σ is the spot radius at the 60% amplitude of the spot intensity distribution.

7.4.2 Television Resolution (TV Limiting Response)

A television wedge pattern measures spot size by determining the point at which the lines of the wedge are just detectable. The number of TV lines per unit distance is then the number of black and white lines at the point of limiting resolution. The wedge pattern is equivalent to a square-wave modulation function, and therefore the TV resolution is often referred to as the *limiting square-wave response.* (Remember that, in television parlance, one cycle of the square wave produces a black interval and a white interval and is considered to be *two* TV lines.) Assuming a Gaussian spot distribution, the limiting square-wave response occurs at a TV line spacing of 1.18σ. Thus there are approxi-

mately 1.7 times as many limiting TV lines per unit distance as shrinking raster lines for a display with the same spot size.

7.4.3 Modulation Transfer Function

In his sine-wave response technique, Schade analyzes the display resolution by the use of a sine-wave test signal, rather than the square-wave signals employed in a TV test pattern or the photographic bar patterns commonly employed in the optical field. The sine-wave response test produces a curve of response called the modulation transfer function (MTF). This is shown in Fig. 7.35. When several devices are cascaded, such as a scan converter and a CRT, the MTFs of the individual devices are multiplied to provide the total system MTF. This capability for computing the system MTF from individual device MTFs is a major advantage of using the MTF resolution measurement. Another advantage of the MTF technique is the graphic capability it provides for the determination of the visual acuity limit of a given display system. The MTF response can be related to the other resolution measurements (shrinking raster and television) if a Gaussian spot shape is assumed. For example, if a sine-wave test signal were set on the display at a half-cycle spacing corresponding to the shrinking raster resolution line spacing, the resultant observable modulation on the display would be approximately 29%. Table 7.12 can be used to convert from one resolution measurement to another.

7.4.4 Shortcomings of Definitions for Flat Panel Displays

At a sensor display workshop, Gurman[106] pointed out that the Department of Defense faces a difficult challenge in the establishment of control/display requirements and image quality assessment methods since, in some recent decisions, the display media of choice for sensor imagery are optically generated images on a helmet-mounted display and a head-down direct-view liquid crys-

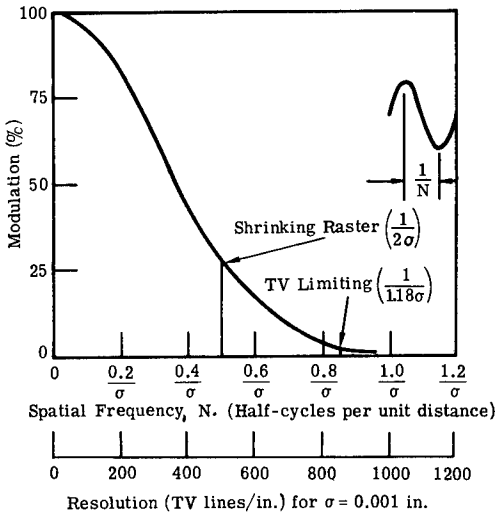


Fig. 7.35 Relative modulation transfer function definition.[10]

Table 7.12 Conversion Table for Various Measures of Display Resolution (from Ref. 10)

To Convert From ↓	To →	TV Limiting	10% MTF	TV_{50}	Shrinking Raster	50% Amplitude	50% MTF	Optical	Equivalent Passband
TV limiting	1.18 σ	—	0.80	0.71	0.59	0.50	0.44	0.42	0.33
10% MTF	1.47 σ	1.25	—	0.88	0.74	0.62	0.55	0.52	0.42
TV_{50} (3 dB)	1.67 σ	1.4	1.14	—	0.84	0.71	0.63	0.59	0.47
Shrinking raster	2.00 σ	1.7	1.36	1.2	—	0.85	0.75	0.71	0.56
50% amplitude	2.35 σ	2.0	1.6	1.4	1.17	—	0.88	0.83	0.66
50% MTF	2.67 σ	2.26	1.8	1.6	1.33	1.14	—	0.94	0.75
Optical (1/*e*)	2.83 σ	2.4	1.9	1.7	1.4	1.2	1.06	—	0.80
Equivalent passband (N_e)	3.54 σ	3.0	2.4	2.1	1.77	1.5	1.33	1.25	—

tal display. This challenge is brought about by the emergence of two technology developments that push the state of the art for both applications and provide little or no database experience in the projected use.

The driving concerns are how we determine the image quality performance requirements and, perhaps more importantly, how we measure the performance.

For example, if a CRT were to be used for the head-down display, we would use MTF as a criterion. To the best of our knowledge, there is no generally accepted and no quantitatively useful metric for measuring flat panel display technology that can be related to MTF.

In the long term, we should conduct a series of laboratory experiments to sample various measurement techniques and establish relational bases for comparison to standard MTF measures. In the short term, approximate procedures could be used. For the measures to be validated, it would be necessary to determine the effects of:

- pixel size, shape, edge characteristic, contrast under both low and high illumination levels, media, rise and fall times, etc.
- motion of targets, symbols, etc.
- color distortion, for both monochrome and color displays, resulting from angle of view, illumination angle, pixel arrangement, etc.

7.4.5 Sensor Resolution

Sensor resolution is usually defined by the 3-dB response, which is equivalent to TV_{50} display resolution and is 1.2 times the shrinking raster solution (Fig. 7.35, Table 7.12).

It is clear that the definitions and measured quantities defined earlier *seem* quantitative and useful. In fact, these are the "specifications" that are used by the more serious and responsible designers and manufacturers. As a matter of fact, Snyder (Ref. 8, Chap. 3) has shown that characteristics discussed above are not time invariant, and Schade (Ref. 8, Chap. 6) has said that properties based on beam half-power points are only valid if the beam cross section is

Gaussian and independent of beam position on the display face, and these criteria have yet to be found to apply in available hardware!

As a first rough cut, the designer may well consider cathode-ray tubes on these factors. But resolution is often stated in terms of spot size (or beam diameter) for some minimal value of beam current, while maximum brightness is specified for a much higher beam current, usually without reference to either beam size or distribution.

A relatively large beam diameter should be chosen to avoid the black unlighted stripes between active lines that cause serious masking and other forms of interference to display observation. One also needs to ensure that such beam size is constant over all the display area or degrades to an *acceptable* degree off axis.

Actually, one would like to have a plot of SNR on the face of the display for a given signal-to-noise input to the display over the entire face of the tube for various SNRs and for various spatial frequencies.

Rosell and Willson[25] conducted their experiments at spatial frequencies sufficiently low that the MTF of the display did not interfere with the experiments. Snyder, on the other hand, had great difficulty in maintaining calibrations over even relatively short experiments even after long warmup times (Ref. 8, Chap. 3).

This handbook cannot list tables of specifications for displays since meaningful specifications have yet to be accepted by the manufacturers and have yet to be used by the commercial display designers.

Detailed methods for measuring spot size as a display criterion are well documented.[107–111] Standards for alphanumerics and other symbols are harder to establish because of the human factors studies required. Shurtleff[112] did such studies. He developed capital letters and numbers (Fig. 7.36) to serve as a reference standard for evaluating the quality of symbols shown on display equipment. Test procedures and criteria were developed to ensure that candidates provide identification performance equivalent to that obtained by Shurtleff using a font designed specifically to minimize intersymbol confusion. For more recent information concerning errors in reading under conditions of transport in various types of vehicles, the reader is referred to Sec. 10.4.5 of Ref. 30 and "Display Tradeoffs" in Ref. 10.

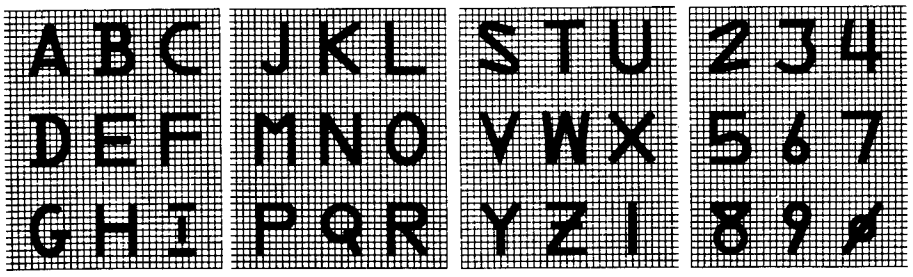


Fig. 7.36 Lincoln/Mitre font.[112]

7.4.6 Display Trade-Offs

Some of the factors influencing the selection of a multisensor display system can be considered quantitatively, but many factors are as yet matters of subjective judgment.[10] Let us compare two sizes (5 and 7 in.) of the Multimode Tonotron® (MMT) to a membrane scan converter tube (SCT) and two sizes of CRTs (8 and 11 in.) for the real-time display of a high-resolution line-scan (strip map) sensor. For this sensor, image storage is required.

The MTF of the four display systems and the SCT alone are plotted in Fig. 7.37. The 7-in. MMT and the 8-in. CRT and SCT have identical MTF curves. The line-scan sensor response alone and the combined sensor and display system response for the four displays are shown in Fig. 7.38.

In Fig. 7.39 the visual acuity of the eye for each of the four image sizes in addition to the display system response is shown. The intersection of the appropriate visual acuity curve and the display curve indicates the maximum usable resolution and is marked with a small circle. Any signal to the left of the intersection has enough display response to be visible to the operators. Conversely, signals to the right of the intersection have insufficient response to be visible. The maximum usable system resolution for each display approach is shown in Table 7.13, which shows that the scan converter plus 11-in. CRT has the best system response. While a significant part of maximum usable resolution results from the total number of resolution elements in the display, the size of the display image is also very important. For example, assume that the number of resolution elements of the 5-in. MMT could be increased so that its total system resolution would be the same as that obtained with the scan converter plus the 11-in. CRT. Under these conditions, it can be seen from Fig. 7.39 that the maximum usable resolution would increase from 505 to only 560 TV lines even though the display system resolution is increased from 550 to 900 TV lines. The small increase in resolution is due to the fact that the MMT display size (3 by 3 in.) is not large enough for the operator to see the finer detail *even though it is present on the tube.*

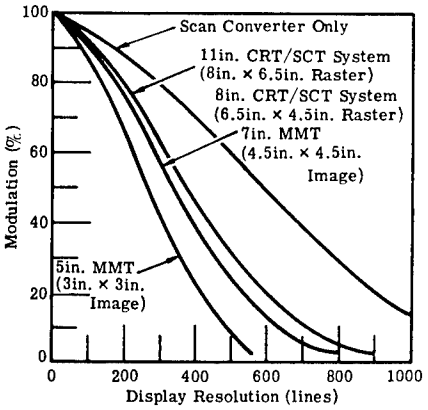


Fig. 7.37 Display system component responses.[10]

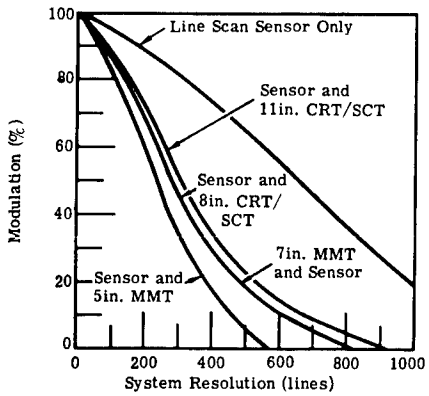


Fig. 7.38 Overall response of sensor plus display system.[10]

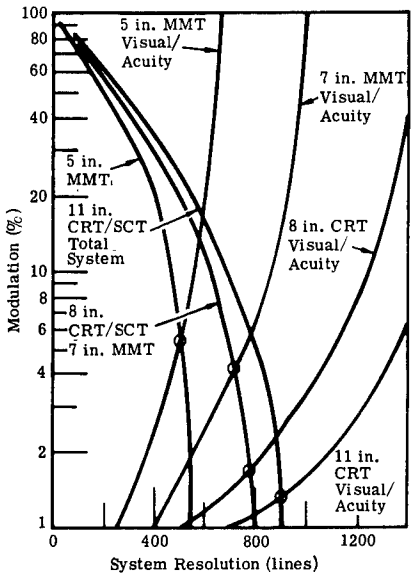


Fig. 7.39 Display system response and visual acuity for four systems.[10]

Table 7.13 Maximum Usable Display System Resolution (from Ref. 10)

	Image Size (in.)	Maximum Usable Resolution (TV lines limiting)
5-in. MMT	3.0 × 3.0	505
7-in. MMT	4.5 × 4.5	725
Scan converter plus 8-in. CRT	6.5 × 4.5	795
Scan converter plus 11-in. CRT	8.0 × 6.5	885

The above reference to the Multimode Tonotron (MMT) is made not because the MMT is the most modern display system, which it is not, but rather because data from the appropriate studies using it are available to illustrate the trade-offs between resolution, scan conversion, display size and, most importantly, *information transfer to the observer.*

7.4.7 Summary of Display Performance

There is no single display that is optimum or even best in all tactical aircraft systems. In laboratories, in factories, and in airport waiting rooms, a wide variety of technology can fill designers' needs. Outdoors in football stadiums or along highways, displays must be very large, and fewer choices exist. But in an aircraft cockpit, where bright, unobscured illumination at 10,000 fc or more can flood a display, few choices exist for good resolution and moderate size. As of June 1990, the double-twisted LCD seemed to the authors to be one of the most obvious choices.

7.5 CAVEATS

Flat panel displays are developing so rapidly that one cannot fairly compare the new liquid crystal, plasma, or electroluminescent displays with the older ones. The new two-dimensionally sampled displays have many advantages, such as a capability to show much more data; a capability to achieve, in the case of LCDs, a much greater contrast that hitherto possible in the presence of high ambient illumination; and an ability to produce relatively faithful color reproduction—better than any other display commercially available at the time this chapter was written.

The new two-dimensionally sampled displays, however, call for a very much more careful look at the needs for image reconstruction and format matching than was required with, say, CRTs. (See the remarks by Tuttle on display interfaces in the latter paragraphs of Sec. 7.3 and in Ref. 51.)

7.5.1 An Opinion

Among the various flat panel displays available in summer 1990, results were most promising for the double-twisted nematic LCD for use in bright sunlight. Though some of the other technologies can and do produce larger displays, their contrast and brightness, as well as their color fidelity, do not yet match the best prototype LCDs. However, for *most* routine good-quality image display purposes, the best and probably the cheapest technological approach for a high-resolution display probably still is the CRT, even with its drawbacks such as its size, its lower contrast in the presence of bright ambient sunlight, and its heat dissipation problems. The CRT's quality and price make it the best choice for most applications in which its two main limitations are not critical.

7.6 DISPLAY DESIGN PROCEDURE

Step 1: Decide Primary Use of Display

Decide:

- What tasks are to be accomplished?
- What is the principal use for the new display?
- What angular resolution is required by the observer?
- What temporal response is required?
- Is interlace acceptable?

Step 2: Determine Operating Conditions

Decide what viewing distance *is* to be used or *can* be used. Remember that in aircraft such factors as ejection seats specify or implicitly require a minimum clearance for knees and feet, thus establishing a minimum eye-to-display distance for a pilot restrained by his flight harness.

- From the considerations of required angular resolution determined above, what linear resolution (under no vibration) does this infer?
- How many pixels need to be provided per observed resolution element?
- How many pixels are required across the entire display?

Step 3: Repeat for Conditions of Expected Vibration

If vibration is to be part of the display environment, repeat the previous step using corrections based on an examination of the Vollmerhausen data, starting with the material in Fig. 7.7 and the associated tables and subsequent figures through Fig. 7.16, or use the corresponding data from Ref. 30.

Step 4: Derive Brightness and Contrast Needed

- Determine what ambient light levels are in the display area and their distribution.
- Determine how much ambient light is potentially reflected by the front surface of the display into the viewing angle of the observer's eyes.
- Refer to Figs. 7.1, 7.2, and 7.3. Use the data they provide or prepare the equivalent for the conditions of your design.
- Specify the brightness and contrast ranges required.

Step 5: Determine Availability of Chosen Parameters

Refer to Ref. 47 or its equivalent and determine if the brightness, contrast, pixel size and number, and display face size are all available in one display device.

Step 6: Review Requirements and Parameters and Reiterate Selection Process

Review the requirements and associated parameters, and redo the selection process as many times as necessary with the understanding that the displays currently available may not be able to present to a human all the data needed for the task.

Acknowledgments

This paper has been assembled from the work of many of the people named in the reference list and bibliography. It has been made possible with the cooperation of several people who worked with us in setting up and documenting the Sensor Display Workshop held in 1989 by the Institute for Defense Analyses (IDA) in conjunction with the U.S. Army CECOM Center for Night Vision and Electro-Optics (CCNVEO), Fort Belvoir, Virginia. In that connection, we especially acknowledge the help of Carolyn Nash and Charles Freeman, of CCNVEO.

These pages reflect the tutoring of the late Otto Schade, supplemented by conversations with Gerry Slocum and Robert Sendall of Hughes Aircraft, who with a few others joined Lucien M. Biberman in giving a course in electro-optics starting in 1971, and supplemented more recently in our meetings with Louis Silverstein and Larry Nelson of Honeywell.

Finally, we must acknowledge the close cooperation of Richard Vollmerhausen of CCNVEO. Vollmerhausen not only contributed in good measure to this collection but also served as a reviewer along with Earl A. Alluisi and Jeffrey A. Nicoll, of IDA; Louis Silverstein, of Honeywell; Mark Tischler, of the U.S. Army Aeroflight Dynamics Directorate; Gerry Slocum, Robert Sendall,

and J. W. Weber of Hughes Aircraft; and James L. Bonomo, of the Rand Corporation.

References

1. M. D. Levine, *Vision in Man and Machine,* McGraw-Hill, New York, 1985.
2. A. B. Watson, A. J. Ahumada, Jr., and J. E. Farrell, "Window of visibility: a psychophysical theory of fidelity in time-sampled visual motion displays," *Journal of the Optical Society of America A* **3**(3), 300–307 (March 1986).
3. C. E. Rash, R. W. Verona, and J. S. Crowley, "Human factors and safety considerations of night vision systems flight using thermal imaging systems," AARL Report 90-10, U.S. Army Aeromedical Research Laboratory (April 1990).
4. L. M. Biberman, "Displays," Chap. 18 in *The Infrared Handbook,* Environmental Research Institute of Michigan, Ann Arbor, MI (Revised 1985).
5. L. M. Biberman, Ed., *Proceedings of the Sensor Display Workshop,* Vols. I–III, IDA Document D-713, Institute for Defense Analyses (December 1989). Volumes I and II contain proprietary information.
6. A. S. Patel, "Spatial resolution by the human system, the effect of mean retinal illuminance," *Journal of the Optical Society of America* **56**(5), 689–694 (May 1966).
7. G. Murch and L. Virgin, "Resolution addressability ratio: how much is enough?" *SID Digest of Technical Papers,* Society for Information Display, pp. 101–103 (1985).
8. L. M. Biberman, Ed., *Perception of Displayed Information,* including Chaps. 3 (H. L. Snyder), 4 (A. Schnitzler), 5 (F. A. Rosell, R. H. Willson), and 6 (Otto Schade, Sr.), Plenum Press, New York (1973).
9. H. L. Task, "An evaluation and comparison of several measures of image quality for television displays," PhD dissertation, Optical Sciences Center, Univ. of Arizona, published by Human Engineering Division, Armstrong Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio (Jan. 1979).
10. Edited and abridged from G. K. Slocum, lectures given at Univ. of Tel Aviv in 1974; derived from G. K. Slocum, "Airborne sensor display requirements and approaches," Report TM-888, Display Systems Dept., Hughes Aerospace Group (Sep. 1967). This material was also presented at an Opto-Electronics Seminar sponsored by AGARD, Paris, France, and Balkesjø, Norway, September 16–20, 1974, published as "Lecture notes" by AGARD.
11. L. D. Silverstein, J. Krantz, F. Gomer, Y.-Y. Yeh, and R. Monty, "Effects of spatial and illumination quantization on the image quality of color matrix displays," *Journal of the Optical Society of America A* **7**(10), 1955–1968 (Oct. 1990).
12. L. D. Silverstein, Honeywell, personal communication (1989).
13. D. F. Kocian, "Design considerations for visually coupled systems and their interface to sensor/computer generated cockpit imagery systems," in *Proceedings of the Sensor Display Workshop,* L. M. Biberman, Ed., Vol. II, pp. 75–194, IDA Document D-713, Institute for Defense Analyses (Dec. 1989).
14. B. H. Tsou, "Visual psychophysical considerations in the design of binocular helmet-mounted displays," in *Proceedings of the Sensor Display Workshop,* L. M. Biberman, Ed., Vol. II, pp. 45–74, IDA Document D-713, Institute for Defense Analyses (Dec. 1989).
15. D. Greene, "Night vision pilotage system field of view/resolution tradeoff study flight experiment report," Report NV-26, U.S. Army CECOM Center for Night Vision and Electro-Optics (Jan. 27, 1988).
16. B. Butterfield, "F-16 helmet-mounted display flight evaluation," *Proceedings of the SPIE* **1290,** 226–235 (April 1990).
17. S. N. Roscoe, "The trouble with HUDs and HMDs," *Human Factors Society Bulletin* **30**(7), 30 (July 1987).
18. S. Hale, "Visual accommodation and virtual images: a review of the issues," Technical Note 3-90, Army Human Engineering Laboratory (Feb. 1990).
19. E. H. Linfoot, *Fourier Methods in Optical Image Evaluation,* Focal Press, London and New York (1964).
20. L. D. Silverstein and R. Merrifield, "Development and evaluation of color display systems, phase I: fundamental visual perception and display considerations," DOT/FAA/PM-85-19,

Honeywell (1985). Available from National Technical Information Service, Springfield, VA 22161.

21. L. D. Silverstein, "Human factors for color displays: concepts, methods and research," in *Color and the Computer,* D. J. Durell, Ed., Academic Press, New York (1987).
22. C. Shannon and W. Weaver, *Mathematical Theory of Communications,* University of Illinois Press, Urbana, IL (1959).
23. F. A. Rosell and R. H. Willson, "Performance synthesis: electro-optical sensors," Report AFAL-TR-72-279, Westinghouse Defense and Electronic Systems (Aug. 1972).
24. F. A. Rosell and R. H. Willson, "Performance synthesis: electro-optical sensors," Report AFAL-TR-73-260, Westinghouse Defense and Electronic Systems (Aug. 1973).
25. F. A. Rosell and R. H. Willson, "Performance synthesis: electro-optical sensors," Report AFAL-TR-74-104, Westinghouse Defense and Electronic Systems (April 1974).
26. F. A. Rosell, R. H. Willson, and H. R. Walmsley, "Effects of vibration and G-loading on airborne E-O sensor augmented observers," Report AFAL-TR-75-172, Westinghouse Defense and Electronic Systems (April 1976).
27. F. A. Rosell and G. Harvey, "The fundamentals of thermal imaging systems," NRL Report 8311, EOTPO Report 46, Naval Research Laboratory (May 10, 1979).
28. W. H. Levison, S. Baron, and A. M. Junker, "Modeling the effects of environmental factors on human control and information processing," AARL Report 76-74, U.S. Army Aeromedical Research Laboratory (Aug. 1976).
29. H. R. Jex, "Problems in modeling man-machine control behavior in biodynamic environments," in *Proceedings of the 7th Annual Conference on Manual Control,* NASA SP-281, National Aeronautics and Space Administration (1971).
30. "Vibration and display perception," Sec. 10.4 in *Engineering Data Compendium: Human Perception and Performance,* K. R. Boff and J. E. Lincoln, Eds., Vol. III, Armstrong Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio (1988).
31. R. Vollmerhausen, U.S. Army CECOM Center for Night Vision and Electro-Optics, personal communication (1990).
32. M. J. Wells and M. J. Griffin, "Benefits of helmet-mounted display image stabilization under whole-body vibration," *Aviation, Space, and Environmental Medicine,* p. 13 (1984).
33. M. J. Wells and M. J. Griffin, "Flight trial of a helmet-mounted display image stabilisation system," *Aviation, Space, and Environmental Medicine,* p. 319 (April 1987).
34. S. Lifshitz, S. J. Merhav, and A. J. Grunwald, G. E. Tucker, and M. Tischler, "Suppression of biodynamic interference in head-tracked teleoperation," presented at European Rotorcraft Forum, Glasgow (Sep. 1990).
35. J. Silk, "The transfer function of a finite sampling process," in *Proceedings of the Sensor Display Workshop,* L. M. Biberman, Ed., Vol. I, pp. 55–63, IDA Document D-713, Institute for Defense Analyses (Dec. 1989).
36. R. Vollmerhausen, "Sampling and Display Processing," in *Proceedings of the Sensor Display Workshop,* L. M. Biberman, Ed., Vol. I, pp. 85–131, IDA Document D-713, Institute for Defense Analyses (Dec. 1989).
37. R. Vollmerhausen, "Display of sampled imagery," in *Proceedings of the IRIS Symposium on Passive Sensors,* Johns Hopkins University Applied Physics Laboratory (May 15, 1990).
38. R. Vollmerhausen, "Application of the sampling theorem to solid state cameras and flat panel displays," Report NV-2-14, U.S. Army CECOM Center for Night Vision and Electro-Optics (Feb. 27, 1989).
39. O. H. Schade, Sr., "A method of measuring the optical sine-wave spatial spectrum of television image display devices," Society of Motion Picture and Television Engineers (Sep. 1959).
40. T. Iki and KI. I. Werner, "CRTs," *Information Display* **5**(12) (Dec. 1989).
41. I. Reingold, "Display devices: a perspective on status and availability," in *Proceedings of the Society for Information Display* **15**(2), pp. 63–73, Society for Information Display, Montvale, NJ (1974).
42. S. Sherr, *Fundamentals of Display System Design,* John Wiley & Sons, New York (1970).
43. G. H. Heilmeier, L. A. Zanoni, and L. A. Barton, "Dynamic scattering: a new electrooptic effect in certain classes of nematic liquid crystals," *Proceedings of the Institute of Electronics Engineers* **56**(7), 1162–1171 (July 1968).
44. M. R. Miller, "Overview of flat panel display device technology," in *Proceedings of the Sensor*

Display Workshop, L. M. Biberman, Ed., Vol. II, pp. 3–7, IDA Document D-713, Institute for Defense Analyses (December 1989).

45. L. A. Jeffries, "Digitally addressed high brightness flat panel display," *SID Digest,* Society for Information Display, Montvale, NJ (1973).
46. L. S. Yaggy and N. J. Koda, *Eighth Technical Session Proceedings,* Society for Information Display, Montvale, NJ (1967).
47. *How to Select a CRT Monitor,* 1990 Designer's Guide Series, and *How to Select a Flat Panel Display,* 1989 Designer's Guide Series, Beta Review, P.O. Box 38, Millwood, VA 22646.
48. L. K. Anderson, "The cathode ray tube display," *Journal of Vacuum Science and Technology* **10**(5), 67 (1973).
49. J. E. Wurtz, "The not-so-amazing survival of the CRT," *Information Display* **5**(9), 5 (Sep. 1989).
50. T. Iki and K. I. Werner, "CRTs," *Information Display* **5**(12) (Dec. 1989).
51. R. Tuttle, "Display Interfaces," in *Proceedings of the Sensor Display Workshop,* L. M. Biberman, Ed., Vol. I, pp. 133–144, IDA Document D-713, Institute for Defense Analyses (December 1989).
52. J. D. Kingsley and G. W. Ludwig, "The efficiency of cathode-ray phosphors, II. correlation with other properties," *Journal of the Electrochemical Society* **117,** 353–359 (1970).
53. G. Garlick, *Luminescence of Inorganic Solids,* P. Goldberg, Ed., Academic Press, New York (1966).
54. S. G. Tomlin, "The back-scattering of electrons from solids," *Proceedings of the Physical Society (London)* **82,** 465–466 (1963).
55. H. A. Bethe, "Theory of the passage of fast corpuscular rays through matter," *Annalen der Physik* **5,** 325–400 (1930).
56. L. V. Spencer, "Theory of electron penetration," *Physical Review* **98,** 1597–1615 (1955).
57. J. H. Jacob, "Penetration and energy deposition of electrons in thick targets," *Journal of Applied Physics* **45**(1), 467–475 (Jan. 1974).
58. C. Feldman, "Range of 1–10 keV electrons in solids," *Physical Review* **117,** 455–459 (1960).
59. W. Ehrenberg and D. L. N. King, *Proceedings of the Physical Society (London)* **81,** 751 (1963).
60. P. H. Hoff and T. E. Everhart, in *10th Symposium on Electron, Ion and Laser Beam Technology,* L. Marton, Ed., San Francisco Press (1969); J. Appl. Phys., Vol 42 (1971).
61. J. D. Kingsley and J. S. Prener, "Voltage dependence of cathode-ray efficiency of phosphors: phosphor particles with nonluminescent coatings," *Journal of Applied Physics* **43**(7), 3073–3079 (1972).
62. P. H. Dowling and J. R. Sewell, *Journal of the Electrochemical Society* **100,** 22 (1953).
63. V. D. Meyer, "Absorption of electron kinetic energy by inorganic phosphors," *Journal of Applied Physics* **41,** 4059–4065 (1970).
64. A. Bril and H. A. Klasens, Philips Research Reports, No. 7 (1952).
65. P. Damon, "The multimode tonotron: a versatile display storage tube," in *Fifth Technical Session Proceedings,* 177–190, Society for Information Display, Montvale, NJ (1965).
66. C. Curtin et al., in *1973 Symposium Digest of Technical Papers,* Society for Information Display, Montvale, NJ (1973).
67. J. W. Sandberg, "Operation of direct view storage tubes," Hughes Aircraft Company, Carlsbad, CA (1972).
68. R. Hayes, R. G. Culter, and K. W. Hawken, "Storage tube with silicon target captures very fast transients," *Electronics,* 97–102 (Aug. 30, 1973).
69. M. J. Cantella, in *Optical Society of America Technical Digest on Optical Displays,* Session WB3 (1975).
70. H. G. Slottow, *SID Digest,* p. 138, Society for Information Display, Montvale, NJ (1975).
71. L. Creagh, "Nematic liquid crystal materials for displays," *Proceedings of the Institute of Electrical and Electronics Engineers* **61**(7), 814–822 (July 1973).
72. G. H. Heilmeier, L. A. Zanoni, and L. A. Barton, "Further studies of the dynamic scattering mode in nematic liquid crystals," *IEEE Transactions on Electron Devices* **ED-17**(1), 22–26 (Jan. 1970).

73. S. Kobayashi, *Conference Record of 1970 Institute of Electrical and Electronics Engineers Conference on Display Devices,* New York, December 2–3, 1970, p. 135 (1971).

74. D. Jones and S. Lu, *SID Digest,* Society for Information Display, Montvale, NJ (1972).

75. L. T. Lipton and N. J. Koda, "Matrix-address liquid crystal panel display," in *SID 1973 Symposium Digest of Technical Papers,* pp. 46–47, Society for Information Display, Montvale, NJ (1973).

76. A. R. Kmetz and F. K. Von Willisen, *Non-emissive Electro-optic Displays,* Plenum Press, New York (1976).

77. L. T. Lipton, Neyer, and Massetti, "A liquid crystal television display using a silicon-on-sapphire switching array," in *SID 1975 International Symposium Digest of Technical Papers,* pp. 78–79, Society for Information Display, Montvale, NJ (1975).

78. M. N. Ernstoff, Aerospace Division, Hughes Aircraft Company, Culver City, California, personal communication.

79. C. J. Nuese, H. Kressel, and I. Ladnay, "Light-emitting diodes and semiconductor material for displays," *Journal of Vacuum Science and Technology* **10,** 772–788 (1973).

80. L. A. Goodman, "The relative merits of LEDs and LCDs," *Proceedings of the Society for Information Display* **16**(1) (1975).

81. E. E. Loebner, "The future of electroluminescent solids in display applications," *Proceedings of the IEEE* **61**(7), 837 (July 1973).

82. C. J. Nuese, J. J. Tietjen, J. J. Gannon, and H. F. Gossenburger, "Optimization of electroluminescent efficiencies for vapor-grown $GaAs_{1-x}P_x$ Diodes," *Journal of the Electrochemical Society* **116,** 248–253 (1968).

83. R. H. Haitz, "Trends in LED display technology," in *Proceedings of the 24th Electronics Components Conference,* Washington, DC, May 13–15, 1974, p. 2, sponsored by the Institute of Electrical and Electronics Engineers.

84. D. A. Laws and R. R. Ady, "Should you use LCD or LED displays?" *Electronic Design* **23,** 88 (1974).

85. L. Cosentino, "On the transient scattering of light by pulsed liquid crystal cells," *IEEE Transactions on Electron Devices* **ED-18,** 1172 (1971).

86. G. Baur, "Angular dependence of transmitted light in deformed twisted nematic cells," in *Conference Record of 1974 Conference on Display Devices and Systems,* p. 139, sponsored by Institute of Electrical and Electronics Engineers and Society for Information Display (1974).

87. A. A. Bergh and P. J. Dean, "Light-emitting diodes," *Proceedings of the Institute of Electrical and Electronics Engineers* **60,** 182 (1972).

88. National Semiconductor Data Sheet for NSN 71 Numeric Display.

89. L. Creagh, A. Kmetz, and R. Reynolds, "Performance characteristics of nematic liquid crystal display devices," *IEEE Transactions on Electron Devices* **ED-18,** 672 (1971).

90. A. Sussman, "Electro-optic liquid crystal devices: principles and applications," *IEEE Transactions on Parts, Hybrids, and Packaging* **PHP-8,** 28 (1972).

91. A. R. Kmetz, "Liquid crystal displays in perspective," *IEEE Transactions on Electron Devices* **ED-20,** 954 (1973).

92. P. M. Alt and P. Pleshko, "Scanning limitations of liquid crystal displays," *IEEE Transactions on Electron Devices* **ED-21,** 146 (1974).

93. C. R. Stein and R. A. Kashnow, "A two frequency coincidence addressing scheme for nematic liquid crystal displays," *Applied Physics Letters* **19,** 343 (1971).

94. P. J. Wild and J. Nehring, "An improved matrix addressed liquid crystal display," *Applied Physics Letters* **19,** 335 (1971).

95. H. K. Bucher, R. E. Klingbeil, and J. P. Van Meter, "Frequency-addressed liquid crystal field effect," *Applied Physics Letters* **25,** 186 (1974).

96. H. Takata, O. Kogure, and K. Murase, "Matrix-addressed liquid crystal display," *IEEE Transactions on Electron Devices* **ED-20,** 990 (1973).

97. K. Nakada, T. Ishibashi, and K. Toriyama, "A design of multiplexing liquid crystal display for calculators," in *Conference Record of 1974 Conference on Display Devices and Systems,* p. 139, sponsored by Institute of Electrical and Electronics Engineers and Society for Information Display (1974).

98. W. E. Good, "Recent advances in the single-gun color television light-valve projector," *Proceedings of the SPIE* **59**, 96–99 (1975).

99. T. T. True, "Color television light valve projection system," presented at IEEE International Convention, Session 26 (1973).

100. T. T. True and W. C. Bates, General Electric, personal communication (1976).

101. J. Coonrod, "Liquid crystal displays," in *Proceedings of the Sensor Display Workshop,* L. M. Biberman, Ed., Vol. II, p. 28, IDA Document D-713, Institute for Defense Analyses (December 1989).

102. R. Flasck, *Flat Panel Display News* **1**(1) (March 1990). Available from Panelight Display Systems, 2440 16th Street, #110, San Francisco, CA, 94103.

103. J. A. Fowler, Hughes Aircraft Company, personal communication (1990).

104. N. Ehrlich et al., "Review of MIL-E-1 test methods for cathode ray storage tubes," Electronic Industries Association, Washington, D.C. (March 1972).

105. P. A. Keller and R. Zavada, "A survey of display standards activities," *Information Display* **5**(12), 21–26 (Dec. 1989).

106. B. Gurman, "Army/LHX concerns," in *Proceedings of the Sensor Display Workshop,* L. M. Biberman, Ed., Vol. I, p. 53, IDA Document D-713, Institute for Defense Analyses (December 1989).

107. J. E. Bryden, "Some notes on measuring performance of phosphors used in CRT displays," in *Proceedings of the 7th National Symposium on Information Display,* Society for Information Display, Montvale, NJ.

108. J. M. Constantine, "Two-slit spot analyzer," in *Proceedings of the 7th National Symposium on Information Display,* Society for Information Display, Montvale, NJ.

109. R. J. Doyle, F. P. Heiman, and M. Kerman, "Modulation transfer function of electrical output cathode ray storage tubes," *SID Journal* **1**(4), 20–22.

110. E. M. Sawtelle and G. W. Gonyou, "Dynamic CRT spot measurement techniques," in *Proceedings of the National Symposium of the Society for Information Display,* Society for Information Display, Montvale, NJ (May 1968).

111. L. E. White, *Measuring Spot Size in High-Resolution Cathode-Ray Tubes,* Sutton Publishing Company (1959).

112. D. Shurtleff, *How to Make Displays Legible,* Human Interface Design, La Mirada, CA (1980).

Bibliography, Classified by Subject

General

How to Select a CRT Monitor, 1990 Designer's Guide Series, and *How to Select a Flat Panel Display,* 1989 Designer's Guide Series, Beta Review, Inc., P.O. Box 38, Millwood, VA 22646.

Credelle, T., "Avionic liquid crystal displays—recent trends," SAE Technical Paper, Aerospace 871-790, Society of Automotive Engineers (Oct. 1987).

Greeson, J. C. "International standards challenge flat-panel displays," *Information Display* **6**(7/8) (July/Aug. 1990).

Information Display **5**(12) (Dec. 1989). This issue contains brief reviews of CRT displays, liquid crystal displays, electroluminescent displays, plasma displays, and display standards. A reproduction of Botticelli's painting, "Birth of Venus," on a Matsushita 17-inch-diagonal color display may well change many minds about the present state of the art.)

Tannas, L., Ed., *Flat Panel Displays and CRTs,* Van Nostrand Rheinhold, New York (1985).

Calibration

Bryden, J. E., "Some notes on measuring performance of phosphors used in CRT displays," in *Proceedings of 7th National Symposium on Information Display,* Society for Information Display, Montvale, NJ.

Constantine, J. M., "Two-slit spot analyzer," in *Proceedings of 7th National Symposium on Information Display,* Society for Information Display, Montvale, NJ.

Doyle, R. J., F. P. Heiman, and M. Kerman, "Modulation transfer function of electrical output cathode ray storage tubes," *SID Journal* **1**(4), 22–22.

Ehrlich, N., et al., "Review of MIL-E-1 test methods for cathode ray charge storage tubes," Electronic Industries Association, Washington, DC (March 1972).

Sawtelle, E. M., and G. W. Gonyou, "Dynamic CRT spot measurement techniques," in *Proceedings of the National Symposium of the Society for Information Display,* Society for Information Display, Montvale, NJ (May 1968).

White, L. E., *Measuring Spot Size in High-Resolution Cathode-Ray Tubes,* Sutton Publishing Company (1959).

Cathode-Ray Tubes and Phosphors

Kingsley, J. D., and J. S. Prener, "Voltage dependence of cathode-ray efficiency of phosphors: phosphor particles with nonluminescent coatings," *Journal of Applied Physics* **43**(7), 3073–3079 (1972).

Meyer, V. D., "Absorption of electron kinetic energy by inorganic phosphors," *Journal of Applied Physics* **41,** 4059 (1970).

Spencer, L. V., "Theory of Electron Penetration," *Physical Review* **98**(6), 1597–1615 (June 15, 1955).

Electroluminescent Displays

Garlick, G., *Luminescence of Inorganic Solids,* P. Goldberg, Ed., Academic Press, New York (1966).

Loebner, E. E., "The future of electroluminescent solids in display applications," *Proceedings of the IEEE* **61**(7), 837–861 (July 1973).

Nuese, C. J., J. J. Tietjen, J. J. Gannon, and H. F. Gossenberger, "Optimization of electroluminescent efficiencies for vapor-grown $GaAs_{1-x}P_x$ diodes," *Journal of the Electrochemical Society* **116** (1968).

HUDs and Disorientation

Hale, S., "Visual accommodation and virtual images: a review of the issues," Technical Note 3-90, Army Human Engineering Laboratory (Feb. 1990).

Iavecchia, J. H., S. N. Roscoe, and H. P. Iavecchia, "Eye accommodation to head-up virtual images," *Human Factors* **30**(6), 689–702 (Dec. 1988).

Marsh, J. S., and L. A. Temme, "Optical factors in judgments of size through an aperture," *Human Factors* **32**(1), 109–118 (Feb. 1990).

Moffitt, K., "Ocular responses to monocular and binocular helmet mounted display configurations," *Proceedings of the SPIE* **1116,** 142–148 (1989).

Newman, R. L., "Response to Roscoe, 'The trouble with HUDs and HMDs'," *Human Factors Society Bulletin* **30**(10), 3 (Oct. 1987).

"Pilot disorientation examined in Canadian CF-18 review," *Aviation Week & Space Technology,* p. 20 (June 25, 1990).

Roscoe, S. N., "The trouble with HUDs and HMDs," *Human Factors Society Bulletin* **30**(7), 30 (July 1987).

Roscoe, S. N., "The trouble with HUDs and HMDs revisited," *Human Factors Society Bulletin* **30**(11), 3 (Nov. 1987).

Weintraub, D. J., "HUDs, HMDs, and common sense: polishing virtual images," *Human Factors Society Bulletin* **30**(10), 1 (Oct. 1987).

Light-Emitting Diodes

Bergh, A. A., and P. J. Dean, "Light-emitting diodes," *Proceedings of the Institute of Electrical and Electronics Engineers* **60,** 182 (1972).

Craford, M. G., and D. L. Keume, "LED technology," *Solid State Technology* (Jan. 1974).

Craford, M. G., and W. O. Groves, "Vapor phase epitaxial materials for LED applications," *Proceedings of the Institute of Electrical and Electronics Engineers* **61** (1973).

Haitz, R. H., "Trends in LED display technology," in *Proceedings of the 24th Electronics Components Conference,* Washington, DC, May 13–15, 1974, sponsored by Institute of Electrical and

Electronics Engineers and Electronics Industries Association (1974).

Nuese, C. J., H. Kressel, and I. Ladnay, "Light-emitting diodes and semiconductor material for displays," *Journal of Vacuum Science and Technology* **10** (1973).

Nuese, C. J., H. Kressel, and I. Ladnay, "The future for LEDs," *IEEE Spectrum,* pp. 28–38 (May 1972).

Liquid Crystal Display Applications

Alt, P. M., and P. Pleshko, "Scanning limitations of liquid crystal displays," *IEEE Transactions on Electron Devices* **ED-21,** 146 (1974).

Creagh, L., A. Kmetz, and R. Reynolds, "Performance characteristics of nematic liquid crystal display devices," *IEEE Transactions on Electron Devices* **ED-18,** 672 (1971).

Ernstoff, M. N., et al., "Liquid crystal pictorial display," in *IEEE 1973 International Electron Devices Meeting Conference Record,* pp. 548–551, IEEE, New York (1973).

Ernstoff, M. N., "Liquid crystal pictorial display," presented at SID 1975 Technical Meeting, Society for Information Display, Montvale, NJ.

Kmetz, A. R., "Liquid crystal displays in perspective," *IEEE Transactions on Electron Devices* **ED-20,** 954 (1973).

Kmetz, A. R., and F. K. Von Willisen, *Non-emissive Electro-optic Displays,* Plenum Press, New York (1976).

Laws, D. A., and R. R. Ady, "Should you use LCD or LED displays?" *Electronics Design* **23** (1974).

Lipton, L. T., et al., "A liquid crystal television display using a silicon-on-sapphire switching array," in *SID 1975 International Symposium Digest of Technical Papers,* pp. 78–79, Society for Information Display, Montvale, NJ (1975).

Lipton, L. T., M. A. Meyer, H. G. Dill, and D. O. Massetti, "SOS liquid crystal TV display," paper presented at 1974 International Electron Devices Meeting, Washington, DC, December 9, 1974.

Sussman, A., "Electron-optic liquid crystal devices: principles and applications," *IEEE Transactions on Parts, Hybrids, and Packaging* **PHP-8,** 28 (1972).

Liquid Crystal Materials

Baur, G., "Angular dependence of transmitted light in deformed twisted nematic cells," in *Conference Record of 1974 Conference on Display Devices and Systems,* p. 139, sponsored by Institute of Electrical and Electronics Engineers and Society for Information Display (1974).

Bucher, H. K., R. T. Klingbiel, and J. P. VanMeter," Frequency-addressed liquid crystal field effect," *Applied Physics Letters* **25** (1974).

Meyer, M. A., L. T. Lipton, G. H. Hershman, and P. G. Hilton, "Processing of a monolithic SOS array for liquid crystal display applications," presented at 1975 Electrochemical Society Meeting, Toronto, Canada, May 1975.

Schadt, M., and W. Helfrich, "Voltage dependent optical activity of a twisted nematic liquid crystal," *Applied Physics Letters* **18** (1971).

Schiekel, M. F., and K. Fahrenschon, "Deformation of nematic liquid crystals with vertical orientation in electric fields," *Applied Physics Letters* **19**(10), 391–393 (Nov. 15, 1971).

Liquid Crystal Technology

Cosentino, L., "On the transient scattering of light by pulsed liquid crystal cells," *IEEE Transactions on Electron Devices* **ED-18,** 1172 (1971).

Goodman, L. A., "Liquid crystal displays—electro-optical effects and addressing techniques," *RCA Review* **35** (1974). The March, September, and December issues of Vol. 35 of *RCA Review* contain a series of papers on the physics, chemistry, and application of liquid crystals.

Observer: Eye, Performance, and Performance Measures

Biberman, L. M., and S. Nudelman, *Photoelectronic Imaging Devices,* Vols. 1 and 2, Plenum Press, New York (1971).

Blackwell, H. R., "Contrast thresholds of the human eye," *Journal of the Optical Society of America* **36**(11), 624–643 (1946).

Blackwell, H. R., "Specification of interior illumination levels," *Illumination Engineering* (June 1959).

Ferrell, R. J., and J. M. Booth, "Design handbook for imagery interpretation equipment," Report D-180-19063-1, Boeing Aerospace (Feb. 1984).

Linfoot, E. H., *Fourier Methods in Optical Image Evaluation,* Focal Press, London and New York (1964).

McLean, W., and S. Smith, "Developing a wide field of view head mounted display for simulators," *Proceedings of the SPIE* **778,** 79–82 (1987).

Schade, O. H., "A method of measuring the optical sine-wave spatial spectrum of television image display devices," Society of Motion Picture and Television Engineers (Sep. 1958).

Silverstein, L. D., "Human factors for color displays: concepts, methods, and research," in *Color and the Computer,* D. J. Durell, Ed., Academic Press, New York (1987).

Silverstein, L. D., J. S. Lepkowski, et al., "Modeling of display color parameters and algorithmic color selection," *Proceedings of the SPIE* **624,** 26–35 (1986).

Silverstein, L. D., and R. Merrifield, "Development and evaluation of color display systems, phase I: fundamental visual perception and display considerations," DOT/FAA/PM-85-19, Honeywell (1985). Available from National Technical Information Service, Springfield, VA 22161.

Silverstein, L. D., R. W. Monty, J. W. Huff, and K. L. Frost, "Image quality and visual simulation of color matrix displays," presented at Society of Automotive Engineers Aerospace Technology Conference and Exposition, Long Beach, CA, October 5–8, 1987.

Slocum, G. K., lectures given at Univ. of Tel Aviv in 1974; derived from G. K. Slocum, "Airborne sensor display requirements and approaches," Report TM-888, Display Systems Dept., Hughes Aerospace Group (Sep. 1967). This material was also presented at an Opto-Electronics Seminar sponsored by AGARD, Paris, France, and Balkesjø, Norway, September 16–20, 1974, published as "Lecture notes" by AGARD.

Plasma Technology for Displays

Bitzer, D. L., and H. G. Slottow, "Principles and applications of the plasma display panel," in *Proceedings of the OAR Research Applications Conference,* Office of Aerospace Research, Arlington, VA, March 1968; see also *Proceedings of the 1968 Microelectronics Symposium,* Institute of Electrical and Electronics Engineers, St. Louis, MO (1968).

Bitzer, D. L., and H. G. Slottow, "The plasma display panel—a digitally addressable display with inherent memory," in *Proceedings of the Fall Joint Computer Conference,* San Francisco, CA, November 1966.

Gregory, R., M. S. Bishop, and R. Weil, "Electron beam addressed plasma display panel," *Proceedings of the Institute of Electrical and Electronics Engineers* (May 1969).

Petty, W. D., "Multiple states and variable intensity in the plasma display panel," Report R-497, Coordinated Science Laboratory, University of Illinois, Urbana, IL (Nov. 1970).

Slottow, H. G., *SID Digest,* p. 138, Society for Information Display, Montvale, NJ (1975).

Slottow, H. G., "The plasma display panel—principles and prospects," in *Proceedings of 1970 Institute for Electrical and Electronics Engineers Conference on Display Devices* (1970).

Stredde, E., "The development of a multi-color plasma display panel," Report R-370, Coordinated Science Laboratory, University of Illinois, Urbana, IL (Nov. 1967).

Projection Displays

Biberman, L. M., Ed., *Perception of Displayed Information,* including Chapters 3 (H. L. Snyder), 4 (A. Schitzler), 5 (F. A. Rosell, R. H. Willson), and 6 (Otto Schade, Sr.), Plenum Press, New York (1973).

Glenn, W. E., "Principles of simultaneous color projection using fluid deformation," *SMPTE Journal* **79** (Sep. 1970).

Good, W. E., "Recent advances in the single-gun color television light-valve projector," *Proceedings of the SPIE* **59,** 96–99 (1975).

True, T. T., "Color television light valve projection systems," presented at IEEE International Convention Session 26 (1973).

van Raalte, J. A., "Survey of developmental light valve systems," presented at IEEE International Convention Session 26 (1973).

Scan Converters

Hayes, R., R. G. Culter, and K. W. Hawken, *Electronics* (Aug. 30, 1973).

Sandberg, J. W., "Operation of scan converter tubes," Hughes Aircraft Company, Carlsbad, CA (1974).

Yaggy, L. S., and N. J. Koda, in *Eighth Technical Session Proceedings,* Society for Information Display, Montvale, NJ (1967).

Storage Devices, Generally Related Material

Kazan, B., and M. Knoll, *Electronic Image Storage,* Academic Press, New York (1968).

Sandberg, J. W., "Operation of direct view storage tubes," Hughes Aircraft Company, Carlsbad, CA (1972).

CHAPTER 8

Photographic Film

H. Lou Gibson
Consultant in Biomedical Photography

CONTENTS

8.1 INTRODUCTION

Infrared film provides a means for registering IR images with essentially the same equipment and procedures of ordinary photography. Unlike many methods for thermal or long-wave IR recording, the film does not have to be shielded from regular, reasonably low, ambient temperatures during handling.

Nevertheless, as with the usual equipment or film, it is necessary to avoid exposing loaded cameras to direct sunlight or to hot enclosures such as the glove compartment of a car.

The symbols, nomenclature, and units used in this chapter are listed in Table 8.1.

8.2 STORAGE

Even temperatures as low as that from the heat of the body can cause fogging after long periods. Therefore, keeping loaded and unloaded film in the refrigerator is necessary for short-term storage. For long-term storage, the sealed package of film should be kept in a freezer. Storage temperatures between 4°C (40°F) and −20°C (−10°F) are recommended. The films are packed in manufacturing conditions of about 50% relative humidity.

8.3 SPECTRAL SENSITIVITY

Infrared films are sensitized by dyes to 900 nm (see Fig. 8.1). Extending this range is not generally practical because the radiation from objects at ambient temperatures will cause fogging. However, for special applications and cooling procedures, Kodak®[a] spectroscopic plate type 1-Z goes to almost 1200 nm.

Since the film is also sensitive to visible radiation, filters over lenses or lights are used to absorb light but pass IR so that the useful recording range is confined to 700 to 900 nm. They are discussed later.

The 700- to 900-nm region is valuable for recording the IR reflected by foliage and other aspects of ecological surveys and for use in the military, remote sensing, biomedical, spectroscopic, forensic, documentary, and many other fields.[2–4]

Objects that have been heated to at least 250°C (480°F) emit radiation in the recording range of IR film. A hot flat iron is a good source of this actinic radiation. [Around 500°C (930°F) "red-hot" visible light is produced.]

Infrared photography is useful for investigating hot metal ingots, cylinder heads, exhaust manifolds, welding operations, and cutting tools in action.

8.4 HANDLING AND PROCESSING

Since the IR film is susceptible to light, operations such as spooling, loading, and developing must be performed in the dark and not too close to a space heater, and certainly not within view of red-hot filaments. No suitable safelight exists for the darkroom, but luminous clock dials are often useful.

[a]The following terms used in this chapter are trademarks of the Eastman Kodak Company: Aerochrome, Aerographic, Estar, Eastman, Kodak, Wratten, D-19, and D-76. Particular text and illustrations reprinted courtesy Eastman Kodak Company.

Table 8.1 Symbols, Nomenclature, and Units

Symbols	*Nomenclature*	*Units*
C	Proportionality constant	–
CI	Contrast index; slope of a straight line drawn between two points on the H-D curve that usually represent the highest and the lowest useful densities in a continuous-tone, black-and-white negative.	–
D	Photographic density*; $= \log_{10} 0 = \log_{10} (\tau_D^{-1})$	–
E_v	Illuminance; photometric irradiance, the incident luminous flux per unit area	lm
$E_{v,\lambda}(\lambda)$	Spectral illuminance; $= \partial E_v / \partial \lambda$	lm nm^{-1}
E_e	Irradiance; the incident radiant flux per unit area, $\partial \phi / \partial A$	W m^{-2}
$E_{e,\lambda}(\lambda)$	Spectral irradiance; $= \partial E_e / \partial \lambda$	W m^{-2} nm^{-1}
H	Exposure; the time integral of E_v or E_e; $= \int_0^t E_e(t)\, dt$	J m^{-2}
H_v	Photometric exposure; $= K_m \int_0^t \int_{\lambda_1}^{\lambda_2} V(\lambda) E_{e,\lambda}(\lambda)\, d\lambda\, dt$	m cd sec lm sec m^{-2}
H_e	Radiometric exposure; $= \int_0^t \int_{\lambda_1}^{\lambda_2} E_{e,\lambda}(\lambda)\, d\lambda\, dt$	J m^{-2}
K_m	Maximum luminous efficacy; = 683	lm W^{-1}
O	Opacity*; $= \tau_D^{-1}$	–
S	Sensitivity	cm^2 erg^{-1}
$S(\lambda)$	Spectral sensitivity; $= [H_e(\lambda)]^{-1}$	cm^2 erg^{-1}
t	Time	sec
$V(\lambda)$	Relative spectral luminous efficiency	–
$\Delta\lambda$	Spectral bandwidth	nm, or μm
γ	Slope of the straight-line portion of the H-D curve	–
λ	Wavelength	μm, or nm
$\bar{\lambda}$	Center wavelength	μm, or nm
ρ	Reflectance	–
$\bar{\rho}$	Reflectance, average	–
τ	Transmittance	–
$\bar{\tau}$	Transmittance, average	–
τ_D	Transmittance, diffuse; the fraction of incident radiation	–
$\tau(\lambda)$	Transmittance, spectral	–
Subscripts		
e	Radiometric quantities	
v	Photometric quantities	
λ	The operation of partial differentiation with respect to wavelength	

*All data in this chapter are based upon diffuse densities (diffuse transmittance).

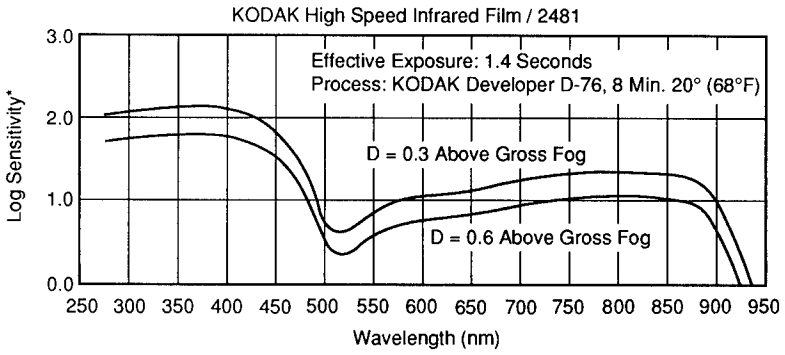


Fig. 8.1 The above curves were derived using Kodak developer D-19 for 8 min at 68°F. The log sensitivity is the reciprocal of exposure (erg/cm^2) required to produce specified density above density of base plus fog.[1]

For large quantities of film, processing machines are available. Information on chemical solutions and running times is provided by the manufacturers.

Certain plastics and woods sometimes used for ordinary photography, which do not pass visible radiation, can transmit IR. Some early cameras, fittings, and dark slides should be checked. Metal dark slides are now universally available.

8.5 TECHNIQUES

In addition to knowing the film and its capabilities and understanding the basics of handling it, each application demands a varying degree of specific technical knowledge and IR procedures. Optics, exposure, sensitometry, color modifications, and densitometry are areas requiring particular attention that depend on the precision required in the results.

Each specialization calls for its own degree of study. The literature of the disciplines covers the IR factors when they are involved. The fundamentals are now discussed.

8.6 FOCUS

The geometric and physical optical properties have been well established by Hardy and Perrin,[5] but are somewhat different for IR images, with the usual lenses designed primarily for visible radiation. Infrared rays have a longer wavelength than those from the visible region. The lens focal positions indicated by barrel markings or range finders are not accurate with IR film. In a single-lens reflex (SLR) camera, the visible image on the screen is not a reliable indicator of focus either. The longer wavelength rays are refracted less by lenses so that the effective focal length of the lens is about 0.5% longer than the visible-light focal length. In addition, aberration corrections made to optimize the sharpness of the image in the visible portion of the spectrum are not generally optimum for the IR portion.

The correction increases the lens-to-film distance—in effect, for a visible image, the focus would be that for a closer object. For moderate close-ups with an SLR, a visual focus on a detail somewhat closer than the main IR detail should be made.

With many complex modern lenses, the focal correction is not so simple to gauge. Nieuwenhuis[6] describes a method for calibrating a lens for exacting work. Modern aerial cameras have complex mechanisms for focusing the invisible image and have lenses corrected for IR.

8.7 EXPOSURE

Ordinary visible-light film-speed ratings cannot be applied to IR materials because of their sensitivity to the unseen region. (Visible light is barred by the filter used.) Again, the usual photographic exposure meters do not usefully register IR.

For medium-distance photography at ground level and for close-ups of scientific and technical subjects, a flat lighting is usually best. In that way, the shadows obscure the least detail. A hazy light, with scarcely noticeable shadows, is better than full sunlight for revealing a disease pattern on leaves. With full cloud cover, however, exposures are difficult to predict, and the use of a tripod will probably be necessary. Dark cloud conditions also will affect the color balance obtained in IR color photography (described later). Since characteristic colors are usually important in this technique, one should take advantage of hazy sunlight or direct sun. The photographer should be with his back to the sun, making sure that a shadow does not fall across the subject.

For plants photographed at 3 ft or closer, the photographer should be particularly careful to find a viewpoint that presents the least amount of obscuring shadows. Flat lighting is usually best; at other times a camera direction at 90 deg to the sun's rays may disclose the best clarity of unshadowed detail. Haze may seem to produce a "lifeless" record, but the purpose of the photograph is to reveal information and this is often best accomplished when shadows are completely absent.

Meter readings for ground photography can only be used as a rough guide; the exposures they indicate should be interpreted by experience. Exact speed recommendations are not possible because the ratio of IR to visible radiation is variable and because photoelectric meters are calibrated only for visible radiation. Use a hand-held meter rather than a through-the-lens type. For critical applications, make trial exposures and determine your own meter calibration. Use the exposure index given in the film data sheet as a starting point.

Table 8.2 will serve as a rough guide for Kodak high-speed IR film and Kodak Ektachrome® IR film. Of course, it is not likely that IR photographs will have to be made when it is raining. Haze, however, does offer a softer lighting than direct sunlight, without shadows. Using a white bed sheet between the sun and specimen will diffuse direct sunlight for close-ups, but this method requires an exposure increase of four times for black-and-white IR photography, or eight times for IR color photography.

For precise work, sensitometric methods for evaluating the effects of, or the requirements for, precise exposure can be adopted. Essentially the same pro-

Table 8.2 Approximate Ground-Level Outdoor Exposure Under Varied Conditions for Kodak High-Speed IR Film with the No. 29 Filter and Kodak Ektachrome IR Film with the No. 12 Filter

Sky Condition	Suggested Trial Exposure*			
	Black-and-White		Color	
Direct Sunlight	1/60	f/16	1/125	f/16
Haze, Shadows Not Quite Discernible	1/125	f/16	1/125	f/11
Light Cloud Cover, Location of Sun Just Discernible	1/25	f/11	1/60	f/11†
Moderate Rain	1/10	f/8	1/30	f/6.3‡
Heavy Thundershower	1/8	f/6.3	1/15	f/5.6‡

*Infrared Intensity of the Sky Varies Hourly and Seasonally

†With KODAK Color Compensating Filter CC10M, to Offset Blueness

‡With KODAK Color Compensating Filter CC20M

cedures used in certain disciplines for using film as an expedient for measuring radiation intensities can be applied to IR investigations. It is not the scope of this chapter to go into detail on such specialized techniques. Some information on the fundamentals of the densitometry involved are discussed later.

8.8 AERIAL PHOTOGRAPHY

While IR aerial photography is primarily useful in enhancing the contrast of the terrain, other distinct advantages exist. For example, bodies of water are rendered very dark in sharp contrast to land, assuming that the day is clear. Fields and wooded areas are rendered very light. Coniferous and deciduous growth is differentiated, the former appearing darker than the latter. Cities are rendered darker than fields. For this reason, in IR pictures taken at very high altitudes, urban areas appear as dark patches surrounded by lighter countryside. Applications of IR aerial photography in agriculture, archaeology, ecology, forestry, geology, and hydrology are numerous.

A high-speed black-and-white film, Kodak Aerographic® infrared film 2424 (Estar® base) is available especially for use in aerial cameras. This negative material, sensitive to IR radiation as well as to the blue light of the visible spectrum, has exceptional sensitivity and is capable of giving high contrast. The diaphragm setting on the camera lens depends on the prevailing light conditions, on the degree of development, and to a great extent on the terrain. In determining this setting, experience is the best guide. Green countryside requires less exposure than an industrial area.

On a bright day a common setting based on a solar altitude of 40 deg and an altitude of 10,000 ft is 1/400 s at *f*/16 with Kodak Aerographic infrared film 2424 (Estar base) and the Kodak Wratten® gelatin filter No. 25-red. These typical camera exposures are based on processing to yield an effective aerial film speed of 400.

The Kodak Wratten filter No. 89B-IR is also utilized for aerial photography. Its use in place of the No. 25 filter may result in a slightly greater reduction of haze effects without an undue increase in exposure.

Kodak Aerographic infrared film 2424 (Estar base) is available in the 70 mm width, in various lengths, and wound on various spools. A 3 ft length of 2424 film can be hand spooled and utilized in conventional cameras. Using Kodak high-speed infrared film in 135-size magazines with the same exposure as that recommended for 2424 film serves the same purpose. This provides a means for making preliminary tests to determine the results that may be expected from a full-scale project of IR aerial photography.

For aerial photography with conventional cameras, the lens should be set at a focus setting of 50 to 80 ft so that IR radiation at infinity distance will be recorded in sharp focus. See Ref. 7 for further data.

8.9 DENSITY AND EXPOSURE

The feature of an IR film that is of paramount importance in all applications is the density generated by the quantity of radiation it receives. The effects of exposure are studied sensitometrically. The characteristics of a representative film, Kodak high-speed IR film, will serve to describe this general response by IR films.

The exposure and development of photographic film produces an image consisting of areas having different transmittances, depending on the number and size of the silver grains present. Specular transmittance may be defined as the ratio of the amount of the undeviated light passing through the plate or film to that of the incident collimated light on the back. Diffuse transmittance is the ratio of the intensity of the undeviated and scattered light together with that of the incident collimated light. The opacity O is defined as the reciprocal of the transmittance. The density D is defined at $\log_{10}O$; it can be either specular or diffuse density. The data in this chapter represent diffuse density values.

The exposure H may be expressed as either the time integral of the illuminance E_v or the time integral of irradiance E_e, so that

$$H_v = K_m \int_0^t \int_0^\infty V(\lambda) E_e(\lambda)\, d\lambda\, dt \tag{8.1}$$

and

$$H_e = \int_0^t \int_0^\infty E_e(\lambda)\, d\lambda\, dt\ , \tag{8.2}$$

where $V(\lambda)$ is the relative spectral luminous efficiency for the CIE-standard photometric observer and K_m is the maximum luminous efficacy.

When exposure is expressed as a photometric quantity in the photographic literature, the units of H_v are customarily meter-candle-seconds. When exposure is expressed as a radiometric quantity, the units of H_e should be joules per square meter.

8.10 SENSITOMETRIC CHARACTERISTICS

The curves presenting density D as a function of $\log_{10}H$ are known as the H and D curves (after Hurter and Driffield), or simply the characteristic curves (see Fig. 8.2). The standard source of radiation for obtaining these curves is an artificial source made to provide a spectral distribution irradiance close to that provided by a 6000 K blackbody in the restricted spectral interval of interest. This source is intended to simulate average daylight, i.e., sunlight plus skylight. From the illuminance from this source, one may derive a unique irradiance in the IR spectral region. Although more direct use of H_e for characteristic curves of IR film makes good sense, H_v has been used in the literature and is a valid indicator of the magnitude of H_e. When a filter is specified along with the characteristic curves, the effects of the filter losses are implicit in the characteristic curves, i.e., the filter is considered to be an integral part of the film.

The slope of the straight-line portion (gamma) of the H and D curve is one parameter used to select appropriate development times. Another parameter that may be used instead of gamma is the contrast index CI, which also is a slope derived from the D versus $\log_{10}H_v$ plots. Contrast index is defined as the slope of a line between the two points on the H and D curve, which represent the highest and lowest useful densities for a continuous-tone black-and-white negative. The CI is used more widely in pictorial photography than in scientific and industrial work.

The characteristic curves are average properties. In addition, these characteristics hold only for the specified conditions of exposure and processing.

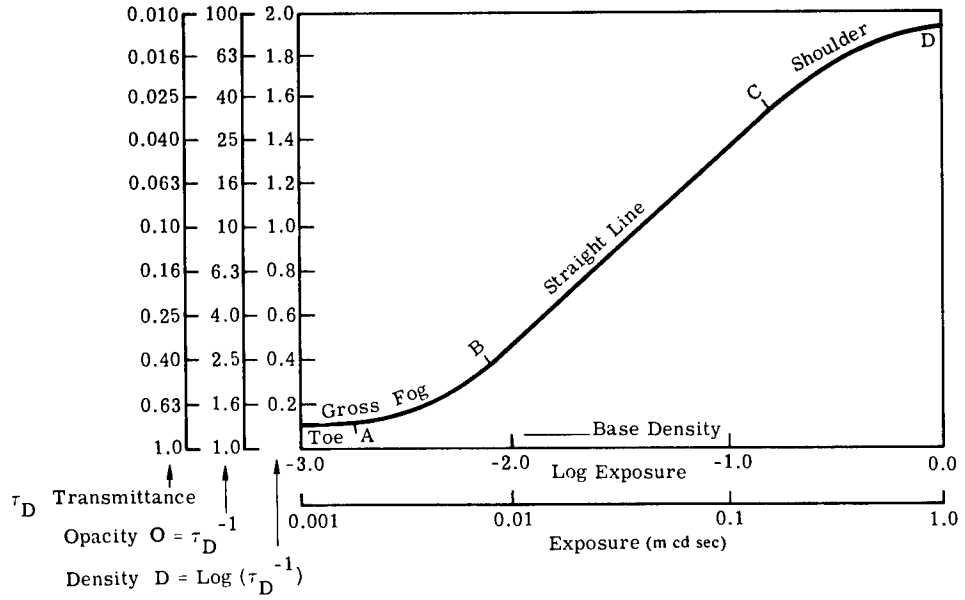


Fig. 8.2 Characteristic curve representative of a negative photographic material.[3]

Changes in light quality or in processing will yield a different set of characteristic curves. The response of photographic materials to parameter variations tends to be nonlinear with almost every parameter.

Apparatus used for making photometric or radiometric measurements must be calibrated for photographic response. Extreme care must be exercised to obtain reproducible processing conditions. When possible, calibrating exposures should be made adjacent to the areas to be measured. A wealth of information concerning the use and problems of photographic materials for photometry may be found in the astronomy literature and that of pertinent disciplines.

The characteristic shape and density levels (for a given exposure) of the H and D curves vary with development time as a dominant parameter. Times are set by the speed and other features of the machine utilized for obtaining the contrast and density required. Representative curves are shown in Fig. 8.3.

Other parameters that govern the shape of the curves are temperature, chemicals, and exposing illuminants.

Manufacturers of films, chemicals, and processing machines can supply related data from which a procedure can be worked out to suit specific appli-

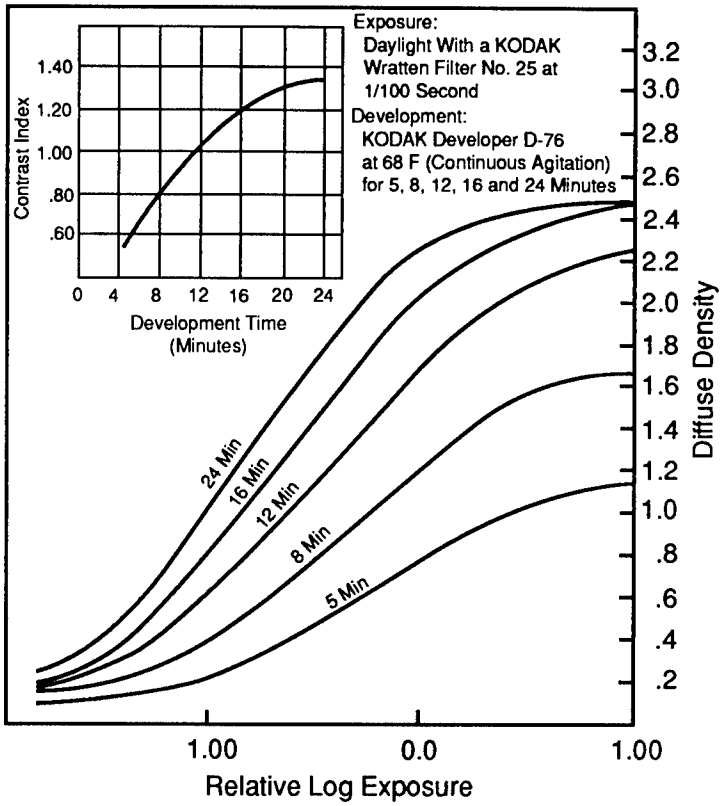


Fig. 8.3 Curves for Kodak high-speed IR film 2481 (Estar base) and Kodak high-speed IR film.[1]

cations. The Eastman® Kodak Company also publishes detailed information regarding specialized applications with spectroscopic materials.

8.11 HYPERSENSITIZING

The speeds of some IR films and plates can be increased by hypersensitizing them. For best results, this should be done just before exposure. Kodak spectroscopic films and plates with *N* sensitizings may be hypersensitized a few days before use. However, type Z plates must be used immediately after treatment. Hypersensitizing has many pitfalls. The characteristics of the emulsion layers have not been established to precisely respond to any modification by the user. An exception is Kodak spectroscopic plate and film, type I-Z; hypersensitizing must be done just prior to exposure. More details can be obtained from the Eastman Kodak Company.

8.12 RECIPROCITY

The reciprocity law states that the product of a photochemical reaction depends on the total energy employed. In photography, this means that the effective exposure *(E)* should equal illuminance *(I)* multiplied by exposure time *(T)*. Through a broad range of I and T, $E = IT$ is a constant that holds true, and any given E value will produce the same image density assuming invariable processing conditions. However, when I is very small and T is larger, or when I is extremely high and T is very small, the formula breaks down. This departure from the rule of constant exposure for a fixed density level is called reciprocity failure (see Fig. 8.4).

Although the reciprocity effect is insignificant in most ordinary applications of photography, some conditions require abnormally long or abnormally short exposures. In reciprocity curve graphics, strict adherence to the reciprocity

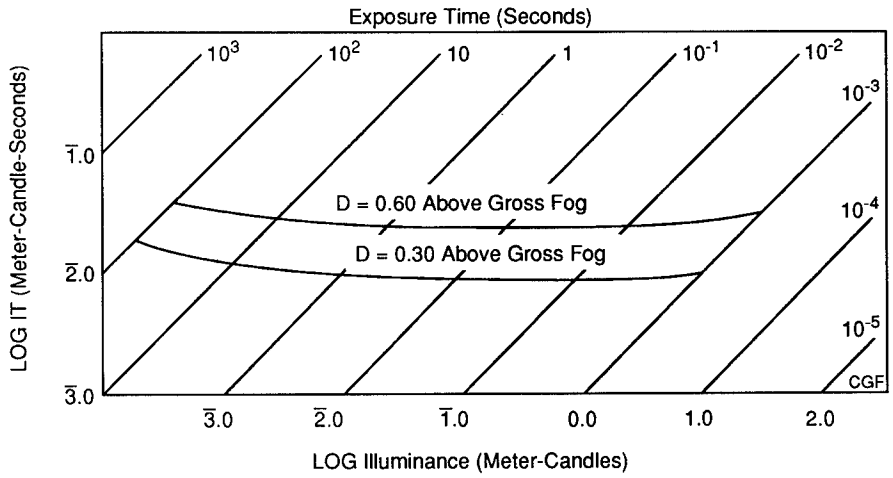


Fig. 8.4 Reciprocity curves for Kodak high-speed IR film 2481 (Estar base) and Kodak high-speed IR film 4143 (Estar thick base).[3]

Table 8.3 Exposure Adjustments for Kodak High-Speed Film

If Indicated Exposure Time (Seconds) Is	Multiply Either Exposure Time or Total Exposure by This Factor
1/1000	1.25
1/100	1.00
1/10	1.00
1	1.00
10	1.00
100	1.60

law would be represented by straight horizontal lines. Upward bending at both ends of the actual curves, however, illustrates the need for increased exposure under abnormal conditions of time or illuminance. Failure to allow for the departure from a straight-line characteristic will result in a reduced density and, in some cases, reduced contrast. In other words, the effective speed of a film varies with the exposure conditions and reaches a maximum at a particular level of illumination. This illumination level varies from one product to another.

Exposure correction factors may be determined from the reciprocity curves or they may be tabulated. Table 8.3 shows the exposure adjustments for the Kodak high speed infrared film, based on processing in a Kodak developer D-76 for 8 min at 20°C (68°F) to a density of 0.60 above gross fog.

8.13 EFFECTIVE SPECTRAL BAND OF FILM-FILTER COMBINATIONS

A photographic camera may sometimes be used as a quantitative radiometric device or to distinguish objects on the ground by film density levels in the photographic image if the approximate spectral-reflectance curves of the objects are known. In either case, a film-filter combination is selected for the particular purpose and a calculation is required to determine the effective spectral band of that choice. The calculation is based on the spectral-additivity assumption. The spectral-additivity curves (Fig. 8.5) represent the relative effectiveness of monochromatic flux at different wavelengths in causing a particular density. Since more flux at some wavelengths is required to produce a certain density than flux at other wavelengths, one assumes that the process is merely less efficient for the former wavelengths. When levels of flux in two different wavelength bands are incident on the film at the same time, it is assumed that the total exposure will be determined by the efficiency-weighted sum of the exposures in each wavelength band.

The accuracy of the additivity assumption is best when the gamma of the sensitometric curve of the film is the same for the bands. It becomes worse as the variation in gamma becomes greater. Though film may respond in a variety of ways, the spectral-additivity assumption has been found accurate for practical purposes. Over relatively wide spectral bands (when the additivity assumption is likely to result in reduced accuracy) the inherent inaccuracy of broadband radiometric measurements is also increased; in narrow spectral

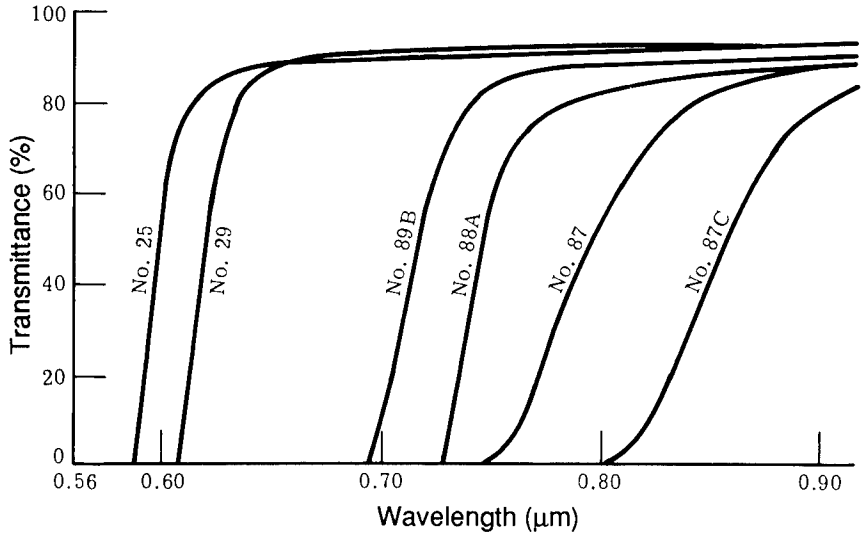


Fig. 8.5 Transmittance of Kodak Wratten filters used for IR photography.[8]

bands, both the validity of the additivity assumption and the radiometric accuracy will improve.

The relation used to calculate the effective exposure H or the efficiency-weighted sum of the exposures for each wavelength is

$$H = C \int_0^{\infty} \int_0^{t} \tau(\lambda) S(\lambda) \rho(\lambda) E_\lambda(\lambda)\, d\lambda\, dt \ , \tag{8.3}$$

where

C = a proportionality constant
$\tau(\lambda)$ = the filter transmittance
$S(\lambda)$ = the relative spectral sensitivity of the film
$\rho(\lambda)$ = the spectral reflectance of the object to be photographed
$E_\lambda(\lambda)$ = the spectral irradiance caused by the source of illumination.

If the value of $\rho(\lambda)$ is not known, one assumes that it is sufficiently constant over the band of operation to be replaced by its average value over that band $\bar{\rho}$. Thus,

$$H = C\bar{\rho} \int_0^{\infty} \int_0^{t} \tau(\lambda) S(\lambda) E_\lambda(\lambda)\, d\lambda\, dt \ . \tag{8.4}$$

The effective spectral bandwidth of operation $\Delta\lambda$ can be found by normalizing the remaining spectrally varying quantity, i.e., $[\tau(\lambda)S(\lambda)E_\lambda(\lambda)]$, to its peak value.

$$\Delta\lambda = \frac{\left[\int_0^{\infty} \tau(\lambda) S(\lambda) E_\lambda(\lambda)\, d\lambda\right]}{[\tau(\lambda) S(\lambda) E_\lambda(\lambda)]_{\text{peak}}} \ . \tag{8.5}$$

The center wavelength of the band is then found from

$$\bar{\lambda} = \frac{\left[\int_0^\infty \tau(\lambda)S(\lambda)E_\lambda(\lambda)\lambda\, d\lambda\right]}{\Delta\lambda[\tau(\lambda)S(\lambda)E_\lambda(\lambda)]_{\text{peak}}} . \tag{8.6}$$

The photographic exposure H should then be proportional to the average reflectance of the object in the effective band from $\bar{\lambda} - \Delta\lambda/2$ to $\bar{\lambda} + \Delta\lambda/2$.

If maximum density separation of two objects with known spectral reflectances is desired, the value of H will depend on $\rho(\lambda)$ for the objects and $\tau(\lambda)$ for the filter transmittance.

The best separation is obtained when the relationship [H(object 1)/H(object 2)] is farthest from a one-to-one ratio. Equation (8.1) is used repeatedly with different filters to compare the effective pairs of exposures for the two objects.

The spectral irradiance $E_\lambda(\lambda)$ is often caused by sunlight. Although the exact spectral distribution of $E_\lambda(\lambda)$ may not be known for a given photographic measurement, usually it will not greatly differ from day to day. Therefore, any reasonable curve of $E_\lambda(\lambda)$ for a clear day may be used. Further, if the spectral band of operation is narrow, the value of $E_\lambda(\lambda)$ may not change significantly with a wavelength in that range. In the latter case, normalization may proceed as if E_λ were constant. The result is identical to Eq. (8.5) in that

$$\Delta\lambda = \frac{\bar{E}_\lambda \int_0^\infty \tau(\lambda)S(\lambda)\, d\lambda}{\{[\tau(\lambda)S(\lambda)]_{\text{peak}}\, \bar{E}_\lambda\}} , \tag{8.7}$$

where $\bar{E}_\lambda$ is the average E_λ in the band $\Delta\lambda$. The value of $\bar{E}_\lambda$ cancels; it does not affect the value of $\Delta\lambda$.

The band center is found as before in Eq. (8.6).

8.14 MODULATION TRANSFER

A realistic means for evaluating the expected definition is the modulation transfer (MT) function. It measures the blurring effects of light scatter in the emulsion and the edge effect of processing.[2] The emulsion function can be multiplied by similar functions obtained from lens and motion effects to determine the final efficiency of the entire system. Fourier mathematics are used.[9] Niederpruem, Nelson, and Yule[10] deal with factors involved in attaining image contrast.

The MT function of an emulsion is measured by exposing a pattern of bars of diminishing widths and spacings having a lateral sinusoidal variation in illuminance—not like the solid strokes of resolution charts. The gradations are spaced by diminishing separations. Their spatial frequency is specified in cycles per millimeter. But the apparent widths of the finest recorded modulations would correspond quite closely to the finest details that could be separated in an image of subject details. The greater the separation of details in the image, the better the transfer. The microdensitometer tracing in Fig. 8.6

shows the variation in MT. The curve in Fig. 8.7 plots a typical function. A value for resolving power cannot be related numerically to MT frequencies. The difference between the two types of data can be appreciated when it is realized that resolving power might be indicated by the single dot in Fig. 8.7, showing 70 lines/mm as the capability under optimum conditions, yet saying nothing about the increased clarity of the coarser details.

An understanding of MT is further aided by analyzing these figures from another viewpoint. Each sinusoidal modulation element in the test pattern has a maximum and minimum illuminance level that is transferred to the film. Were the emulsion a perfect receiver, and were it developed to a sensitometric reproduction of 1:1, the same sinusoidal luminance would be transmitted by the illuminated negative—but displaced by half a wavelength. However, scattering and blurring in the emulsion reduce the effective modulation

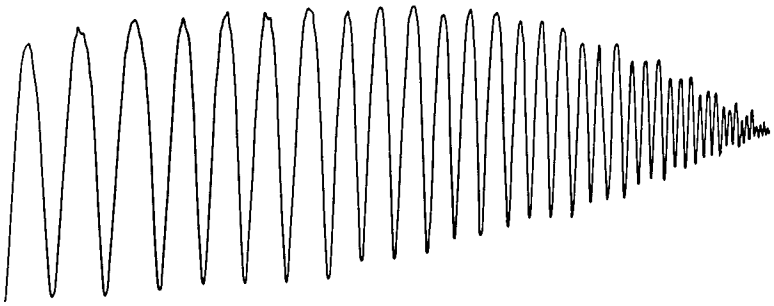


Fig. 8.6 Microdensitometer tracing made from a negative record of a MT transfer pattern.[2]

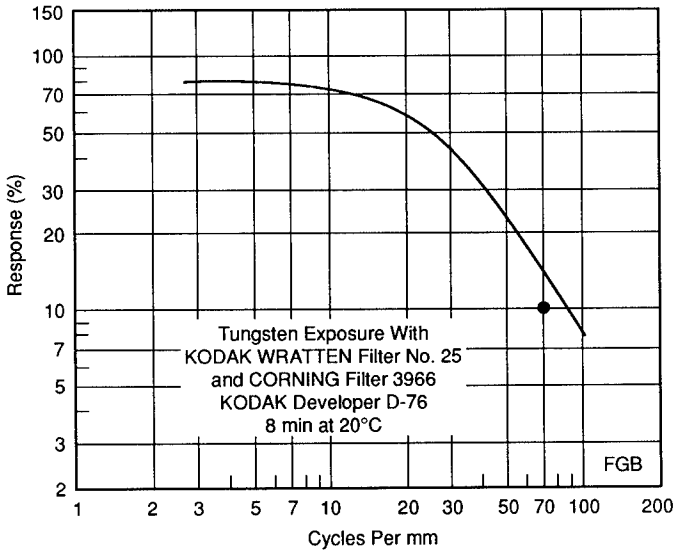


Fig. 8.7 Modulation transfer curve for Kodak high-speed IR film 4143 (Estar thick base).[4]

amplitude by adding illuminance in the valleys and subtracting it from the peaks. The higher the wave frequency (the closer the spacing of the peaks), the more pronounced the deterioration. This is depicted by the microdensitometer traces and indicated by the efficiency of the recordings at the spacings. Since the blurring tends to lighten the dark valleys and darken the light peaks, when the pattern of bars is recorded by a film, the intensities of the transferred maximum and minimum illuminances tend to merge, which reduces the contrast that provides resolution of the bars. The efficiency shown by the MT curve is derived from the expression:

$$\mathrm{MT} = \frac{E_{\max} - E_{\min}}{E_{\max} + E_{\min}} . \tag{8.8}$$

More detailed information and help in working with MT (sine-wave response) can be found in technical papers by Lamberts,[11,12] Nelson,[13] and Perrin.[14] Nelson[13] describes a method for predicting the densities of fine detail in photographic images that makes use of the light-spread function for a lens-and-film combination and a chemical-spread function related to the diffusion of reaction by-products during photographic development. Additional information, including an extensive bibliography, appears in the *SPSE Handbook of Photographic Science and Engineering.*

8.15 DENSITOMETRY

Photographic film generally responds to exposure in a nonlinear fashion. The quantitative result of exposure may be measured by the transmittance of the resulting processed transparency. The commonly derived expression for this transmittance is the density $D = \log_{10}(1/\tau)$. The density of a processed transparency depends in part on the method used to measure transmittance. Consequently, different densitometers may not produce identical density readings of the same transparency. However, any densitometer that produces self-consistent density readings is entirely suitable for use in radiometric photography. The film must be calibrated by exposure to a sensitometric step wedge, which provides the relationship between exposure and final density readings. As long as the functional relationship between exposure and density is single valued, a direct relationship can be established between output signal, input exposure, or target radiance for unknown targets.

8.16 RADIOMETRICS

Using the film and its exposure for the measurement of illuminance and its effects calls for meticulous procedures.[2] Representative applications are studying the nature of distant subjects by evaluating their reflectance, establishing a signal-to-noise ratio, detecting weak signals, and other scientific studies.

8.16.1 Black-and-White Film

The densitometer and the photographic camera with an associated filter and single-emulsion film can be combined to make a useful and economical radio-

metric system provided that certain features of the camera and film are taken into account. These features are those that alter the single-valued relationships between density and scene radiance.

The primary, camera-related variations are off-axis vignetting (or fall-off), flare, and in the case of cameras with focal-plane shutters, shutter-speed variations across the film plane. Because of these variations, exposure varies on the film itself. A photograph of an infinite, perfectly uniform, diffuse (i.e., Lambertian) target will result in nonuniform density across the processed film. Such variations may be taken into account by making one of the exposures of a photographic sequence an exposure of a reasonable uniform diffuse scene. A sheet of matte white typing paper black-illuminated by daylight and placed directly over the camera lens or filter will produce the effect of a uniform Lambertian illuminator or diffuse scene in the photographic spectral range. The relative variation in exposure as a function of film position may be found using the calibrated step-wedge exposure data along with the photograph of the Lambertian scene.

Primary film-related variations are the result of the limiting spatial resolution of the film in combination with the camera and the chemical processing. High-contrast edges tend to develop preferentially. As a result, regions appear (near the edges) with higher densities than would be predicted by a density step-wedge calibration curve. Also, the density of detail near the limit of resolution approaches the density corresponding to the average scene radiance of that detail and the surrounding area, rather than to the radiance of the detail. This is a result of the size of the grains and the scattering of light in the film causing a reduction in the contrast MT function. A bar-pattern resolution of 50 lp/mm marks the point at which the averaging is nearly complete. Thus, the centers of regions no smaller than 0.5 to 1.0 mm should be used to obtain an accurate density-to-scene-radiance correspondence for a particular object.

To obtain the average scene radiance of a fairly large area, one must measure the scene radiance of each part separately by using a densitometer on the photograph. The use of a large densitometer spot will usually lead to an incorrect scene-radiance average unless the variation in density within the spot is much less than 0.1. The transmittance measurement using the large spot size is equivalent to an average of film transmittance, which is nonlinearly related to scene radiance. The most accurate method of achieving the average scene radiance of objects that contain much high-contrast texture, such as forests and agricultural fields, is to take the photograph in such a way that the texture falls below the resolving power of the camera and film system. A small-format camera, hand-held and set at a small aperture with a long exposure time, can produce ideal images of textured areas suitable for radiometric photography.

8.16.2 Color Film

All of the features of single-layer film relating to radiometric densitometry apply also to trilayer color film. Unfortunately, additional peculiarities of trilayer films exist that make the use of such film for radiometric purposes not only difficult, but also less accurate.

In the color-film layers, density is produced by the deposition of dye colorants (instead of grains of metallic silver) in the finished image. Dye colorants are chosen that will produce visual color from the white light used for the image illuminant by means of the tristimulus color-subtractive system. Generally, the spectral absorption coefficient of the dye colorant in one layer is intended to control only one visual band—red, green, or blue—in accordance with the dye concentration within that layer. However, the spectral absorption coefficients are significantly nonzero in the other visual bands that are controlled by the dye colorants, as shown in Fig. 8.8 for Kodak Aerochrome® infrared films 2443 and 3443 (see Sec. 8.18).

Densitometry of trilayer film requires the use of color-separation filters for transmittance measurements. A red filter, for instance, is used over the densitometer light source to obtain the density in the visual tristimulus red band of the image. Because the film layers are stacked together, the products of the transmittances of all three layers must be measured at the same time. The spectral overlap of the colorant spectral absorption coefficients permits the exposures in all three layers to control, to some extent, the density in one visual tristimulus band. The image densities taken with color-separation filters are called *integral densities*.

The spectral absorption coefficients of the dye colorants may be used to compute the densities one would have found for the isolated layers. These computed densities are called *analytic densities* and are computed by using appropriate linear combinations of the three integral densities. If there were no further difficulties, trilayer film could be used for radiometric purposes in the same way as single-layer film.

However, interimage effects appear in trilayer film beyond those of the overlapping spectral absorption coefficients. A fundamental asymmetry to the film processing exists. Processing chemicals must enter the emulsion stack by diffusion from only one side. Exhausted chemicals and by-products must leave by the same route. The speed of chemical reactions depends on the concentration of the constituents and, for diffusion to occur, a concentration gradient must exist through the trilayer emulsion. Therefore, the extent of the reaction in one layer will depend on its location and the extent of the reaction in adjacent

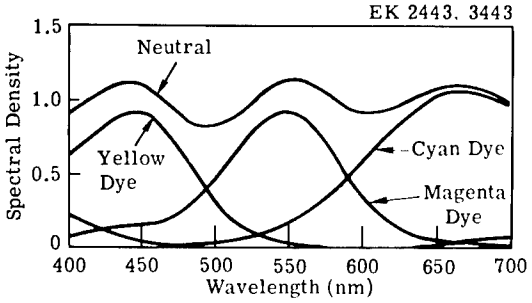


Fig. 8.8 Spectral dye density curves of Kodak Aerochrome IR film 2443 (Estar base) and Kodak Aerochrome IR film 3443 (Estar thin base).[7]

layers. Image processing will differ depending on the spectral balance and magnitude of the exposure. The consequences are that sensitometric curves derived from a gray step-wedge source will not apply accurately to nongray sources. The internal chemical processing is not always the same.

8.17 INFRARED LUMINESCENCE

The fluorescence of objects irradiated with UV is well known. Light of other colors can also excite fluorescence—always at a longer wavelength than the excitation. Infrared fluorescence can also be induced but cannot be seen, hence, it was unsuspected. Dhéré and Biermacher[15] reported a faint presence under UV in geranium leaves during spectroscopic investigations. However, other workers found it too faint to photograph. Gibson[2] discovered that blue-green light could excite a recordable response from a host of materials and worked out a way to photograph it. Figure 8.9 compares the components for ordinary IR photography and those for recording IR luminescence (a term adopted to avoid confusion with visible fluorescence). Corning blue-green molded glass filters C.S. No. H.R.1-59 were used over the lights.

The technique proved valuable in many fields.[8] Biomedical investigations in medicine and in natural history disciplines provided a host of subjects. The detection of forgeries in forensic, art, legal, and other examinations was often made possible when other means failed. Determination of the nature and

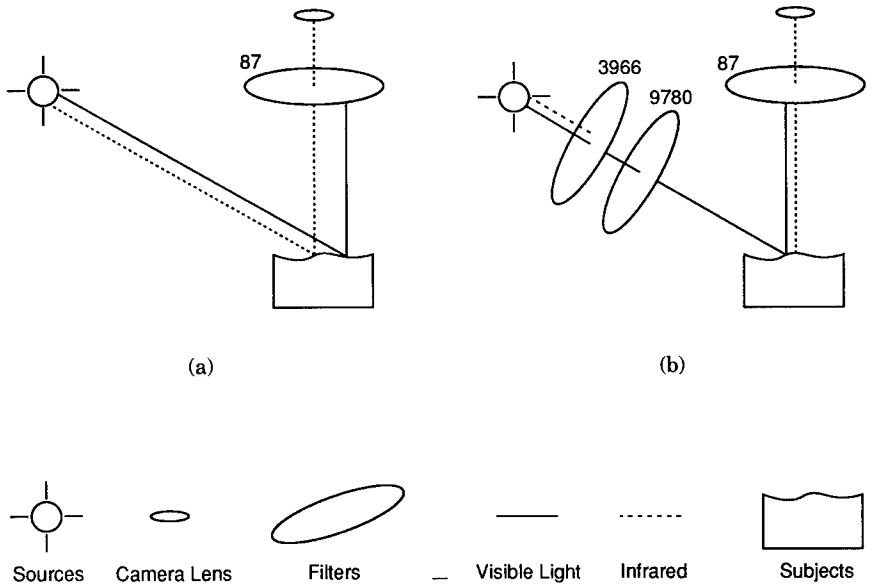


Fig. 8.9 Diagram illustrating the basic difference in setups for (a) IR-reflection and (b) IR-luminescence photography.[8]

condition of wood resins and the differentiation of coal samples was aided. Unsuspected details in fossils and other geological specimens were revealed.

For photomacrographic subjects, spotlights with small beams or the focusing lamps used in photomicrography are useful for lighting small areas. However, they must be enclosed or baffled to prevent stray light from escaping into the room. The blue-green filters should be fitted tightly enough to eliminate light leaks.

Infrared luminescence can also be photographed in a photomicrographic setup. The excitation filters are placed in the light patch between the lamp and the substage condenser. The IR filter (No. 87) is located at the eyepiece. As a sample exposure, found for sectioned human dentin, the following factors were involved: ×90, N.A. 0.35, Kohler illumination—15 min with the high-speed IR film.

8.18 INFRARED COLOR FILM

A relatively new capability of IR film is that of forming usefully modified ("false") color renditions of numerous subjects. It is widely used in remote-sensing surveys.

Tristimulus dye layers (see Fig. 8.8) in the emulsions serve to render the basic original colors one step away from the invisible IR as follows: IR to red, red to green, and green to blue. Blue colors are recorded black by the use of a yellow Kodak Wratten No. 12 filter over the camera lens. Figure 8.10 traces the colors of the subject through the various layers. Figure 8.11 indicates the sensitivity characteristics of the various emulsions.

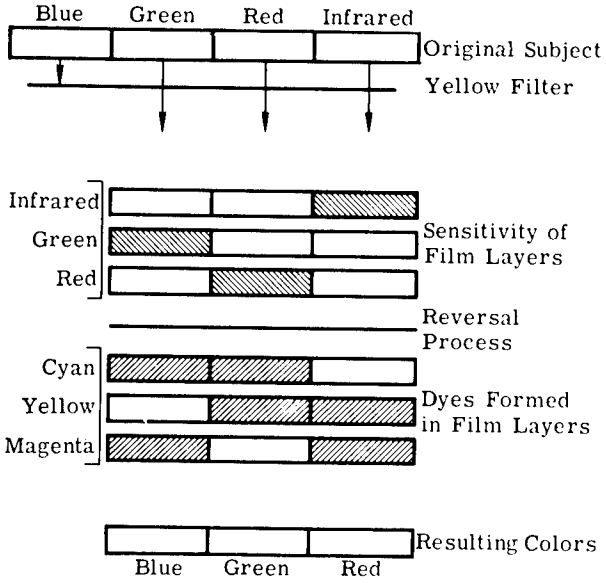


Fig. 8.10 Color formation with Kodak Ektachrome IR film.[7]

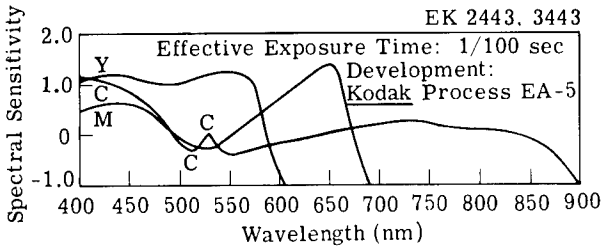


Fig. 8.11 Spectral sensitivity curves of Kodak Aerochrome IR film 2443 (Estar base) and Kodak Aerochrome IR film 3443 (Estar thin base).[7]

The detection of very small changes in the reflectance properties of distant objects in the green, red, and near-IR spectral bands can be enhanced. The gamma of the three layers may be greater than the gamma of normal color film. The greater gamma results in greater density changes and, hence, greater color shifts in the layers (and, therefore, in the image) in response to small changes in the spectral distribution of the spectral radiance of the scene.

These large values of gamma make proper exposure difficult. Noticeable color changes in the processed image occur when the exposure of these films is changed by one camera stop (a factor of 2 in one exposure). Small changes in the spectral irradiance of the illuminant are also detectable by the film. The sun's angle and the sky's conditions may alter the balance between irradiance in the near-IR band and the visible spectral range. Such alteration will not be evident to the eye or to a normal photographic light meter. By limiting the use of color IR film to clear days and high-sun-elevation conditions, one may be fairly confident that the spectral balance between visible and IR bands will be consistent.

8.19 KODAK LISTINGS

The major source of IR materials is the Eastman® Kodak Company. Some former suppliers have indicated to the author that they no longer provide such items; others have not replied to inquiry. Kodak materials are listed herewith. Because of susceptibility to heat, some types are stocked at the factory rather than at the dealers.

- Kodak high-speed IR film 2481 (Estar base): on 4-mil base, 135–36, 35-mm rolls, in one-roll units
- Kodak high-speed IR film 4143 (Estar thick base): on 7-mil base, 4- × 5-in. sheets, 25 per package
- Kodak Aerographic IR film 2424 (Estar base): effective aerial film speed EAFS 400, no filter, for haze conditions, 70 mm × 150 ft, factory stocked, or 5 in. × 150 or 350 ft, 9.5 in. × 125, 350, and 500 ft, 250 factory stocked, in single rolls

- Kodak Aerochrome® IR film 2443 (Estar base): false-color IR emulsion on 4-mil Estar base, fast-drying backing, process EA-5, EAFS 40 with Kodak Wratten filter No. 12, 5 in. × 100, 300, and 600 ft, 9.5 in. × 125, 200, and 400 ft; other sizes: 125 and 200 ft, factory stocked
- Kodak Aerochrome MS film 2448 (Estar base): false-color IR emulsion on 4-mil Estar base, fast-drying backing, EAFS 32 with Kodak Wratten filter No. 12, 70 mm × 150 ft, 1 roll, factory stocked, 5 in. × 100, 300, and 600 ft, 9.5 in. × 200 and 400 ft, and factory stocked—125 ft
- Kodak spectroscopic films and plates are manufactured to order, type I-N peaks at around 800 nm and responds to about 900 nm. Type I-Z responds to 1150 nm and peaks at 1080 nm.

Information regarding processing and special factors for specific work can be obtained from the Eastman Kodak Company. The illustrations in this chapter depict the fundamental characteristics of the films and show readers the basic measure of the parameters involved. For precise quantitative values in meticulous investigations the photographer has recourse to personal densitometry and experimentation.

8.20 LASER IMAGE SETTING

The 3M Printing and Publishing Systems Division provides film and paper for image setting with an IR (780-nm) laser diode exposure source. While the techniques involved in their use are not related to the main topic of this chapter, IR readers from the graphic arts are likely to be interested in investigating these imaging items.

The materials are 3M Imagesetting IR film (negative acting) and 3M Imagesetting IR paper (resin coated). They are available in roll form up to 40 in. wide and in sheets. They are both suitable for graphics and type. The film can be used for color separations and the paper for halftones.

They are handled, before processing, under green safelighting such as that provided by the Kodak Wratten No. 7 filter or Encapsulite T-40 fluorescent safelight. The materials are compatible with the solutions for the rapid-access, hybrid, and lithography processes.

References

1. *Kodak Infrared Films,* Kodak Publication No. N-17, Eastman Kodak Company (1973).
2. H. L. Gibson, *Photography by Infrared,* John Wiley & Sons, Inc., New York (1978).
3. *Kodak Plates and Films for Scientific Photography,* Kodak Publication No. P-9, Eastman Kodak Company (1973).
4. *Scientific Imaging with Kodak Films and Plates,* Kodak Publication No. P-315, Eastman Kodak Company (1987).
5. A. C. Hardy and F. H. Perrin, *The Principles of Optics,* McGraw-Hill, New York (1932).
6. G. Nieuwenhuis, "Lens focus shift required for ultraviolet and infrared photography," *Journal of Biological Photography* **59,** 17–20 (1991).
7. *Kodak Data for Aerial Photography,* Kodak Publication No. M-29, Eastman Kodak Company (1971, 1976).
8. *Applied Infrared Photography,* Kodak Publication No. M-28, Eastman Kodak Company (1967, 1970, and 1980).

9. E. Welander, "Some methods and investigations for determining the quality of aerial photographs," *Photogrammetrie* **14,** 28–36 (1962).
10. C. J. Niederpruem, C. N. Nelson, and J. A. C. Yule, "Contrast index," *Photographic Science and Engineering* **10,** 35–41 (1966).
11. R. L. Lamberts, "Measurements of sine-wave response for a photographic emulsion," *Journal of the Optical Society of America* **69,** 425–428 (1959).
12. R. L. Lamberts, "The production and use of variable-transmittance sinusoidal test objects," *Applied Optics* **2,** 273–276 (1963).
13. C. N. Nelson, "Prediction of densities in fine detail in photographic images," *Photographic Science and Engineering* **49,** 151–156, 239–249 (1971).
14. F. H. Perrin, "Methods of appraising photographic systems," *Photographic Science and Engineering* **15,** 82–97 (1960).
15. C. Dhéré and O. Biermacher, "Spectrochimie bioloque, *Comptes Rendus* **203,** 412–414 (1936).
16. H. L. Gibson, "The photography of infrared luminescence," *Medical and Biological Illustration* **12,** Part I (1962); Parts II and III (1963).

CHAPTER 9

Reticles

Richard Legault
Institute for Defense Analyses
Washington, D.C.

CONTENTS

9.1 INTRODUCTION

This chapter introduces some analytic tools useful in the design and analysis of electro-optical systems. Emphasis is placed on the spatio-temporal analysis of reticles. The designer or analyst of electro-optical systems will not find models treated in this chapter. He or she must then rely on other data: tables and computer-assisted computation.

Table 9.1 lists the symbols and definitions used in this chapter.

9.2 FOURIER ANALYSIS

Fourier methods provide the analyst with methods for determining the frequency content of a signal.

9.2.1 Fourier Series—One Dimensional

If $v(t)$ is a periodic function of period T, i.e., $v(t) = v(t + T)$, then $v(t)$ may be written in exponential form

$$v(t) = \sum_{-\infty}^{\infty} c_n \exp(i2\pi nt/T) \ , \tag{9.1}$$

where

$$c_n = (1/T)\int_0^T v(t) \exp(-i2\pi nt/T)\, dt \ ,$$

or in trigonometric form

$$v(t) = \tfrac{1}{2}a_0 + \sum_{1}^{\infty}\left(a_n \cos\frac{2\pi nt}{T} + b_n \sin\frac{2\pi nt}{T}\right) \ , \tag{9.2}$$

where

$$a_n = 2/T\int_{-T/2}^{T/2} v(t) \cos(2\pi nt/T)\, dt$$

$$b_n = 2/T\int_{-T/2}^{T/2} v(t) \sin(2\pi nt/T)\, dt \ .$$

The term $2\pi/T$ may be written as ω_0 in units of radians per unit time.

9.2.2 Fourier Integral—One Dimensional

If the integrals $|v(t)|$ or $v^2(t)$ exist over the integration path, then

$$v(t) = \int_{-\infty}^{\infty} V(f) \exp(2\pi itf)\, df \ , \tag{9.3}$$

Table 9.1 Symbols and Definitions

Symbol	Definition
A	Parallelogram defined by $\mathbf{a}_1$ and $\mathbf{a}_2$; region defined by the aperture
$A(\mathbf{k})$	Fourier transform of $a(\mathbf{x})$; and optical response
$A_m(\mathbf{k})$	Fourier transform of $a_m(\mathbf{x})$
a,b,c	Coefficients
$\mathrm{a}(\mathbf{x})$	Aperture transmission pattern function
$a_m(x)$	Coefficients
$a(\rho)$	Optimal scanning aperture
$b(\mathbf{x})$	Display response
$d(\mathbf{x})$	Response of detector over its surface
f	Temporal frequency
G	Fourier transform of g
g	Function
H	Hadamard matrix
I	Intensity; also, identity matrix
$\mathbf{k}$	Spatial frequency, $\mathbf{k} = (k_1 k_2)$, in units of cycles per unit length
k_ρ,ϕ	Polar coordinate transforms
m	Running index; 0,1,2,3...
n	Running index; 0,1,2,3 ...
$O(\mathbf{k})$	Fourier transform of $o(\mathbf{x})$
$o(\mathbf{x})$	Display output spatial signal
p	Number of spoke pairs
R	Matrix, (r_{ij})
$R(\mathbf{k})$	Fourier transform of $r(\mathbf{x})$
r_{ij}	Matrix element
$r(\mathbf{x})$	Reticle transmission coefficient
$S(\mathbf{k})$	Fourier transform of $s(\mathbf{x})$
$s(\mathbf{x})$	Average scene radiation distribution
$s_\omega(\mathbf{x})$	Random scene radiation distribution
T	Period; as superscript, transpose of matrix; as subscript, target
$T(\mathbf{k})$	Fourier transform of target $t(\mathbf{x})$
t	Time
$t(\mathbf{x})$	Target function
V,k,λ	Integers
$v(t)$	Voltage signal output from sensor detector
$v_\omega(t)$	Random voltage signal
$W(\mathbf{k})$	Wiener spectrum
x,y	Coordinates
β	Parallelogram defined by $\mathbf{b}_1$ and $\mathbf{b}_2$
β_m	Modulation coefficients
δ	Phase
$\delta(\mathbf{x} - \mathbf{x}_o)$	Delta function
λ	Wavelength
ρ,θ	Polar coordinates
σ	Bar width in translating reticle
$\phi(\mathbf{x})$	Second-order correlation statistic
Ω	A set of real numbers
ω	Average domain
ω_0	Angular frequency, $\omega_0 = 2\pi/T$
-1	As superscript, inverse of matrix
$*$	As superscript, complex conjugate

with $v(t)$ and $V(f)$ called transform pairs. The transform pair $V(f)$ may be obtained from $v(t)$ by the relation

$$V(f) = \int_{-\infty}^{\infty} v(t) \exp(2\pi itf)\, dt \tag{9.4}$$

A number of excellent compilations of Fourier series expansions and Fourier transform pairs exist. References 1 and 2 contain two such important compilations.

9.2.3 Fourier Series—Two Dimensional

The definition of two-dimensional (2-D) periodicity and the 2-D counterpart of the one-dimensional (1-D) reciprocal $1/T$ are less familiar. Let $\mathbf{a}_1 = (x_1,x_2)$ and $\mathbf{a}_2 = (y_1,y_2)$ be any two noncolinear vectors in the plane, and n_1,n_2 any two positive or negative integers, then

$$\mathbf{a}_n = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 \ . \tag{9.5}$$

The vectors (points) defined by $\mathbf{a}_n$ in Eq. (9.5) form a lattice in the plane x_1', x_2', depicted in Fig. 9.1. The definition of periodicity is given by

$$s(\mathbf{x}) = s(\mathbf{x} + \mathbf{a}_n) \tag{9.6}$$

for all $\mathbf{n} = (n_1, n_2)$ integers, and can be predicted by referring to the figure. The vectors (points) $\mathbf{a}_n$ break the plane up into parallelograms. A periodic pattern is one that repeats itself from one parallelogram to the next. The pattern in one parallelogram equals the pattern in another.

The vectors defining the reciprocal lattice are

$$\mathbf{b}_n = n_1\mathbf{b}_1 + n_2\mathbf{b}_2, \quad n_1,n_2 \text{ integer} \ . \tag{9.7}$$

The generating vectors $\mathbf{b}_1$, $\mathbf{b}_2$ are derived from $\mathbf{a}_1$, $\mathbf{a}_2$ by use of the notion of dot or inner product:

$$\mathbf{a}_1 \cdot \mathbf{a}_2 = (x_1,x_2) \cdot (y_1,y_2) = x_1,y_1 + x_2y_2 \ . \tag{9.8}$$

This results in four linear equations in four unknowns:

$$\begin{aligned} \mathbf{b}_1 \cdot \mathbf{a}_2 &= 0 \qquad & \mathbf{b}_1 \cdot \mathbf{a}_1 &= 1 \\ \mathbf{b}_2 \cdot \mathbf{a}_1 &= 0 \qquad & \mathbf{b}_2 \cdot \mathbf{a}_2 &= 1 \ . \end{aligned} \tag{9.9}$$

The first two equations specify that the spatial and reciprocal lattice generating vectors must be perpendicular, as illustrated in Fig. 9.2.

The 2-D Fourier series expansion of a spatial pattern is given by

$$s(\mathbf{x}) = \sum_n s_n \exp(2\pi i\mathbf{x} \cdot \mathbf{b}_n) \ , \tag{9.10}$$

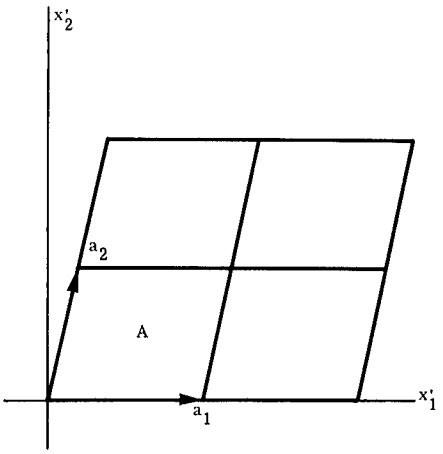


Fig. 9.1 Spatial periodic lattice a_n.

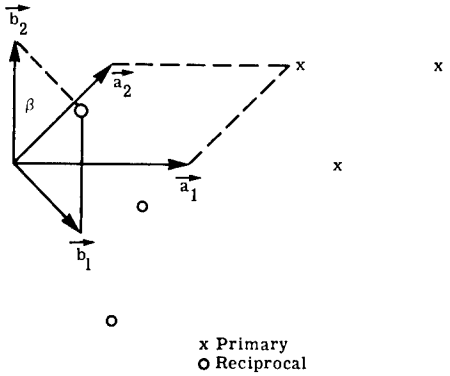


Fig. 9.2 Primary and reciprocal lattice.

where

$$s_n = \int_A s(\mathbf{x}) \exp(-2\pi i \mathbf{x} \cdot \mathbf{b}_n)\, d\mathbf{x}$$

A = the parallelogram defined by the vectors $\mathbf{a}_1$ and $\mathbf{a}_2$.

9.2.4 Fourier Transform—Two Dimensional

If the integral $\int_{-\infty}^{\infty} s^2(\mathbf{x})\, d\mathbf{x} < \infty$ exists, then

$$s(\mathbf{x}) = \int_{-\infty}^{\infty} S(\mathbf{k}) \exp(2\pi i \mathbf{k} \cdot \mathbf{x})\, d\mathbf{k} \ , \tag{9.11}$$

where $\mathbf{k}$ equals (k_1,k_2) and $s(\mathbf{x})$ and $S(\mathbf{k})$ are Fourier transform pairs. If $s(\mathbf{x})$ is known, then $S(\mathbf{k})$ can be found from the relation

$$S(\mathbf{k}) = \int_{-\infty}^{\infty} s(\mathbf{x}) \exp(2\pi i \mathbf{k} \cdot \mathbf{x})\, d\mathbf{x} \ . \tag{9.12}$$

A spatial frequency representation $S(\mathbf{k})$ gives the amplitude of the sinusoidal wave with spatial frequency k_1 in the x_1 direction and spatial frequency k_2 in the x_2 direction. Figure 9.3 depicts an approximation to such a wave with $k = k_1 = k_2$. Clearly, negative spatial frequencies have a physical interpretation.

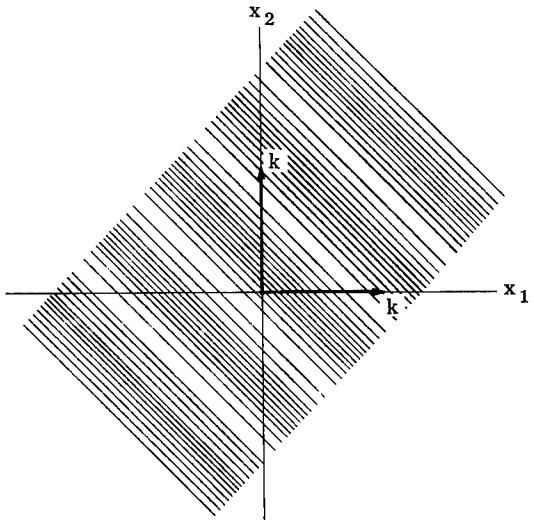


Fig. 9.3 Bar approximation to $\cos 2\pi(k_x x + k_y y)$.

9.2.5 Properties of Fourier Transform Pairs

Certain relationships between Fourier transform pairs will be required in the further development.

Scale changes: $s(ax_1,bx_2) \leftrightarrow \dfrac{1}{|a||b|} S\left(\dfrac{k_1}{a}, \dfrac{k_2}{b}\right)$,

Translation: $s(\mathbf{x} + \mathbf{x}_0) \leftrightarrow S(\mathbf{k}) \exp(-2\pi i \mathbf{k} \cdot \mathbf{x}_0)$,

Conjugates: $s^*(\mathbf{x}) \leftrightarrow S^*(-\mathbf{k})$, $s(\mathbf{x})$ complex

$s^*(\mathbf{x}) = s(\mathbf{x}) \leftrightarrow S(\mathbf{k})$, $s(\mathbf{x})$ real ,

$S^*(\mathbf{k}) = S(-\mathbf{k})$,

where $s^*(\mathbf{x})$ is the complex conjugate of $s(\mathbf{x})$.

9.2.6 Fourier Transforms—Polar Coordinates

Many times it is convenient to perform the analysis in polar coordinates, particularly when spatial patterns are constant radially or angularly. The usual coordinate transforms for $\mathbf{x} = (x_1,x_2)$ are

$$x_1 = \rho \cos\theta \qquad (9.13a)$$
$$x_2 = \rho \sin\theta$$

and for $\mathbf{k} = (k_1,k_2)$ are

$$k_1 = k_\rho \cos\phi$$
$$k_2 = k_\rho \sin\phi \ . \tag{9.13b}$$

The Fourier transform pair relationship is

$$\begin{aligned} s(\rho,\theta) &= \int_0^\infty \int_0^{2\pi} S(k_\rho,\phi) \exp[2\pi i k_\rho \rho \cos(\theta-\phi)] k_\rho \, d\phi \, dk_\rho \ , \\ S(k_\rho,\phi) &= \int_0^\infty \int_0^{2\pi} s(\rho,\theta) \exp[-2\pi i k_\rho \rho \cos(\theta-\phi)] \rho \, d\theta \, d\rho \ . \end{aligned} \tag{9.14}$$

9.2.7 Correlation and Convolution

A major reason for utilizing Fourier methods in the analysis of electro-optical systems is the analytic convenience for some types of systems. The reason for this convenience is found in the following relationships.

Parseval's Theorem. If $s^{(1)}(\mathbf{x})$, $s^{(2)}(\mathbf{x})$ are periodic over the parallelograms A, then

$$\int_A s^{(1)}(\mathbf{x}) s^{(2)}(\mathbf{x}) \, d\mathbf{x} = \sum_{n=-\infty}^{\infty} s_n^{*(1)} s_n^{(2)} \ . \tag{9.15}$$

If $s_1(\mathbf{x})$ and $s_2(\mathbf{x})$ have Fourier transforms, then

$$\int_{-\infty}^{\infty} s_1(\mathbf{x}) s_2(\mathbf{x}) \, d\mathbf{x} = \int_{-\infty}^{\infty} S_1^*(\mathbf{k}) S_2(\mathbf{k}) \, d\mathbf{k} \ . \tag{9.16}$$

Convolution. The signal at a spatial point may be the weighted sum of the signals at surrounding points. Then the output spatial signal $o(\mathbf{x})$ is written as a convolution integral

$$o(\mathbf{x}) = \int_{-\infty}^{\infty} g(\mathbf{x} - \mathbf{y}) s(\mathbf{y}) \, d\mathbf{y} \ . \tag{9.17}$$

The Fourier transform of $o(\mathbf{x})$ takes a simple form

$$O(\mathbf{k}) = G(\mathbf{k}) S(\mathbf{k}) \ . \tag{9.18}$$

Correlation. A similar result is found for the correlation integral

$$o(\mathbf{x}) = \int_{-\infty}^{\infty} g(\mathbf{x} + \mathbf{y}) s(\mathbf{y}) \, d\mathbf{y} \ . \tag{9.19}$$

The Fourier transform of $o(\mathbf{x})$ takes the form

$$O(\mathbf{k}) = G(\mathbf{k}) S^*(\mathbf{k}) \ . \tag{9.20}$$

9.2.8 Wiener Spectrum

Analysis may require only the statistics of the output signal. Further, one may find that background scenes do not have a Fourier transform in the usual sense. Each value of ω identifies a sample $s_\omega(\mathbf{y})$ from some 2-D process. The second-order correlation statistic can be defined as

$$\underset{\omega}{\text{average}}\, s_\omega(\mathbf{y}) s_\omega(\mathbf{x} + \mathbf{y}) \ . \tag{9.21}$$

It is usually assumed that the process is stationary. Then the average in expression (9.21) depends only on the displacement, i.e.,

$$\underset{\omega}{\text{average}}\, s_\omega(\mathbf{y}) s_\omega(\mathbf{x} + \mathbf{y}) = \phi(\mathbf{x}) \ . \tag{9.22}$$

If the statistics of a single realization of the random process represent the process (ergodicity), then the average $\phi(\mathbf{x})$ can be written as

$$\phi(\mathbf{x}) = \int_{-\infty}^{\infty} s(\mathbf{y}) s(\mathbf{x} + \mathbf{y})\, d\mathbf{y} \ . \tag{9.23}$$

The Fourier transform of $\phi(\mathbf{x})$ is called the Wiener spectrum of the scene and is defined by

$$W(\mathbf{k}) = \int_{-\infty}^{\infty} \phi(\mathbf{x}) \exp(-2\pi i \mathbf{k} \cdot \mathbf{x})\, d\mathbf{x} \ . \tag{9.24}$$

R. C. Jones[3] postulated a model for the Wiener spectrum defined by

$$W(\mathbf{k}) = \frac{B}{(k_0^2 + \mathbf{k} \cdot \mathbf{k})^q} \ , \tag{9.25}$$

where

k_0 = constant
q = a measure of the fuzziness of the disk edges
B = total scene radiance.

Note that the Wiener spectrum is the frequency representation of the auto-correlation of the scene. It is the distribution of the total power over the spatial frequencies and is a second-moment statistic. It is also a representation of the noise power in the scene, but only in specialized cases can it be used as an estimator of false alarms that usually are signals exceeding a high threshold.

9.3 SCANNING APERTURE

Initially, the aperture transmission pattern $a(x_1,x_2)$ is centered on the origin of the scene-coordinate system. Suppose at time t the aperture is moved, but not rotated, and centered on the scene point $[x_1(t), x_2(t)]$. The output signal is given by

$$g[\mathbf{x}(t)] = \int_{-\infty}^{\infty} a(\mathbf{x}) s[\mathbf{x} + x(t)]\, d\mathbf{x} \ , \tag{9.26}$$

where $a(\mathbf{x})$ is the aperture transmission pattern and $s(\mathbf{x})$ is the radiant signal from the scene point $\mathbf{x}$. If the aperture is translated so that the center of the pattern is pointed in turn at every scene point $\mathbf{x}$, the function $g[\mathbf{x}(t)]$ may be considered as a filtered version of the original imaged scene $s(\mathbf{x})$.

The Wiener spectrum $W_g(\mathbf{k})$ of $g(\mathbf{x})$ is given by

$$W_g(k_1,k_2) = |A(k_1,k_2)|^2 W_S(k_1,k_2) \ , \tag{9.27}$$

where $A(k_1,k_2)$ is the Fourier transform of $a(\mathbf{x})$ and $W_S(k_1,k_2)$ is the Wiener spectrum of the scene. The filtering of the scene by the reticle system is obvious and the Fourier transform of the aperture transmission pattern is seen to be descriptive of scanning aperture performance.

It is of considerable interest that this formulation of the scanning aperture model permits an optimization of the scanning aperture. The criterion for optimization is the maximization of the ratio of the instantaneous target signal squared to the mean-squared background signal, that is, maximization of

$$\frac{\left| \int_{-\infty}^{\infty} A^*(\mathbf{k}) T(\mathbf{k})\, d\mathbf{k} \right|^2}{\int_{-\infty}^{\infty} |A(\mathbf{k})|^2 W_B(\mathbf{k})\, d\mathbf{k}} \ , \tag{9.28}$$

where

$A(\mathbf{k})$ = aperture Fourier transform
$T(\mathbf{k})$ = target Fourier transform
$W_B(\mathbf{k})$ = Wiener spectrum of background.

Using the Schwarz inequality, one finds the aperture $A(\mathbf{k})$, which maximizes the ratio (9.28). The Schwarz inequality is given by

$$\left| \int g(\mathbf{x}) s(\mathbf{x}) \right|^2 \leq \int g^2(\mathbf{x})\, d\mathbf{x} \int s^2(\mathbf{x})\, d\mathbf{x} \ . \tag{9.29}$$

The upper bound is obtained when $g(\mathbf{x}) = s(\mathbf{x})$. Ratio (9.28) is now written as

$$\begin{aligned} \frac{\left| \int_{-\infty}^{\infty} A^*(\mathbf{k}) W_B^{1/2}(\mathbf{k}) \dfrac{T(\mathbf{k})}{W_B^{1/2}}\, d\mathbf{k} \right|^2}{\int_{-\infty}^{\infty} |A(\mathbf{k})|^2 W_B(\mathbf{k})\, d\mathbf{k}} &\leq \frac{\int |A(\mathbf{k})|^2 W_B(\mathbf{k})\, d\mathbf{k} \int \dfrac{T^2(\mathbf{k})}{W_B(\mathbf{k})}\, d\mathbf{k}}{\int_{-\infty}^{\infty} |A(\mathbf{k})|^2 W_B(\mathbf{k})\, d\mathbf{k}} \\ &= \int_{-\infty}^{\infty} \frac{T^2(\mathbf{k})}{W_B(\mathbf{k})}\, d\mathbf{k} \ . \end{aligned} \tag{9.30}$$

Fortunately, the upper bound is independent of $A(\mathbf{k})$ and is obtained when

$$A(\mathbf{k}) = \frac{T^*(\mathbf{k})}{W_B(\mathbf{k})} \ . \tag{9.31}$$

The optimal scanning aperture specified in Eq. (9.31) is the 2-D counterpart of the matched filter of electronics. This result supports the feeling that a scanning aperture should essentially be matched to the target shape but modified by the spatial characteristics of the background.

Consider an example proposed by Jones. The target is assumed to be a Gaussian pulse:

$$t(x_1,x_2) = a \exp[-(x_1^2 + x_2^2)/2b] \ , \tag{9.32}$$

where a is the peak radiance of Gaussian pulse and b is the second moment of radiance density of Gaussian pulse. The Wiener spectrum of the background is given by

$$W_B(k_1,k_2) = \frac{b}{k_1^2 + k_2^2} \ . \tag{9.33}$$

The optimal scanning aperture is given by

$$a(\rho) = \left(1 - \frac{\rho^2}{2b}\right) \exp(-\rho^2/2b) \ . \tag{9.34}$$

Unfortunately, for some targets the time scale of events is so short that it is not possible to scan the required field of view with a small instantaneous field of view within the time required and keep within the response time of available detectors.

In general, two solutions exist to the dilemma. The straightforward one is to scan many detectors, each scanning some smaller part of the total field of view, the number selected to permit covering the total field of view in the required time. Another, and more commonly employed solution, is the use of a reticle.

9.4 RETICLE SYSTEMS

A reticle system is essentially a mask or pattern placed in the image plane of an optical system. The transmission of this mask varies spatially. Usually the mask transmits certain portions of the imaged scene and completely blocks other portions of the scene. The radiation from the transmitted portions is focused on a detector. The detector output is assumed to be proportional to the total incident radiation. The reticle mask may be moved in the image plane, the imaged scene may be moved over a fixed reticle mask, or both. Reference 4 contains a fairly complete exposition of the current uses of reticle systems.

The reticle mask in scene coordinates is specified by a real-valued function $r(\mathbf{x},t)$. The function $r(\mathbf{x},t)$ specifies the transmission coefficient for the intensity of an image scene point $\mathbf{x}$ at time t. The radiation distribution of the image scene in scene coordinates is represented by a positive real-valued function

$s(\mathbf{x})$. Since the transmitted fluxes are integrated, the output $v(t)$ from a reticle system is

$$v(t) = \int_{-\infty}^{\infty} r(\mathbf{x},t)s(\mathbf{x})\, d\mathbf{x} \ . \tag{9.35}$$

Consequently, the scene $s(\mathbf{x})$ is encoded into a temporal signal $v(t)$ by a reticle system. Equation (9.35) is a general model of a reticle system.

The correlation properties of the output signal give

$$\begin{aligned} &\underset{\omega}{\text{average}}\ v_\omega(t)v_\omega(t+\tau) \\ &= \underset{\omega}{\text{average}} \int_{-\infty}^{\infty} r(\mathbf{x},t)s_\omega(\mathbf{x})\, d\mathbf{x} \int_{-\infty}^{\infty} r(\mathbf{x}',\, t+\tau)s_\omega(\mathbf{x}')\, d\mathbf{x}' \ , \\ &= \int_{-\infty}^{\infty} \underset{\omega}{\text{average}}\ s_\omega(\mathbf{x})s(\mathbf{x}+\mathbf{x}'') \int_{-\infty}^{\infty} r(\mathbf{x},t)r(\mathbf{x}+\mathbf{x}'',t+\tau)\, d\mathbf{x}\, d\mathbf{x}'' \ . \end{aligned} \tag{9.36}$$

If the process is stationary, average $s_\omega(\mathbf{x})s(\mathbf{x}+\mathbf{x}'')$ is independent of $\mathbf{x}$, and its transform pair is the Wiener spectrum $W(\mathbf{k})$. Using the Parseval relation [Eq. (9.16)], one obtains

$$\underset{\omega}{\text{average}}\ v_\omega(t)v_\omega(t+\tau) = \int_{-\infty}^{\infty} W(\mathbf{k})R(-\mathbf{k},t)R(\mathbf{k},\, t+\tau)\, d\mathbf{k} \ . \tag{9.37}$$

The correlation and ultimately the power spectrum of $v(t)$ are weighted averages of the scene's Wiener spectrum. The weights are derived from the reticle patterns.

Reticles are commonly used in guidance systems. The reticle and its motion encodes the coordinates of the object to be located. Generally, this object is assumed to be a point source. It is necessary to determine how an object's coordinates modulate the signal from the reticle system. The models considered are too general for this purpose.

9.4.1 Analysis of Reticle Modulation

A stylized reticle system is illustrated in Figs. 9.4 and 9.5. The general reticle system considered consists of a transparent aperture in the image plane with a reticle pattern moving across the aperture. The area A and shape of the aperture are independent of time. Take a coordinate system fixed in the aperture A. Since the reticle moves across the aperture, the reticle pattern transmission will be time dependent in aperture coordinates. As the aperture is scanned, the scene will be time dependent in aperture coordinates. Then

$$v(t) = \int_A r(\mathbf{x},t)s(\mathbf{x},t)\, d\mathbf{x} \ , \tag{9.38}$$

where

$\mathbf{x}$ = a point in A
$r(\mathbf{x},t)$ = reticle transmission
$s(\mathbf{x},t)$ = scene radiation distribution in the aperture-limited plane
$d\mathbf{x}$ = an elemental area in A
$v(t)$ = voltage output from the detector.

Here, $\mathbf{x}$ may be any set of two coordinates specifying a point in the plane; for example, $\mathbf{x} = (x_1,x_2)$ if Cartesian coordinates are used, or $\mathbf{x} = (\rho,\phi)$ if polar coordinates are used. The integral on the right side of Eq. (9.38) is independent of the coordinate choice. The coordinate system most convenient for calculation is dictated by the geometry of the aperture. If A is rectangular, Cartesian coordinates make calculation simplest; if A is circular, polar coordinates are the choice.

A reticle pattern is finite. The reticle seen through the aperture is a moving pattern. This pattern is repeated at regular temporal intervals. The assumption of periodicity stated in Eq. (9.39) below does not restrict the class to realizable reticle systems:

$$r(\mathbf{x},t) = r\left(\mathbf{x}, t + \frac{2\pi}{\omega_0}\right) \quad \text{for all } \mathbf{x} \ . \tag{9.39}$$

The fundamental frequency ω_0 should not be confused with any rotational or translational frequency; they may or may not correspond. Generally, a simple relation exists between the two frequencies. The reticle function defined in aperture coordinates has a Fourier series representation in time, with fundamental frequency ω_0:

$$\begin{aligned} r(\mathbf{x},t) &= \sum_{m=-\infty}^{\infty} a_m(\mathbf{x}) \exp(im\omega_0 t) \ , \\ a_m(\mathbf{x}) &= \frac{\omega_0}{2\pi}\int_0^{2\pi/\omega_0} r(\mathbf{x},t) \exp(-im\omega_0 t)\, dt \ . \end{aligned} \tag{9.40}$$

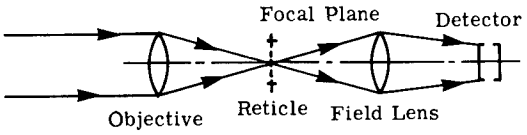


Fig. 9.4 Conceptual reticle system.

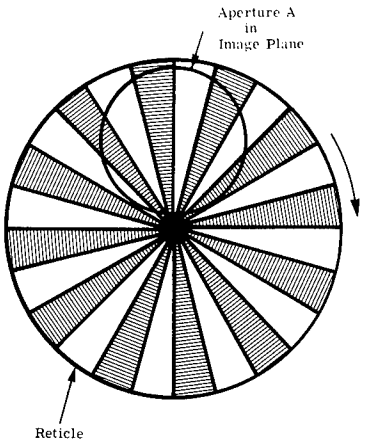


Fig. 9.5 Reticle aperture and motion.

The substitution of Eq. (9.40) in Eq. (9.38) yields

$$\begin{aligned} v(t) &= \sum_{m=-\infty}^{\infty} \beta_m(t) \exp(im\omega_0 t) \ , \\ \beta_m(t) &= \int_A a_m(\mathbf{x}) s(\mathbf{x},t)\, d\mathbf{x} \ . \end{aligned} \tag{9.41}$$

At this level, Eqs. (9.40) and (9.41) are revealing. Equation (9.41) is generally not a Fourier series even though Eq. (9.40) always is. The coefficients in the summation of Eq. (9.41) are time dependent. Each component $\exp(im\omega_0 t)$ of Eq. (9.41) may be considered as a carrier. Each carrier is modulated by a temporal function $\beta_m(t)$ reflecting the scene information, in particular an object's coordinates. When the aperture is not scanned, the scene is time dependent. The analysis of tracking systems generally assumes $s(\mathbf{x},t)$ is not time dependent, i.e., the scene does not change much before the target is located. The reticle pattern is reflected in the coefficients $a_m(\mathbf{x})$ and in the interaction of scene and reticle pattern by $\beta_m(t)$.

The time dependence of $v(t)$ has two origins; reticle and scan motion. A portion of the temporal dependence, in particular ω_0, arising from reticle motion, is contained in the exponential terms, and all time dependence arising from aperture scan (and inherent scene time dependence) is contained in $\beta_m(t)$. However, $\beta_m(t)$ is affected by $a_m(\mathbf{x})$, and $a_m(\mathbf{x})$ is generally not independent of reticle motion. Inspection of Eq. (9.40) indicates the relation between reticle motion and $a_m(\mathbf{x})$. Reticle motion determines both ω_0 and the time dependence of $r(\mathbf{x},t)$. Generally, reticle pattern and reticle motion are inextricably interdependent in producing time dependence in $v(t)$. Equation (9.41) shows that a similar remark can be made about scene pattern and scan motion.

In some important cases, $a_m(\mathbf{x})$ is independent of reticle motion. For example, if a reticle is uniformly rotating, then ω_0 is its rotational frequency, and $a_m(\mathbf{x})$ is independent of ω_0.

In summary, the reticle and its motion are completely specified by ω_0 and the set $a_m(\mathbf{x})$. The voltage output is completely specified by ω_0 and the set $\beta_m(t)$. The harmonics of ω_0 are the carrier frequencies and $\beta_m(t)$, the modulation placed on the m'th harmonic by the interaction of scene and reticle patterns. To define a reticle system, one must calculate ω_0 and $a_m(\mathbf{x})$. To specify $v(t)$ for a given scene, one must calculate $\beta_m(t)$. In most cases, the convergence of Eq. (9.41) is sufficiently rapid to ensure the practicality of computing $v(t)$ numerically.

The basic method of reticle system analysis as represented in Eqs. (9.40) and (9.41) is to determine a fundamental frequency ω_0 for the system. The interaction of the system with the scene is seen as a modulation of the carrier frequencies. Target location is based on the relation between the target location in aperture-fixed coordinates and the modulation $\beta_m(t)$. Discrimination between targets and backgrounds is based on modulation differences. References 5 and 6 consider the modulation from various sources for a number of reticle systems.

9.4.2 Aperture Effects

The modulation coefficients $\beta_m(t)$ of Eq. (9.41) indicate integration over an aperture area A. The integral expression may be written with infinite limits and the introduction of an aperture function $a(\mathbf{x})$:

$$\beta_m(t) = \int_A a_m(\mathbf{x})s(\mathbf{x},t)\,d\mathbf{x} = \int_{-\infty}^{\infty} a_m(\mathbf{x})a(\mathbf{x})s(\mathbf{x},t)\,d\mathbf{x}\ , \tag{9.42}$$

where

$a(\mathbf{x})$ = 1, with $\mathbf{x}$ contained in A
= 0, elsewhere.

The calculation of modulation characteristics is usually facilitated by the use of Eq. (9.41). Introduction of the aperture function provides some insight and, occasionally, computational ease. Using the Parseval relation and the convolution theorem, one may write

$$\beta_m(t) = \int_{-\infty}^{\infty} A_m^*(\mathbf{k})S'(\mathbf{k})\,d\mathbf{k}\ , \tag{9.43}$$

where

$A_m^*(\mathbf{k})$ = the conjugate of the Fourier transform of $a_m(\mathbf{x})$

$$S'(\mathbf{k}) = \int_{-\infty}^{\infty} S(\mathbf{k}')O(\mathbf{k} - \mathbf{k}')\,d\mathbf{k}'\ .$$

The aperture smears the scene in the spatial frequency.

The transform of a rectangle with sides of length a and b is

$$A(\mathbf{k}) = A(k_1,k_2) = \frac{\sin\pi ak_1 \sin\pi bk_2}{\pi_2 k_1 k_2}\ . \tag{9.44}$$

The transform of a circular aperture of radius a is

$$A(\mathbf{k}) = A(k_1,k_2) = \frac{aJ_1[2\pi a(k_1^2 + k_2^2)^{1/2}]}{(k_1^2 + k_2^2)^{1/2}}\ , \tag{9.45}$$

where J_1 is a Bessel function of the first order. The most common aperture is circular.

9.4.3 Reticle Motion

While many types of reticle motion are mathematically possible, convenient mechanical implementation tends to drive the designer to translation, rotation, and nutation of the reticle.

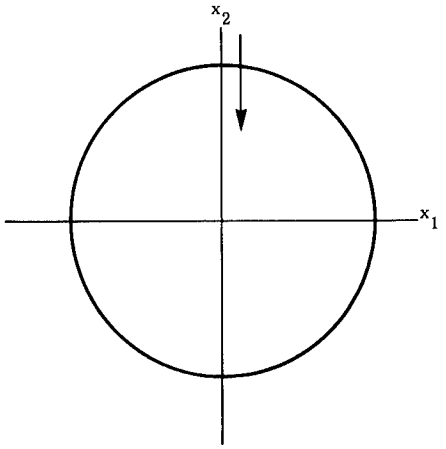


Fig. 9.6 Aperture coordinates.

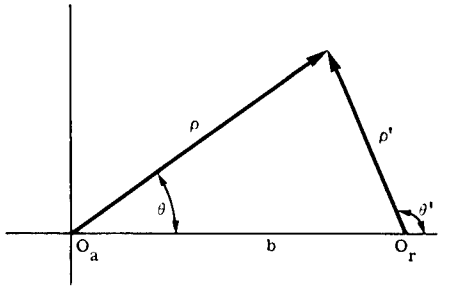


Fig. 9.7 Coordinate transforms for rotation. The variable O_a equals aperture center and O_r equals center of rotation, usually taken as the center of the reticle pattern.

The simplest translational motion is along one axis of the aperture-fixed coordinates that are used to express the reticle pattern (see Fig. 9.6). If the translational motion is in the negative x_2 direction,

$$r(x_1,x_2) = r(x_1, x_2 + st) \ , \tag{9.46}$$

where s is the velocity. Thus, if the point $(x_1,x_2) = \mathbf{x}$ in the aperture is seen at time t, the point $(x_1, x_2 + st)$ is seen on the reticle.

The next form of motion to be considered is the rotating reticle. The center of rotation may not coincide with the aperture center, as shown in Fig. 9.7. The reticle pattern is usually obtained in rotational-center coordinates (ρ',θ'), but the relation can be transformed to aperture-centered coordinates (ρ,θ) by

$$r(\rho,\theta, t) = r'[\rho'(\rho,\theta),\theta'(\rho,\theta),t] \ . \tag{9.47}$$

From the law of cosines, and the fact that three complex vectors, which form a triangle, sum to zero, one obtains

$$\begin{aligned} \rho' &= \rho'(\rho,\theta) = (\rho^2 + b^2 - 2\rho b \cos\theta)^{1/2} \ , \\ \exp(-i\theta') &= \exp[-i\theta'(\rho,\theta)] = [\rho \exp(-i\theta) - b] \\ &\quad \times (\rho^2 + b^2 - 2\rho b, \cos\theta)^{1/2} \ . \end{aligned} \tag{9.48}$$

A nutating system centers the reticle in the aperture and nutates the scene, as shown in Fig. 9.8. Reticle motion is the same as for an aperture-centered reticle (Fig. 9.7). The scene center O_S is rotated around the aperture center O_a. The radius of scene nutation is c. Generally, scene nutation is at a uniform rate:

$$\theta(t) = \Omega_0 t \ , \tag{9.49}$$

where

$\Omega_0 = 2\pi/T$
T = period.

The temporal dependence of the scene $s(\rho,t)$ becomes

$$s(\rho,t) = s(\rho,\theta,t) = s[\rho'(\rho,\theta)\theta'(\rho,\theta)t] = s[\rho'(\rho, \theta - \Omega_0 t)\theta'(\rho, \theta - \Omega_0 t)] \tag{9.50}$$

Analysis of the modulation of point sources requires the relations

$$\begin{aligned} \rho &= [c^2 + \rho'^2 - 2c\rho\cos(\theta' - \Omega_0 t)]^{1/2} \ , \\ \exp(-i\theta) &= \exp(-i\Omega_0 t)\{\rho' \exp[-i(\theta' - \Omega_0 t)] - c\}[\rho'^2 + c^2 - 2\rho' c \\ &\quad \times \cos(\theta' - \Omega_0 t)]^{-1/2} \ . \end{aligned} \tag{9.51}$$

This permits writing the intensity of a scene point (ρ',θ') in aperture coordinates. Note that Cartesian coordinates were chosen for translation, while polar coordinates were chosen for rotation and nutation. These are natural choices for each type of motion. The prevalent reticle motions are rotation, with the rotational center displaced from the aperture center, and simple nutation.

9.4.4 Reticle and Motion Representation

There are, of course, many possible reticle patterns. This section introduces four patterns that are commonly used. The motions here are motions in the direction of one coordinate. To derive the Fourier series expansion of $r(\mathbf{x},t)$, one first finds the Fourier series expansion of the reticle pattern with respect to the coordinate involved in the motion.

The first pattern considered is the translating bar reticle (Fig. 9.9). The alternately opaque and transparent bars of dimension σ are translated in the negative x_2 direction. There is no pattern variation in the x_1 direction:

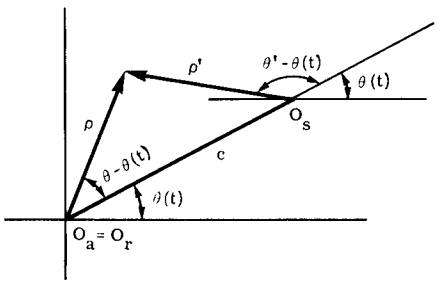


Fig. 9.8 Coordinate transforms for nutating scene.

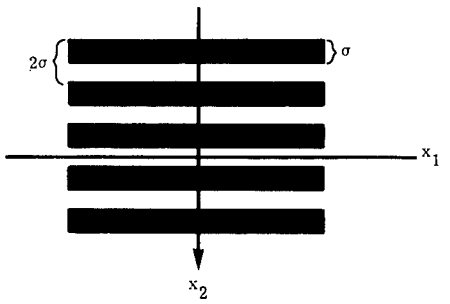


Fig. 9.9 Translating bar reticle.

$$r'(x_1',x_2') = r'(0,x_2') \ . \tag{9.52}$$

The pattern periodicity in x_2' is now exploited:

$$\begin{aligned} r'(x_1',x_2') &= \sum_{k=\infty}^{\infty} a_m \exp(im\pi x_2'/\sigma) \ , \\ a_m &= \frac{1}{2\sigma}\int_0^{2\sigma} r_c'(x_1',x_2') \exp(-im\pi x_2'/\sigma)\, dx_2' \ . \end{aligned} \tag{9.53}$$

Evaluation of a_m requires the introduction of an arbitrary phase δ, illustrated in Fig. 9.10. The evaluation of a_m shows that $a_m = 0$ for even m, and one finds that

$$\begin{aligned} a_0 &= \tfrac{1}{2} \ , \\ a_{(2k+1)} &= \frac{\exp[-i(2k+1)\pi\delta]}{i\pi(2k+1)} \ , \quad k = \pm 1, \pm 2, \ldots \\ &= 0, \text{ otherwise} \ . \end{aligned} \tag{9.54}$$

Substituting Eq. (9.54) in Eq. (9.53) and using Eq. (9.45), one obtains

$$\begin{aligned} r(x_1, x_2 + st) = \frac{1}{2} + \frac{1}{i\pi}\sum_{k=-\infty}^{\infty} &\frac{\exp[-i(2k+1)\pi\delta}{2k+1} \\ &\times \exp[i(2k+1)\pi x_2/\sigma] \exp[i(2k+1)\pi st/\sigma] \ . \end{aligned} \tag{9.55}$$

From Eqs. (9.40) and (9.41)

$$\begin{aligned} \beta_0(t) &= \frac{1}{2}\int_A s(\mathbf{x},t)\, d\mathbf{x} \ , \\ \beta_{2k+1}(t) &= \frac{1}{i\pi}\int_A \frac{\exp[-i(2k+1)\pi\delta]}{2k+1} \exp[i(2k+1)\pi x_2/\sigma] s(\mathbf{x},t)\, d\mathbf{x} \ . \end{aligned} \tag{9.56}$$

The second pattern, the radial, uniformly rotating reticle known as the wagonwheel or episcotister, is commonly employed. On occasion, the basic pattern is modified but not in important ways. Figure 9.11 illustrates the pattern. The reticle's radial property simplifies the reticle description

$$r'(\rho',\theta',t) = r(0,\theta',t) \ . \tag{9.57}$$

If there are p transparent and opaque spoke pairs, the basic period of θ' is $2\pi/p$. The 1-D Fourier series expansion of $r(0,\theta',t)$ follows the same analytic pattern as the bar reticle with the same definition of the arbitrary phase δ:

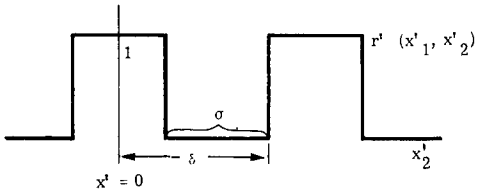


Fig. 9.10 Bar reticle phase.

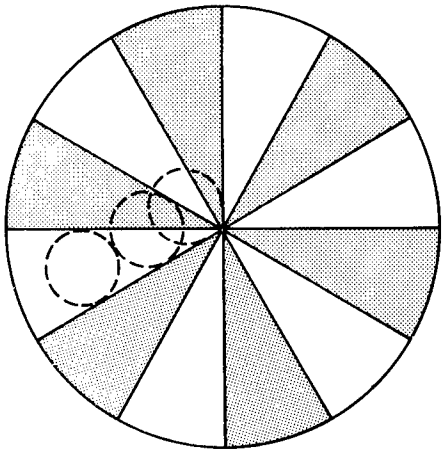


Fig. 9.11 Episcotister pattern.

$$r'(0,\theta') = \sum_{k=-\infty}^{\infty} a_m \exp(im\theta' p) \ ,$$
$$a_m = \frac{p}{2\pi}\int_0^{2\pi/p} r'(0,\theta') \exp(-im\theta' p)\, d\theta' \ . \tag{9.58}$$

After the phase shift of $\exp(im\pi\delta)$, $r'(0,\theta') = 1$, $(0 \leq \theta' \leq \pi/p)$, and 0 for the rest of the period. Thus, one finds $a_m = 0$ for even m and one has

$$a_0 = \frac{1}{2} \ ,$$
$$a_{(2k+1)} = \frac{\exp[-i(2k+1)\pi\delta]}{i\pi(2k+1)} \ , \quad k = \pm 1, \pm 2, \ldots \ . \tag{9.59}$$

Substituting Eq. (9.58) into Eq. (9.57) and using the rotational definition of Eq. (9.49), one has

$$r(\rho', \omega_0 t - \theta') = \frac{1}{2} + \frac{1}{i\pi}\sum_{k=\infty}^{\infty} \frac{\exp[-i(2k+1)\pi\delta}{2k+1} \times \exp[-i(2k+1)p\theta'] \exp[i(2k+1)p\omega_0 t] \ . \tag{9.60}$$

Then from Eqs. (9.40), (9.41), and the definition of $\exp(-i\theta')$ from Eq. (9.48), one obtains

$$\beta_0(t) = \frac{1}{2}\int_0^{2\pi}\int_0^a \rho s(\rho,\theta,t)\, d\rho\, d\theta \ ,$$

$$\beta_{2k+1}(t) = \frac{\exp[-i(2k+1)\pi\delta]}{i\pi(2k+1)}\int_0^{2\pi}\int_0^a \times \frac{\rho[\rho\exp(-i\theta) - b]^{(2k+1)p}}{(\rho^2 + b^2 - 2b\rho\cos\theta)^{(2k+1)p/2}} s(\rho,\theta,t)\, d\rho\, d\theta \ . \tag{9.61}$$

The carrier frequency is $p\omega_0$ and clearly amplitude modulation will occur.

The third pattern considered is the sun-burst or rising-sun reticle. This reticle adds a phasing sector to the episcotister (Fig. 9.12) to obtain angular-positional information. The phasing sector has a transmission of 1/2. The 1-D pulse train is shown in Fig. 9.13. The period of this reticle is 2π and one has

$$r(\rho', \omega_0 t - \theta') = \sum a_m \exp(-im\theta')\exp(im\omega_0 t)$$

$$a_0 = \frac{1}{2}$$

$$a_m = -\frac{\exp(-im\pi\delta)}{im2\pi}\left(\sum_{k=0}^{p-1}\{\exp[-im\pi(2k+1)/2p] - \exp(-im\pi k)\} + \frac{1}{2}[\exp(-2im\pi) - \exp(-im\pi)]\right) \ . \tag{9.62}$$

The last term, (½) $[\exp(-2im\pi) - \exp(-im\pi)]$, is 0 if m is even, and 1 if m is odd. Intuition suggests considerable modulation at $2p\omega_0$ and its harmonics

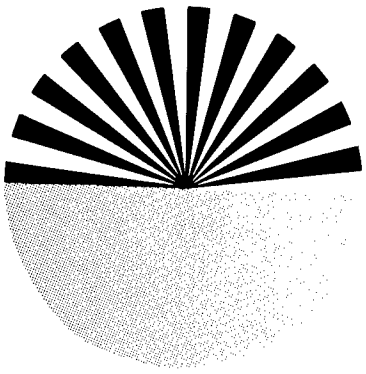


Fig. 9.12 Episcotister with phasing sector.

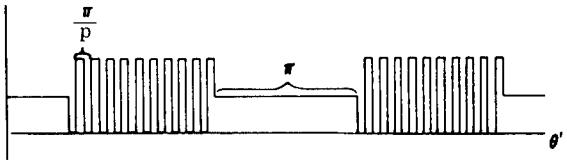


Fig. 9.13 Annular pattern of sectioned episcotister.

$2np\omega_0$. Examination of Eq. (9.62) shows $a_m = 0$ for n even, and a_m a maximum at $n = 1$, which is $\exp(-2p\pi\delta)/i2\pi$. One finds that

$$\begin{aligned}\beta_m(t) &= \frac{1}{2}\int_0^{2\pi}\int_0^a \rho s(\rho,\theta,t)\, d\rho\, d\theta\ ,\\ &= \frac{\exp(-im\pi\delta)}{im2\pi}\int_0^{2\pi}\int_0^a \frac{\rho[\rho\exp(-i\theta) - b]^m}{(\rho^2 + b^2 - 2\rho b\cos\theta)^{m/2}}\\ &\quad \times s(\rho,\theta,t)\, d\rho\, d\theta\ .\end{aligned} \tag{9.63}$$

Again, this is amplitude modulation.

The last reticle pattern considered is more general than the preceding three. It is presented both as a potential reticle system and as a method for approximating more complex patterns. The basic motion is a reticle pattern that is a sum of concentric ring reticles. Figure 9.14 illustrates such a reticle pattern.

The reticle description is

$$r(\rho,\theta,t) = \sum_{n=1}^{N} r_n(\rho,\theta,t)\ , \tag{9.64}$$

where $r_n(\rho,\theta,t)$ has p_n equal transparent and opaque spokes in the ring $\rho_{n-1} < \rho < \rho_n$. It is assumed that the reticle is uniformly rotated with the center of the aperture as the center of rotation. The values of δ_n may differ, but they are equal for Fig. 9.14. The expression for $v_n(t)$ is

$$\begin{aligned}v_n(t) &= \int_0^{2\pi}\int_{\rho_{n-1}}^{\rho_n} \rho r_n(\rho,\theta,t)s(\rho,\theta,t)\, d\rho\, d\theta\\ &= \beta_0^{(n)}(t) + \sum_{k=-\infty}^{\infty}\beta_{2k+1}^{(n)}\exp[i(2k+1)p_n\omega_0 t]\ ,\end{aligned} \tag{9.65}$$

where

$$\begin{aligned}\beta_0^{(n)}(t) &= (1/2)\textstyle\int_0^{2\pi}\int_{\rho_{n-1}}^{\rho_n}\rho s(\rho,\theta,t)\, d\rho\, d\theta\ ,\\ \beta_{(2k+1)}^{(n)} &= \{\exp[-i(2k+1)\pi\delta_n]/i\pi(2k+1)\}\textstyle\int_0^{2\pi}\int_{\rho_{n-1}}^{\rho_n}\rho\\ &\quad \times \exp[-i(2k+1)p_n\theta]\, s(\rho,\theta,t)\, d\rho\, d\theta\ .\end{aligned}$$

Then $v(t)$ is the sum of $v_n(t)$:

$$v(t) = \sum_{1}^{N} v_n(t)\ . \tag{9.66}$$

The analysis of a pattern such as that shown in Fig. 9.15 would be tedious. Clearly, if N is made large enough, the pattern's signal modulation $v(t)$ can be approximated as closely as desired.

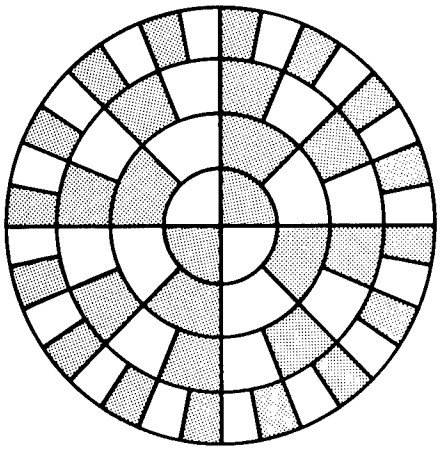


Fig. 9.14 Rotating concentric annular ring reticle.

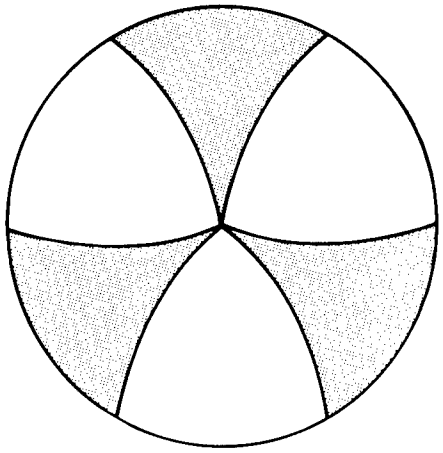


Fig. 9.15 Complex reticle pattern that is approximately annular.

9.4.5 Modulation of Point Sources

The delta function representation for a point source scene $I\delta(\mathbf{x} - \mathbf{x}_0)$ is a convenient total for the analysis of modulation characteristics of reticle systems:

$$I\int_A a_m(\mathbf{x})\delta(\mathbf{x} - \mathbf{x}_0)\, d\mathbf{x} = \begin{cases} I\, a_m(\mathbf{x}_0) , & \text{for } \mathbf{x}_0 = A , \\ 0 , & \text{otherwise} , \end{cases} \tag{9.67}$$

where I is the target intensity. The delta function representation for a point source in polar coordinates is $I\delta(\rho - \rho_T, \theta - \theta_T)$. For the episcotister, the modulation is

$$v(t) = \frac{I(\rho_T)}{2} + I(\rho_T) \sum_{k=-\infty}^{\infty} \frac{\exp[-i(2k + 1)\pi\delta]}{i\pi(2k + 1)} \times \frac{\rho_T[\rho_T \exp(-i\theta_T) - b]^{(2k+1)p}}{(\rho_T^2 + b^2 - 2b\rho_T \cos\theta_T)} \exp[i(2k + 1)p\omega_0 t] . \tag{9.68}$$

This is an amplitude modulation of the carrier $\rho\omega_0$. The modulation of I occurs as in Fig. 9.11 because the target blur-disk is not completely modulated by the reticle.

A point source in a nutated scene has the following representation:

$$s\rho(\rho',\theta')s(\rho',\theta') = S[\rho(\rho',\theta') - \rho'_T,\theta(\rho',\theta') - \theta'_T] .$$

The nutated scene with the centered reticle has, with $b = 0$ in Eq. (9.61),

$$\beta_0(t) = \frac{I}{2}[c^2 + \rho_T'^2 - 2c\rho_T' \cos(\theta_T - \Omega_0 t)]^{1/2} ,$$

$$\beta_{2k+1}(t) = \frac{I \exp[-i(2k+1)\pi\delta]}{i\pi(2k+1)} \exp(-\Omega_0 t) \tag{9.69}$$
$$\times \{\rho_T' \exp[-i(\theta_T - \Omega_0 t)] - c\}^{(2k+1)p} .$$

The modulation clearly depends on the relative size of the error ρ_T' and the radius of nutation c. For $\rho_T' = c$, amplitude modulation occurs. For small values of the ratio ρ_T'/c, one may approximate Eq. (9.51) by

$$\rho \sim c , \tag{9.70}$$
$$\theta \sim \frac{\rho_T'}{c} \sin(\theta' - \Omega_0 t) + \Omega_0 t .$$

Using the approximation (9.70) instead of Eq. (9.51) in Eq. (9.61), with $b = 0$, one has, for a stationary reticle, i.e., $\omega_0 = 0$,

$$\beta_0(t) = \frac{Ic}{2} ,$$

$$\beta_{2k+1}(t) = \frac{I \exp[-i(2k+1)\pi\delta]}{i\pi(2k+1)} \exp[-i(2k+1)p] \tag{9.71}$$
$$\times \left[\frac{\rho_T'}{c} \sin(\theta_T - \Omega_0 t) + \Omega_0 t\right] ,$$
$$v(t) = \frac{Ic}{2} + \frac{I}{i\pi} \sum_{k=-\infty}^{\infty} \frac{\exp[-i(2k+1)\pi\delta]}{2k+1}$$
$$\times \exp[i(2k+1)p] \left[\frac{\rho_T'}{c} \sin(\theta_T - \Omega_0 t) + \Omega_0 t\right] .$$

Clearly this is frequency modulation (FM). The carrier frequency is $\rho\Omega_0$ where ρ is the number of spokes and Ω_0, the nutation frequency, becomes the modulation frequency. The target's radial position ρ_T' is proportional to the amplitude (modulation index) of the modulating frequency Ω_0. The amplitude is $p/c\ \rho_T'$. The phase of the modulation is determined by θ_T.

Another FM reticle system can be constructed from the translating bar-reticle and rotation of the scene with the center of the scene's rotation, the aperture center. Then

$$s(\rho,\theta,t) = s(\rho, \theta + \Omega_0 t) . \tag{9.72}$$

Figure 9.16 represents such a system. Changing to polar coordinates in Eq. (9.56) and introducing the delta function representation for the scene, one obtains

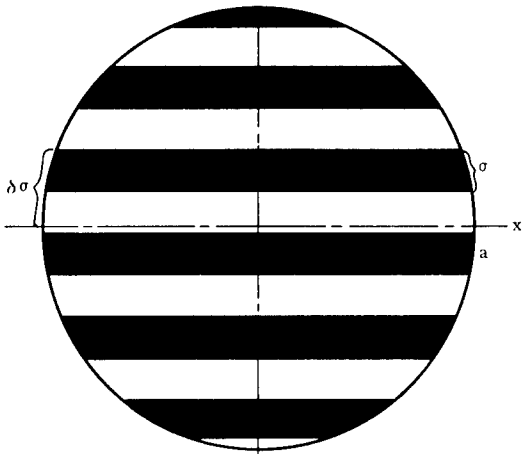


Fig. 9.16 Translating bar-reticle circular aperture and rotating scene at aperture center.

$$s(\rho, \theta - \Omega_0 t) = \delta(\rho - \rho_T, \theta - \theta_T - \Omega_0 t) \ . \tag{9.73}$$

The modulation of the point source is, by substitution into Eq. (9.56),

$$\beta_0(t) = \frac{I}{2} \ ,$$

$$\beta_{2k+1}(t) = \frac{I}{i\pi} \frac{\exp[-i(2k + 1)\pi\delta]}{2k + 1} \exp[i(2k + 1)\pi\rho_T \sin(\theta_T - \Omega_0 t)] \ ,$$

then

$$v(t) = \frac{I}{2} + \frac{I}{i\pi} \sum_{k=-\infty}^{\infty} \frac{\exp[-i(2k + 1)\pi\delta]}{2k + 1} \exp[i(2k + 1)\pi] \times \left[\frac{\rho_T}{\sigma} \sin(\theta_T - \Omega_0 t) + st/\sigma\right] \ , \tag{9.74}$$

where

$\pi s/\sigma$ = carrier frequency
$\Omega_0 t$ = modulation frequency
$\pi\rho_T/\sigma$ = amplitude of modulation (carries information about ρ_T)
ρ_T = radial error
θ_T = angular error.

The phase of the modulated signal carries the information about θ_T.

9.4.6 Reticles Coding Spatial Frequencies in Temporal Frequencies

The analytic tools developed so far are discussed toward the analysis of the modulation characteristics of a reticle system. This section covers some reticle system synthesis results.

Doubly Periodic Reticles. This method of reticle synthesis was proposed by Ulrich, Montgomery, and Alward.[5] These reticles are fairly simple to design. The pattern of openings in the reticle mask is assumed to have 2-D periodicity. (For the definition of 2-D periodicity, see Sec. 9.2.3.) The aperture is considered arbitrarily large. This reticle system may be regarded as a device that codes linear combinations of certain spatial frequencies of the scene into the time frequencies of the voltage output.

Let the infinite reticle mask have a periodicity defined by $\mathbf{a}_1$ and $\mathbf{a}_2$. The reticle pattern may be represented by a series, as in Sec. 9.2.3:

$$r(\mathbf{x}) = \sum_n r_n \exp(2\pi i \mathbf{x} \cdot \mathbf{b}_n) \ . \tag{9.75}$$

The output voltage $v(t)$, when the reticle is translated, may be represented by

$$v(t) = \int_{-\infty}^{\infty} s(\mathbf{x}) r[\mathbf{x} - \mathbf{x}(t)] \, d\mathbf{x} \ . \tag{9.76}$$

Substituting Eq. (9.75) into Eq. (9.76) and using the fact that $r(\mathbf{x})$ is real, and consequently that the complex conjugate of r_n, r_n^*, is r_{-n}, one has

$$v(t) = \sum_n r_n^* S(\mathbf{b}_n) \exp[2\pi i \mathbf{b}_n \cdot \mathbf{x}(t)] \ , \tag{9.77}$$

where $S(\mathbf{b}_n)$ is the Fourier transform of $s(\mathbf{x})$ evaluated at the lattice point $\mathbf{b}_n$. For the signal to be periodic with period T, $\mathbf{x}(t)$ must equal $t/T\mathbf{a}_k$ and $\mathbf{a}_k$ must be associated with only one primary lattice point. Using the $\mathbf{x}(t)$ defined above, and the definition of the primary and reciprocal lattice, one has

$$\exp[2\pi i \mathbf{b}_n \cdot \mathbf{x}(t)] = \exp[2\pi i \mathbf{k} \cdot \mathbf{n}(t/T)] \ , \tag{9.78}$$

where

$$\mathbf{k} = (k_1, k_2)$$
$$\mathbf{n} = (n_1, n_2) \ .$$

Using the periodicity of $v(t)$, the series expansion may be written for $v(t)$ as

$$v(t) = \sum_m V_m \exp[2\pi i m(t/T)] \ . \tag{9.79}$$

Equating Eqs. (9.79) and (9.77) after substituting Eq. (9.78) in Eq. (9.77), one obtains

$$\sum_m V_m \exp[2\pi i m(t/T)] = \sum_n r_n^* S(\mathbf{b}_n) \exp[2\pi i \mathbf{n} \cdot \mathbf{k}(t/T)] \ . \tag{9.80}$$

Equating coefficients, the m'th harmonic of $v(t)$ is expressed as

$$V_m = \sum_{n \cdot k = m} r_n^* S(\mathbf{b}_n) \ . \tag{9.81}$$

Equation (9.81) shows that the m'th harmonic of the output is a linear combination of the spatial frequencies along a line ($\mathbf{n} \cdot \mathbf{k} = m$) perpendicular to the scan direction $\mathbf{a}_k$. Since $\mathbf{k}$ has been chosen so that k_1 and k_2 are relatively prime, there will be at least one lattice point on every line $\mathbf{n} \cdot \mathbf{k} = m$. If only a finite number of spatial frequencies are required to distinguish between targets and their backgrounds, then the coefficients r_n are set to zero on the remaining reciprocal lattice points. Fig. 9.17 represents the discrimination set in reciprocal space. The coefficients r_n are set equal to zero for all lattice points not in S. Choose a scan direction $\mathbf{a}_k$ so that for each pair of lattice points in S, $\mathbf{a}_k$ is not perpendicular to a line joining them:

$$(\mathbf{n} - \mathbf{m}) \cdot \mathbf{k} \neq 0, \quad \text{for all } \mathbf{n}, \mathbf{m} \text{ in } S \ . \tag{9.82}$$

Then, since each harmonic v_m of the output voltage is the sum of spatial frequencies perpendicular to the scan direction, the sum of the right side of Eq. (9.81) will contain only one term:

$$V_m = r_n^* S(\mathbf{b}_n) \ , \tag{9.83}$$

where $\mathbf{n} \cdot \mathbf{k} = m$. Discrimination is achieved by analyzing the output signal $v(t)$ for the harmonics m/T. Some combination of harmonic amplitudes will indicate a target while other, hopefully distinct combinations, simply a background.

Consider a reticle that will separate a point source from a straight edge. A point source will have equal spatial frequency content in all directions. The spatial frequency content of a straight edge is maximum in the direction perpendicular to the edge and zero in the direction parallel to the edge. Our discrimination set contains only four, perpendicular, reciprocal lattice points (see Fig. 9.18). Let the temporal period of the scan be $2T$ seconds. Then the

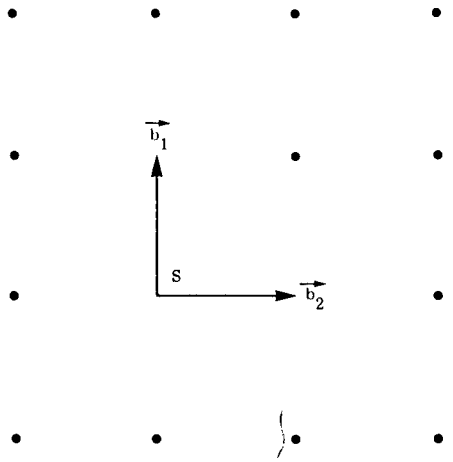


Fig. 9.17 Discrimination set in reciprocal space.

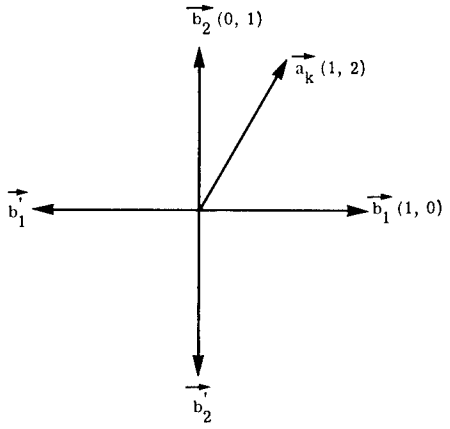


Fig. 9.18 Discrimination set and scan direction.

spatial frequency pair $\mathbf{b}_1, -\mathbf{b}_1$ is encoded in temporal frequency $1/2T$, since $m = \mathbf{n} \cdot \mathbf{k} = (1,0)(1,2) = 1$; and the spatial frequency pair $\mathbf{b}_2, -\mathbf{b}_2$ is encoded in temporal frequency $1/T$, since $m = \mathbf{n} \cdot \mathbf{k} = (0,1)(1,2) = 2$. This includes conjugate pairs and requires $r_n^* = r_{-n}$, real values for $r(x)$. The reticle mask $r(\mathbf{x})$ is given by

$$r(\mathbf{x}) = \frac{1}{2}\exp(2\pi i x_1) + \frac{1}{2}\exp(2\pi i x_2) \ . \tag{9.84}$$

The two sinusoidal patterns could be approximated by the superposition of two bar patterns. Figure 9.19 shows the resulting reticle pattern and scan direction. The dark areas have zero transmission, the lightly shaded areas have transmission 1/2, and the unshaded areas have transmission 1. Suppose a point source is scanned by the reticle, making a path $\mathbf{a}_k$. The output signal depicted in Fig. 9.20(a) is the sum of the signals depicted in Fig. 9.20(b) and (c). If the point is shifted, the relative phase of (b) and (c) will change, but the output signal will remain the same. Discrimination of the symmetric source from the straight edge is achieved by comparing the amplitude of frequency $1/T$ with the amplitude of frequency $1/2T$. A symmetric source exists if the two amplitudes are equal. Such an experimental reticle has been constructed and works as predicted.[6]

9.4.7 Coded Imaging Reticles

The output signal $v(t)$ of the reticles considered so far cannot be used to recreate an image of the scene from $v(t)$. There do exist reticle codes that permit recovery of the image. The simplest and most familiar is the Nipkow scanner, the first TV camera. Such a system produced N raster lines with M resolution elements in each line as illustrated in Fig. 9.21. The reticle contains N transparent apertures. The first aperture scans the image and produces a signal $v_1(t)$ representing the first line. The second aperture scans the second line until finally the frame is complete, with the scanning of the N'th aperture producing $v_N(t)$.

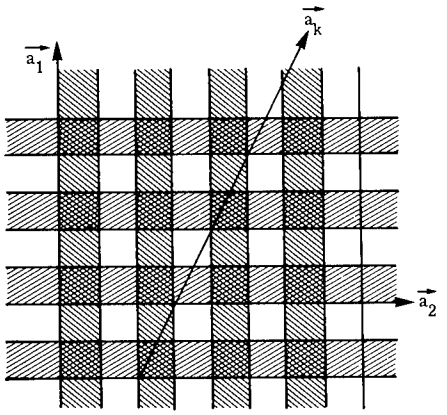


Fig. 9.19 Bar approximation to reticle pattern Eq. (9.85).

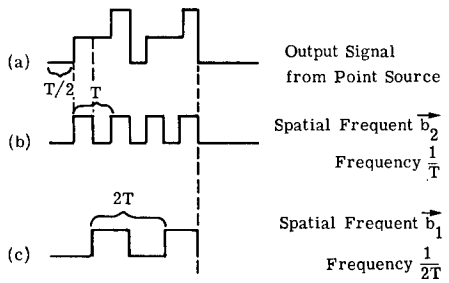


Fig. 9.20 Output form point source.

The reticle was implemented on a rotating disk. To analyze such reticle codes an analytic model is required.

So far both the scene and reticle have been treated as continuous entities. A discrete model is more appropriate for coded reticle analysis. The scene representation is presented in Fig. 9.22. The intensity s_{ij} represents the signal from the i'th resolution element from the j'th line. The representation of the scene is in column vector. The superscript T denotes the transpose of the row vector:

$$\mathbf{S} = (S_{11},...,S_{1M},...,S_{i1},...,S_{ij},...,S_{iM},...,S_{N1},...,S_{NM})^T \ . \tag{9.85}$$

The reticle pattern is depicted in Fig. 9.23. The reticle pattern is superimposed on the scene. As the reticle is moved to the left, a new block of the reticle is superimposed on the scene. The k'th reticle block superimposes on the scene the reticle code given by

$$\mathbf{r}_k = (\mathbf{r}_{1k},...,\mathbf{r}_{1,k+M-1},\mathbf{r}_{2k},...,\mathbf{r}_{2,k+M-1},...,\mathbf{r}_{N,k},...,\mathbf{r}_{N,k+M-1}) \ . \tag{9.86}$$

The voltage v_k observed when the k'th reticle block is superimposed on the scene is simply the inner or dot product:

$$v_k = \mathbf{s}^T \cdot \mathbf{r}_k, \quad k = 1, ..., l - M + 1 \ . \tag{9.87}$$

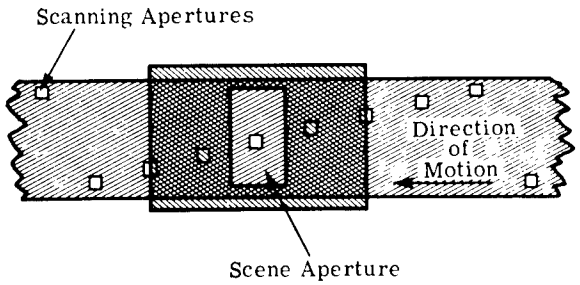


Fig. 9.21 Idealized Nipkow reticle.

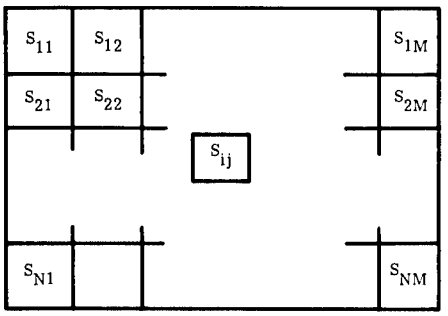


Fig. 9.22 Discrete scene representation.

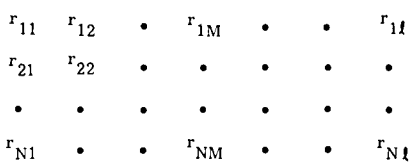


Fig. 9.23 Physical reticle pattern.

The number of v_ks produced by a reticle of length l is $l - M + 1$. The definition of inner product Eq. (9.7) shows that Eq. (9.87) is a set of $l - M + 1$ linear equations for the unknown NMs_{ij}. For a unique solution to exist there must be NM equations. Thus, the reticle length l is

$$l = M(N + 1) - 1 \ . \tag{9.88}$$

If the $M(N + 1) - 1$ vectors $\mathbf{r}_k$ are given, the reticle pattern can be constructed using the definition of $\mathbf{r}_k$ given in Eq. (9.86) and Fig. 9.23. Henceforth, attention will be directed toward defining the reticle in terms of the vector $\mathbf{r}_k$.

The most compact method for defining the coded imaging reticle is in matrix notation. The reticle matrix $R = (r_{ij})$ has as its k'th row the vector $\mathbf{r}_k$ with $k = 1, ..., NM$. One should understand at least the rules for matrix multiplication, addition, identity matrix, and the definition of a matrice's inverse. Reference 7 (or any elementary text on linear algebra) contains this information. The output vector from a coded reticle in matrix notation is

$$\mathbf{v} = \mathbf{s}R \ . \tag{9.89}$$

The encoded signal $\mathbf{v}$ must be decoded to recover the scene $\mathbf{s}$. The decoding transform is the inverse of the matrix R denoted by R^{-1}. Then, operating on $\mathbf{v}$ with R^{-1}, the scene is recovered:

$$\mathbf{v}R^{-1} = \mathbf{s}RR^{-1} = \mathbf{s}I = \mathbf{s} \ , \tag{9.90}$$

where I is the identity.

There are important restrictions on the matrices R that can be implemented as reticles. The rows of R, $\mathbf{r}_k$, are representations of *overlapping* reticle blocks [see Eq. (9.86) and Fig. 9.23]. The key restriction is overlapping. This leads to a restriction on the values $(\mathbf{r}_{ij})$ of the matrix R:

$$\mathbf{r}_{ij} = \mathbf{r}_{i-1,j+1} \ . \tag{9.91}$$

A matrix R with the property shown in Eq. (9.91) is called a *circulant.* A circulant matrix is represented in Fig. 9.24. The circulant requirement for R is a direct and inescapable consequence of the reticle implementation. The circulant has the diagonal elements from the lower left to upper right equal. For example, inspection shows the matrix R associated with the Nipkow scanner has elements in the diagonal $r_{k,NM-k+1}$ (see Fig. 9.25).[a] The mathematical development of coded reticles may be found in Ref. 8.

A further restriction on the coded reticle is that the transmission r_{ij} is such that $0 \leq r_{ij} \leq 1$. In fact, the easiest reticles to construct are those that set r_{ij}

[a]The critical reader will note that while every circulant matrix can be implemented in a reticle, not every matrix associated with a reticle is a strict circulant. The circulant restriction must hold for submatrices of R. The less restrictive condition on R has, to this author's knowledge, not been studied. On the other hand, the literature on circulants is rich and the electro-optical engineer may draw on a large mathematical literature. Coded reticles that are not full circulants do not appear in the literature.

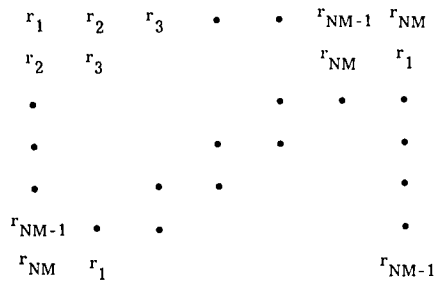


Fig. 9.24 Circulant reticle matrix.

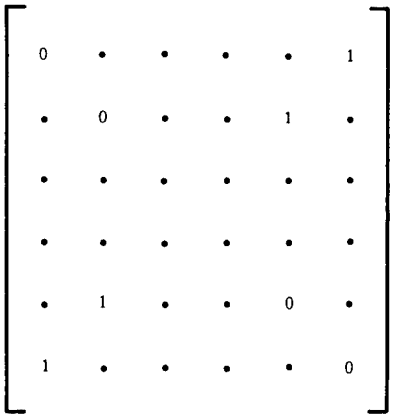


Fig. 9.25 Reticle matrix for Nipkow.

$= 0$ or 1, i.e., either opaque or transparent. Letting $V = NM$, one has a class of widely studied matrices, the (V, k, λ) configuration. (See Chapter 8 of Ref. 9.) These configurations have been studied mathematically[9] and have found application in statistical design of experiments.[10] The matrices of zeroes and ones associated with (V, k, λ) have the following important properties:

1. R is of order $V \quad 0 < \lambda < k < V - 1$
2. $\mathbf{r}_i \cdot \mathbf{r}_i = k \quad i = 1, ..., V,$ where $\mathbf{r}_i$ is the i'th row of R
3. $\mathbf{r}_i \cdot \mathbf{r}_j = \lambda \quad i = j$
4. $RR^T = R^R R = (k - \lambda)I + \lambda J,$ where J is the $V \times V$ matrix of ones
5. $$R^{-1} = \frac{1}{(k - \lambda)} R - \frac{\lambda}{k} J \ . \tag{9.92}$$

Relation (9.90) says that each row of R contains k ones. The Nipkow scanner has $k = 1$, $\lambda = 0$, and $R^{-1} = R$, which is easily checked. Of great importance is the fact that the matrix R associated with a (V, k, λ) configuration has an inverse R^{-1}.

Not all combinations of integers (V, k, λ) form a (V, k, λ) configuration. If there exists a (V, k, λ) configuration, the V, k, and λ must satisfy

1. $\lambda = k(k - 1)/V - 1$
2. If V is even, then $(k - \lambda)$ is a square
3. If V is odd, then the equation

$$x^2 = (k - \lambda)y^2 + (-1)^{(V-1)/2}\lambda z^2 \tag{9.93}$$

has a solution in integers x, y, and z not equaling zero.

The coded reticles used in the current literature are called Hadamard codes.[11,12] It can be shown that such Hadamard codes are a special case of (V,k,λ) configurations. A Hadamard matrix has only ones and minus ones as entries. They can be normalized to having only ones in the first row and column of the matrix. The Hadamard of order n matrix satisfies the relation

$$HH^T = nI \ . \tag{9.94}$$

A Hadamard reticle code matrix is obtained by dropping the first row and column and changing the ones to zeros. Hadamard matrices of order $n = 2^m$ are easily constructed.[13,14] In fact, all known Hadamard matrices have order divisible by 4, i.e., $n \equiv 0 \bmod 4$. An unproven conjecture has *all* Hadamard matrices of order $n \equiv 0 \bmod 4$. It can be proven that Hadamard matrices of order $n = 4t$ lead to Hadamard reticle codes that are equivalent to (V,k,λ) configurations where

$$\begin{aligned} V &= 4t - 1 \qquad & RR^T &= tI + (t-1)J \\ k &= 2t - 1 \qquad & R^{-1} &= \frac{1}{t}R - \frac{t-1}{2t-1}J \\ \lambda &= t - 1 \ . \end{aligned} \tag{9.95}$$

The relation between Hadamard codes and maximal length sequences (*m*-sequences) is discussed in Ref. 13.

So far, the discussion has not introduced circulant (V, k, λ) configurations. Such circulants do exist (see Chapter 9 of Ref. 9). Circulant (V, k, λ) configurations are derived from perfect different set[15] or cyclic projective planes.[16] If $4t - 1$ is a prime, then a circulant Hadamard reticle code can be found by the method of quadratic residues. A detailed discussion of methods for constructing such circulant reticle codes lies outside the scope of this chapter. References 14 and 15 plus some computational skills enable the reader to construct circulant coded reticles. The reader should note that the Hadamard configurations are a subset of the more general (V, k, λ) configurations that the reticle designer has at his disposal.

Attention must now be paid to decoding the encoded scene $\mathbf{v}$. The quantity $\mathbf{v}$ may be recorded digitally and the inverse R^{-1} implemented by a general or special purpose computer. Equations (9.92) and (9.95) show that R^{-1} is also a circulant. Since R^{-1} is a circulant, the decoding operation can be implemented by shift registers. The encoded image $\mathbf{v}$ must be stored. The price and size of digital storage and shift registers today warrant considering the construction of special purpose decoders.

An early method of decoding[8] was to make an encoded photographic image of $\mathbf{v}$ and then to decode the image with a decoding reticle. The natural decoding reticle is R^T. This decoding reticle produced a signal pedestal $\lambda \mathbf{S}J$, Eqs. (9.92) and (9.95), which had to be reduced by a bias or dc correction.

One should ask what advantages can be gained from a coded reticle as compared with a single small detector scanned in the object or image plane. The answer that can be given is equivocal. The coded reticle increases the dwell time on the scene element by a factor k, the number of ones in the reticle. If an imaging system's limiting noise is photon noise from the housing, or preamplifier noise, then there is a gain in signal to noise of $\sqrt{k}$. Again, one would maximize k. The coded reticle shows no improvement for photon noise arising from the scene. If the system is detector noise limited, and the *f*-numbers of the scanned element and coded reticle system are equal, then the

Table 9.2 Signal-to-Noise Gains from Coded Reticles

Gain in S/N	Noise Source
1	Scene photon noise
$\sqrt{k}$	Preamplifier, or housing photon noise
$\sqrt{k/V}$	Detector noise

detector area must increase by a factor V. In this case, the signal-to-noise gain is $\sqrt{k/V}$, which is always less than one (a loss). These results are summarized in Table 9.2.

The factors considered so far lead to maximizing k. The configuration that maximizes k sets $k = N - 1$. This trivial circulant makes the transparent portions of the Nipkow reticle opaque, the remainder of the reticle transparent in Fig. 9.21. This would be a practical solution except that the modulation amplitude is very small compared to the average signal. Electronic circuits do not handle such a signal well. A compromise solution is to set $k \approx 1/2V$, a reticle code implemented by a Hadamard configuration. The designer must make his own compromise.

References

1. F. Oberhettinger, *Fourier Expansions: A Collection of Formulas,* Academic Press, New York (1973).
2. I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals Series and Products,* 4th ed., Academic Press, Washington, DC (1965).
3. R. C. Jones, "Theoretical studies of infrared target detection problems," AFCRL Report No. 1373, Polaroid Corporation, Cambridge, MA (July 1963).
4. L. M. Biberman, *Reticles in Electrooptical Devices,* Pergamon Press, Elmsford, NY (1966).
5. J. P. Ulrich, W. D. Montgomery, and J. L. Alward, "Analysis of reticle system," Report No. 6054-2-T, Background Analysis Center, Willow Run Laboratories, University of Michigan, Ann Arbor, MI (Oct. 1965).
6. S. Sternberg, J. Ulrich, and R. Hamilton, "Analysis and testing of a special rotating-translating reticle," Report No. 7102-1-T, Willow Run Laboratories, University of Michigan, Ann Arbor, MI (June 1967).
7. G. Birkhoff and S. MacLane, *A Survey of Modern Algebra,* Macmillan, New York (1941).
8. S. Sternberg, "Designing reticles for imaging systems by use of linear algebra," ARO-D Report 67-1, Transactions of the Twelfth Conference of Army Mathematicians, U.S. Army Research Office, Durham, NC (Feb. 1976).
9. H. J. Ryser, *Combinatorial Mathematics,* Carus Mathematical Monographs No. 14, John Wiley & Sons, New York (1963).
10. R. C. Bose, "On the construction of balanced incomplete block designs," *Annals of Eugenics,* Vol. 9, pp. 353–399, Cambridge University Press, New York (1939).
11. R. D. Swift, R. B. Wattson, J. A. Decker, Jr., R. Paganetti, and M. Harwit, "Hadamard transform imaging and imaging spectrometer," *Applied Optics* **15**(6), 1595–1609 (1976).
12. E. D. Nelson and M. L. Fredman, "Hadamard spectroscopy," *Journal of the Optical Society of America* **60**(12), 1665 (1970).
13. W. W. Peterson, *Error Correcting Codes,* p. 106, MIT Press, Cambridge, MA, and John Wiley & Sons, New York (1961).
14. F. J. McWilliams and N. J. A. Sloane, "Pseudo-random sequences and arrays," *Proceedings of the IEEE* **64**(12), 79–80, 1715–1729 (1976).

15. M. Hall, "A survey of different sets," *Proceedings of the American Mathematical Society* **7**, 975–986 (1956).

16. J. Singer, "A theorem infinite projective geometry and some applications to number theory," *Transactions of the American Mathematical Society* **43**, 377–385 (1938).

Bibliography

Blackwell, D., and M. A. Girshick, *Theory of Games and Statistical Decisions,* John Wiley and Sons, New York (1954).

Davenport, W. D., and W. L. Root, *Random Signals and Noise,* McGraw-Hill, New York (1958).

Lee, Y. W., *Statistical Theory of Communications,* John Wiley and Sons, New York (1960).

Montgomery, W. D., "Some consequences of sampling in FLIR systems," Research Paper P-543, Institute for Defense Analyses, Arlington, VA (September 1969).

Oberhettinger, F., *Fourier Expansions: A Collection of Formulas,* Academic Press, New York (1973).

CHAPTER 10

Lasers

Hugo Weichel
Nichols Research Corporation
Vienna, Virginia

CONTENTS

10.1 INTRODUCTION

To trace the evolution of the ideas behind the discovery of the laser, one must go back to 1917. In that year Einstein proposed the existence of stimulated emission of radiation. The most remarkable aspect of stimulated radiation is that it is emitted in the same direction and phase as the incident radiation, while spontaneous radiation is not related to incident radiation in any way.

Experimental evidence of stimulated emission was, however, difficult to obtain. With ordinary light sources ($T \sim 10^3$ K) the rate of stimulated emission is extremely small in the visible region of the spectrum. The radiation of such sources is primarily due to spontaneous transitions that occur in a random manner. The preponderance of these spontaneous transitions is the reason why light from ordinary sources is incoherent.

In 1940 Fabrikant, while writing his doctoral dissertation, claimed that under certain conditions stimulated emission can lead to the coherent amplification of light. He suggested that for amplification to occur it is necessary that the number of atoms[a] in a higher energy state exceed those in a lower energy state. This situation, which requires the energy distribution among atoms to deviate from thermal equilibrium, is called a population inversion.

A decade later Purcell and Pound obtained experimental evidence of stimulated emission of a 50-kHz signal from the nuclear spin system of lithium fluoride. A population inversion was created by the sudden reversal of an external magnetic field. This achievement led to a search for methods of establishing population inversions between energy levels located far enough apart that transition between them would produce radiation in the microwave region. In the years 1952 to 1954, several papers were published on the amplification of microwaves by stimulated emission of radiation from excited molecules. One such paper by Townes, Gordon, and Zeiger described the construction and operation of a device that they called microwave amplification by stimulated emission of radiation (MASER).

Following the development of the MASER, the idea of applying the same principles to the optical region of the spectrum was proposed by Schawlow and Townes in 1958. The first successful operation of a laser was reported in 1960 by Maiman of the Hughes Aircraft Company, whose device used a ruby crystal. The helium-neon (HeNe) gas laser, proposed in 1959 by Javan, was continuously operated for the first time before the end of 1960.

The immense potential of the laser was quickly recognized, and in subsequent years the device was studied in many laboratories. The intense interest and explosive growth that followed was based partly on the phenomenon of laser action itself (as an interesting example of the collective interaction between atoms and light), but mostly on the many foreseeable uses of laser beams.

10.1.1 Comparison of Laser Beams to Ordinary Light Beams

Laser light is different from ordinary light. It is much more directional, monochromatic, coherent, and intense. The directionality of a laser beam arises

[a]The word *atom* as used here is the generic sense, which includes molecules as well.

from the geometry of the optical resonator. The coherence, chromatic purity, and intensity of a laser beam are because excited atoms are stimulated to radiate light cooperatively rather than spontaneously and independently.

Laser light shows itself to be different from ordinary light even when it merely illuminates a piece of paper. The area illuminated looks grainy and seems to sparkle. The reason lies in the coherence of the laser beam. As the laser light waves are scattered from neighboring points on the paper, they interfere with one another everywhere, producing bright spots where the waves reinforce each other in phase, and leaving dark spots where they destructively interfere. This "speckle" pattern depends on the angle at which the paper is viewed; the pattern changes with a slight movement of the head and the shifting bright spots seem to sparkle. This speckle effect is well known and can in fact be a nuisance in certain laser applications such as imaging laser radars.

The directionality of laser beams is due to the fact that the gain medium is placed in the optical resonator in such a way that only a wave propagating along the resonator axis can be amplified. If this wave has perfect spatial coherence, an arbitrary irradiance distribution, and passes through an aperture of diameter D, the emerging beam will spread by the amount

$$\theta = C\lambda/D \ , \tag{10.1}$$

where λ is the wavelength of the beam, and C is a constant of the order of unity whose exact value depends on the intensity distribution and on the way in which both the beam divergence angle and beam diameter are defined. A beam whose divergence angle is given by Eq. (10.1) is said to be diffraction limited. One important but not often realized goal of laser designers is to construct lasers whose beams are diffraction limited.

If the wave has partial spatial coherence, the resulting beam will have a divergence angle that is larger than the diffraction-limited angle. The beam divergence angle is now given by

$$\theta = C'\lambda/\sqrt{S} \ , \tag{10.2}$$

where C' is a numerical coefficient of the order of unity whose value depends on the way in which the divergence angle θ and the coherence area S are defined. Typical values for beam divergence angles of laser beams are 10^{-3} to 10^{-6} rad. On the other hand, the beam divergence angle of a good searchlight is typically between 10^{-1} to 10^{-2} rad.

Another important characteristic of laser light is its narrow bandwidth or monochromaticity, which is essentially due to two phenomena. One property of stimulated emission of radiation is the addition in phase of the stimulated radiation to the incident radiation. This results in the conservation of frequency. Thus, if the incident radiation is monochromatic, then the stimulated radiation would be likewise. Furthermore, the laser resonator can only support those electromagnetic waves whose frequencies match the resonant frequencies of the resonator. Because these frequencies have a much smaller bandwidth than the atomic transition line, laser beams can have bandwidths that are considerably smaller than the linewidths of spontaneous atomic transitions.

For example, the light of a single spectral line from a low-pressure gas discharge lamp is spread over a band of frequencies that is about 10^9 Hz wide. A beam from a gas laser on the other hand can have a bandwidth of less than 10^4 Hz. This corresponds to a spectral purity ($\Delta f/f_0$, where f_0 is the line center frequency) of 10^{-11} for laser light and 10^{-6} for the most monochromatic light from a low-pressure gas lamp.

The theoretical limit of the bandwidth of a laser beam of power P and resonator bandwidth of Δf_{cav} full width at half maximum (FWHM) is

$$\Delta f_{\text{osc}} = \frac{\pi h f (\Delta f_{\text{cav}})^2}{P} , \tag{10.3}$$

where hf is the photon energy. This result is known as the Schawlow-Townes formula. For a typical HeNe laser ($P = 1$ mW, $\Delta f_{\text{cav}} \simeq 5 \times 10^5$ Hz) the beam's spectral width is $\Delta f_{\text{osc}} = 3 \times 10^{-4}$ Hz. The measurement of such narrow bandwidths is extremely difficult (if not impossible) and is affected by thermal fluctuations and vibrations.

A third major characteristic that distinguishes a laser from ordinary light sources is the high degree of coherence of the laser radiation. The coherence of electromagnetic radiation can be specified by its spatial and temporal coherence. Spatial coherence refers to a definite phase relationship between different points in a cross section of the beam. To illustrate, consider two points P_1 and P_2 that lie on the same cross section, and let $E_1(t)$ and $E_2(t)$ be the corresponding electric fields at these two points. If the difference between the phases of the two fields remains constant at any time $t > 0$, then by definition there exists perfect spatial coherence between these two points. If this situation exists for any two points in the cross section, the beam has perfect spatial coherence. In reality, for any point P_1, the point P_2 lies within some suitably defined coherence area $S(P_1)$, and the beam is said to have partial spatial coherence. For an ordinary light source, S is typically of the order of the area of a pinhole, whereas S_L for a laser beam is of the order of the cross-sectional area of the beam, and $S_L/S \geq 100$.

Temporal coherence can be defined with the aid of a Michelson interferometer such as the one shown in Fig. 10.1. A parallel beam of light is split by a beamsplitter into two beams propagating in different directions. After traversing their individual paths, the two beams are recombined on a screen where they may form interference fringes. The visibility of the fringes is given by

$$V = \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}} , \tag{10.4}$$

where I_{max} and I_{min} are the maximum and minimum observed intensities, respectively. A visibility of one (i.e., $I_{\text{min}} = 0$) is associated with full temporal coherence, whereas a visibility of zero (i.e., $I_{\text{max}} = I_{\text{min}}$) is observed when no temporal coherence exists between the two beams. Partial temporal coherence exists when V is between zero and one.

In practice, interference fringes are observed only when the path differences of the two beams remain within a certain maximum length. For ordinary light sources it may be as much as a few centimeters. On the other hand, for laser light the path difference may be many meters. The maximum path-length

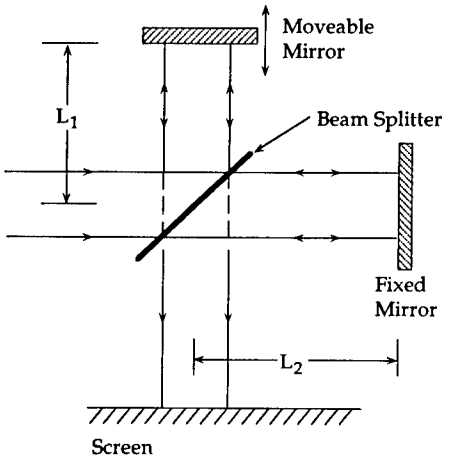


Fig. 10.1 Diagram of the Michelson interferometer. When $2(L_1 - L_2)$ is larger than the coherence length of the optical wave, the fringes begin to disappear.

difference for which fringes are still observed is called the coherence length l_c, and the maximum time delay of the electromagnetic wave corresponding to this path-length difference (l_c/c, where c is the speed of light) is the coherence time t_c. The concept of temporal coherence is, of course, directly related to that of monochromaticity. A continuous-wave (cw) beam with a coherent time t_c has a bandwidth $\Delta f = 1/t_c$. Measured linewidths of some common multifrequency lasers are shown in Table 10.1. These linewidths are observed if no special line-narrowing precautions are taken. As a rule, gas lasers are in general more monochromatic than solid-state lasers.

We conclude the discussion of coherence by observing that the two concepts of temporal and spatial coherence are independent of each other, and by emphasizing that lasers operating in a single longitudinal mode (corresponding to one resonant frequency of the laser resonator) have extremely narrow linewidths. For a single-frequency HeNe laser, linewidths of 50 to 500 Hz have been achieved.[1] Finally, in terms of beam brightness, even a very low power 1-mW HeNe laser produces a spectral irradiance that is orders of magnitude larger than the sun's. For instance, the spectral irradiance of the sun at 0.63 μm

Table 10.1 Linewidths of Several Common Lasers

Laser	Line (μm)	Linewidth (GHz)
Ar	0.448	~5
HeNe	0.6328	~1.5
Ruby	0.6943	~30
Nd:glass	1.06	~1500
Nd:YAG	1.064	~13
CO_2	10.6	~0.1

is 1570 W m^{-2} μm^{-1}. In contrast, the HeNe laser with an output of 1 mW, a beam diameter of 1 mm, and a bandwidth of 2×10^{-6} μm has a spectral irradiance of 600 MW m^{-2} μm^{-1}. Thus, the 1-mW HeNe laser has a spectral irradiance that is nearly a million times larger than that of the sun.

10.1.2 Essential Elements of the Laser

The laser is an optical device that emits an intense, highly collimated beam of nearly monochromatic radiation. The device consists basically of three elements—an external energy source or *pump* that excites an *amplifying medium* placed inside of an *optical resonator*.

The amplifying medium may be a gas, liquid, or solid. It determines the wavelength of the laser radiation. Because of the large selection of amplifying media, the range of available laser wavelengths extends from the UV well into the IR, up to wavelengths that are a sizable fraction of a millimeter. For example, laser action has been observed in more than half of the known elements with more than a thousand laser transitions in gases alone. Two of the most widely used transitions in gases are the 0.6328-μm visible transition of neon and the 10.6-μm IR transition of the carbon dioxide molecule.

Aside from the wavelength of the laser transition, one of the most important features of the amplifying medium is the gain coefficient versus the frequency curve. This bell-shaped curve, which is simply called the gain curve or gain distribution, shows the gain coefficient as a function of frequency. It generally fits a Gaussian or Lorentzian envelope function. For gaseous media the bandwidth of the curve depends on the gas temperature and pressure.

In some lasers the amplifying medium consists of two parts—a host medium and the laser atoms. For example, the host medium of the Nd:YAG laser is a crystal of yttrium aluminum garnet (commonly called YAG), while the laser atoms are the trivalent neodymium ions. In gas lasers consisting of a mixture of gases, the distinction between host and laser atoms is generally not made.

The most important requirement of the amplifying medium is the ability to support a population inversion between two energy levels of the laser atoms. This is accomplished by exciting (or pumping) more atoms into the higher energy level than exist in the lower level. Here, then, we see the need for the second element of the laser, the excitation mechanism or the *pump*. For gaseous lasers, the most commonly used pump is an electric discharge. The important parameters governing this type of pumping are the electron excitation cross sections and the lifetimes of the various energy levels. In some gas lasers, the free electrons in the discharge collide with and excite the laser atoms, ions, or molecules directly, while in others, excitation occurs by means of inelastic atom-atom (or molecule-molecule) collisions. In this latter approach, a mixture of two gases is used such that the two different species of atoms, say A and B, have excited states A^* and B^* that have more or less the same energy values. Energy is transferred from the excited species to the unexcited species in the following way:

$$A^* + B \rightarrow A + B^* ,$$

where atom A obtained its excitation energy from a free electron or by means of some other excitation process. A notable example is the HeNe laser, where

the laser active neon atoms are excited by resonant transfer of energy from helium atoms in a metastable state. The helium atoms receive their energy from discharge electrons via collisions.

Although other excitation processes exist, some of which are described elsewhere in this chapter, we cite one more process that has some historical significance. The first laser, developed by Maiman, was a pulsed ruby laser that operates at the red wavelength $\lambda = 0.6943\ \mu m$ (see Fig. 10.2). To excite the Cr^{+3} ions in the ruby rod, Maiman used a helical flashlamp filled with xenon gas. This particular method of exciting the amplifying medium is known as optical pumping. It is the only practical method that can be used to pump liquid or solid (i.e., dielectric) media.

The third element of the laser is the optical resonator. In its most basic form, it consists of a pair of carefully aligned plane or curved mirrors. Their relative orientation and location with respect to the laser medium is shown in Figs. 10.2 and 10.3. The purpose of the optical resonator is to provide optical feedback that sustains laser oscillation. The reflectivity of one mirror is chosen to be as close to 100% as possible, while the reflectivity of the output mirror is less than 100%. The structure of the electromagnetic field inside the optical resonator depends on the boundary conditions at the mirrors and the requirement that the total phase change of a wave for one round-trip through the resonator equal an integer multiple of 2π. For a resonator with plane mirrors separated by a distance L, this means that

$$2kL = 2\pi q \ , \tag{10.5}$$

where q is an integer (usually of the order of 10^5 to 10^6), $k = (2\pi/\lambda)$ is the propagation constant, and λ is the wavelength of the electromagnetic field in the amplifying medium of refractive index n. The resonant frequencies of the optical resonator are obtained from Eq. (10.5); thus,

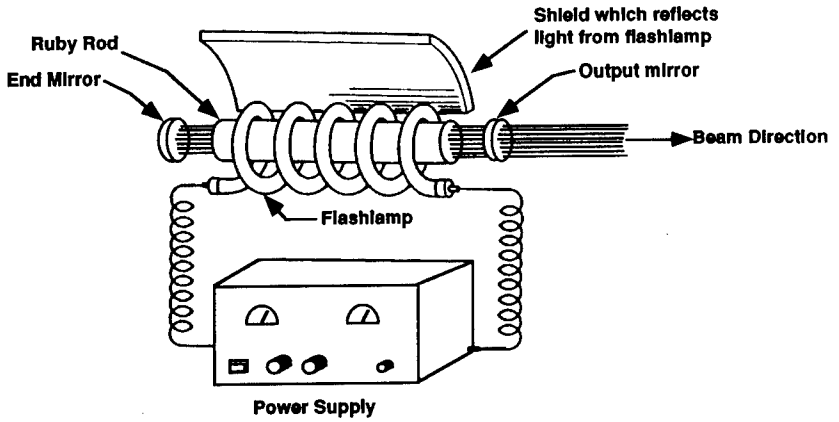


Fig. 10.2 Components of a ruby laser system. Note that, while the ruby rod, flashlamp, and mirrors may fit into an ordinary shoe box, the power supply may be as large as an office desk.

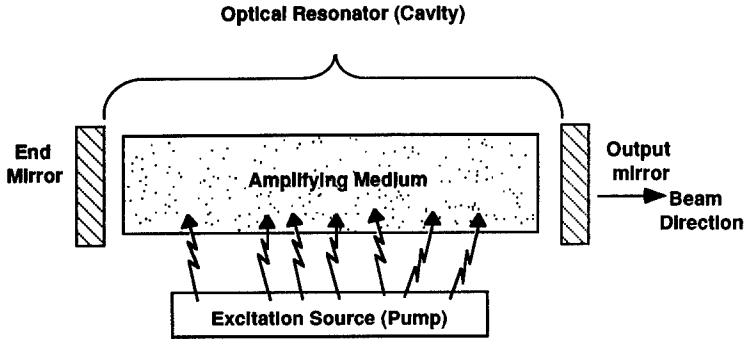


Fig. 10.3 The three essential elements of a laser.

$$f_q = \frac{v}{\lambda} = \frac{qc}{2nL} , \tag{10.6}$$

where $v = c/n$ is the speed of light in a medium of refractive index n, and c is the speed of light in vacuum. Note that the resonator can oscillate at an infinite number of equally spaced frequencies $f_q, f_{q+1}, f_{q+2}, \ldots$. The frequency difference between adjacent frequencies (i.e., the free spectral range) is given by

$$\Delta f_q = f_{q+1} - f_q = \frac{c}{2nL} . \tag{10.7}$$

This result is of considerable significance because it determines whether a given laser will operate. The frequency response of the laser medium is determined by the gain curve. Depending on the useful bandwidth of this gain curve in relation to the free spectral range, the laser may not oscillate, or it may oscillate at one or more of the resonant frequencies. These various situations are depicted in Fig. 10.4. In practice one finds that most lasers oscillate simultaneously at several resonant frequencies.

So far we have described very briefly the three essential elements of the laser. How do these elements, when put together in a certain way, lead to laser oscillation? At the start, when the excitation source is turned on, many laser atoms are rapidly excited to some higher energy state until a population inversion between an upper and lower level is established. Some excited atoms will spontaneously decay to the lower level emitting light of frequency $f = (E_2 - E_1)/h$, where h is Planck's constant and $(E_2 - E_1)$ is the energy difference between the upper and lower energy level. The light emitted in a direction perpendicular to the mirrors (along the resonator axis) and having a frequency that corresponds to a resonant frequency of the optical resonator is amplified by stimulated emission of radiation. That is, the initially spontaneously emitted light stimulates other atoms to add their internally stored energy to the electromagnetic field in such a way that direction, frequency, polarization, and phase of the field are conserved. If the energy gain of the field during one round-trip between the mirrors is larger than the energy losses (caused by diffraction, reflection, scattering, absorption, and mirror transmission), the

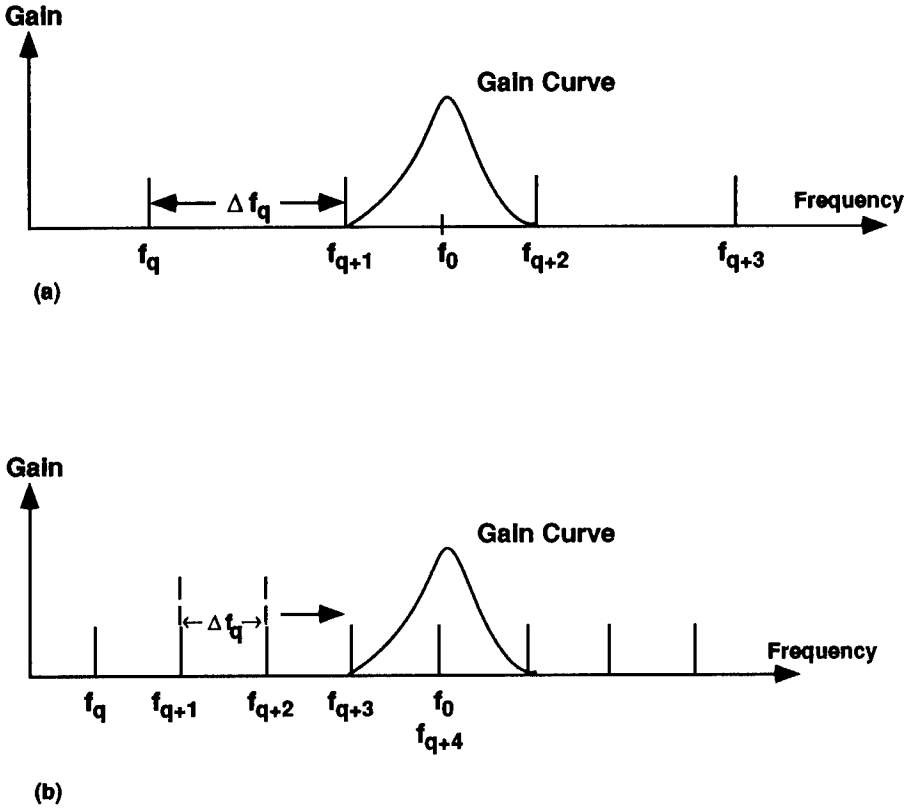


Fig. 10.4 Relationship of gain curve to resonant frequencies of the optical resonator: (a) The laser cannot oscillate because the frequency response of the laser medium does not coincide with a resonant frequency of the resonator and (b) laser can oscillate. The output will consist of a narrow frequency distribution about f_0.

amplitude of the electromagnetic field grows at the expense of the excited atoms, which are forced by the field to emit their quanta of energy in a particular direction and at a particular frequency before they emit their quanta spontaneously in any arbitrary direction. If the excitation source is sufficiently intense to ensure an adequate rate of supply of upper level atoms and if the removal rate of lower level atoms can be maintained, steady-state operation will result with a cw output beam emerging through the output mirror. On the other hand, if the pumping process exists for only a short interval of time (about 10^{-3} s for the ruby laser), then the laser emits a shorter pulse of radiation, and the laser is said to be pulsed. In the next section we take a closer look at the amplifying medium and the condition for laser oscillation.

10.2 GAIN MEDIUM

10.2.1 Gain Coefficient

Because the laser is a device that amplifies light, one of our first tasks is to determine the conditions in the amplifying medium that must be achieved for

coherent light amplification. Let us begin by considering the propagation of a plane electromagnetic wave of frequency f through a collection of atoms (or molecules) with resonant frequency f_0, such that $f_0 \simeq f$. The irradiance $E(z)$ at a point z inside the atomic medium is

$$E(z) = E_0 \exp[-\alpha(f)z] \ . \tag{10.8}$$

This equation is known as Beer's law.

The frequency-dependent absorption coefficient in Eq. (10.8) is given by

$$\alpha(f) = \frac{\lambda^2}{8\pi n^2 \tau_R} [(g_2/g_1)N_1 - N_2]\, g(f) \ , \tag{10.9}$$

where

n = index of refraction of the gain medium
λ = wavelength of the electromagnetic wave
τ_R = radiative lifetime of energy level 2
N_1, N_2 = number density of atoms or molecules in energy levels 1 and 2, respectively
g_1, g_2 = degeneracy of energy levels 1 and 2, respectively
$g(f)$ = lineshape function.

Equations (10.8) and (10.9) show that amplification of the beam irradiance will occur if $N_2 > (g_2/g_1)N_1$. This condition is known as a population inversion and is a critical requirement for the operation of a laser.

According to the Boltzmann formula, the populations of energy levels 1 and 2 are given by

$$\frac{N_2}{g_2} = \frac{N_1}{g_1} \exp[-(E_2 - E_1)/kT \ , \tag{10.10}$$

where k is the Boltzmann constant (1.38×10^{-23} J/K) and T is temperature. This equation shows that a population inversion is not possible unless we assign a negative value to the temperature. This is not a particularly worthwhile idea. Rather, note that the Boltzmann formula only applies to atomic systems in thermal equilibrium. Under certain experimental conditions to be described later, it is possible to produce a population inversion. When this is the case, $\alpha(f)$ takes on a negative value and, according to Beer's law, light is amplified rather than absorbed. It is customary then to define a gain coefficient $\gamma(f)$, where

$$\gamma(f) = -\alpha(f) = \frac{\lambda^2}{8\pi n^2 \tau_R} [N_2 - (g_2/g_1)N_1]g(f) \ . \tag{10.11}$$

The radiative lifetime τ_R for an atomic transition is given by

$$\tau_R = \frac{\tau_R'}{3F_{21}} \ , \tag{10.12}$$

where F_{21} is the oscillator strength and τ_R' is the radiative damping time of the classical electron oscillator. The radiative damping time of the classical electron oscillator is

$$\tau_R' = \frac{3\ \varepsilon_0 m_e c^3}{2\pi e^2 f^2}\ , \tag{10.13}$$

where

m_e = electron rest mass (9.1×10^{-31} kg)
e = electron charge (1.6×10^{-19} C)
ε_0 = permittivity of vacuum (8.8×10^{-12} $C^2/N\ m^2$)
c = speed of light (2.9×10^8 m/s).

The oscillator strength is a measure of the strength of an atomic transition. It is defined as

$$F_{21} = \frac{4\pi m_e f r_{21}^2}{3h}\ , \tag{10.14}$$

where r_{21} is the quantum mechanical matrix element for the transition from level 2 to level 1.

If $\Psi_2(\bar{r})$ and $\Psi_1(\bar{r})$ are the normalized wave functions corresponding to the upper and lower energy levels, then the matrix element is given by

$$r_{21} = \int \Psi_2^*(\bar{r})\ \bar{r}\ \Psi_1(\bar{r})\ d\bar{r}\ . \tag{10.15}$$

When the gain coefficient is expressed in terms of the matrix element r_{21}, we obtain

$$\gamma(f) = \frac{\pi e^2 r_{21}^2}{3\varepsilon_0 h\lambda}\,[N_2 - (g_2/g_1)N_1]\,g(f)\ . \tag{10.16}$$

Thus, we see that the gain coefficient is proportional to r_{21}^2. This result may lead us to conclude that the gain of an amplifying medium can be increased by selecting a transition that has a large matrix element. However, a large matrix element implies a strongly allowed transition with a large atomic response and a short radiative lifetime. We will see in a later section that a small τ_R implies a large pumping power for achievement of a given population inversion. Indeed, we will find that an inversion is likely to be achieved, if at all, primarily on those transitions having small matrix elements or, what amounts to the same thing, having long radiative lifetimes.

An alternative way of expressing the gain coefficient is with the stimulated transition cross section σ, such that

$$\gamma(f) = \sigma[N_2 - (g_2/g_1)N_1]\ , \tag{10.17}$$

where

$$\sigma = \frac{\pi e^2 r_{21}^2}{3\varepsilon_0 h\lambda_0}\,g(f_0)\ . \tag{10.18}$$

Note that σ is generally evaluated at line center. The unit for σ is area.

The gain coefficient $\gamma(f)$ that we have described so far is called the unsaturated gain coefficient. This is the gain coefficient that would be measured if a very low intensity probe beam were passed through the laser medium with the resonator mirrors removed. The actual gain coefficient that would be measured when the laser is operating in a cw fashion is referred to as the saturated gain coefficient; it is always less than the unsaturated gain coefficient. The reason for this will become apparent when we study the rate equations for N_2 and N_1. Suffice it to say here that $[N_2 - (g_2/g_1)N_1]$ is in practice determined by the pump. In the absence of the laser's electromagnetic field, the pump produces a population inversion that has an initial value of $[N_2 - (g_2/g_1)N_1]_0$. As the laser begins to oscillate, its electromagnetic field reduces the number of excited atoms through stimulated emission until the rate of stimulated downward transitions is just balanced by the replenishment rate of excited atoms through pumping minus relaxation. The reduction in the population inversion and, hence, the gain coefficient, brought about by the laser's field, is called gain saturation. It is the mechanism that reduces the gain coefficient of the laser medium to a level where it just balances the losses so that steady-state oscillation can result.

The frequency dependence of the gain coefficient is expressed by the lineshape function $g(f)$. In the next section, we obtain exact expressions for this important function.

10.2.2 Lineshape Function

To account for either absorption or amplification at frequencies that differ slightly from the resonant frequency f_0, we found it necessary in the previous section to introduce the lineshape function $g(f)$, which is defined as

$$g(f) = \frac{E(f)}{\int E(f)\, df} = \frac{E(f)}{E} , \tag{10.19}$$

where $E(f)$ is the isotropic spectral irradiance of the emission at frequency f and E is the integrated irradiance of the emission. The lineshape function $g(f)$ is normalized such that

$$\int_{+\infty}^{-\infty} g(f)\, df = 1 . \tag{10.20}$$

The function $g(f)$ formally recognizes that in actual practice all spectroscopic emission lines are broadened. A plot of irradiance versus frequency of the transition is bell shaped; the exact shape depends on the particular cause that is responsible for the broadening.

10.2.2.1 Natural Broadening. Natural broadening is due to the finite duration of the atomic radiation process. Because the power radiated by an excited atom decreases exponentially in time with a characteristic decay time τ_R, the emitted electric field diminishes as $\exp(-t/2\tau_R)$. The exponentially decaying field yields an irradiance distribution given by

$$E(f) \propto \left[\frac{1}{(1/2\tau_R)^2 + 4\pi^2(f - f_0)^2}\right] . \tag{10.21}$$

The function within the square brackets is named after Lorentz. Thus, a spectral line whose intensity versus frequency profile matches Eq. (10.21) is said to have a Lorentzian shape. The Lorentzian lineshape is shown in Fig. 10.5 and compared with a Gaussian lineshape, which we shall encounter later in our discussion of Doppler broadening.

The full width at half height is

$$\Delta f = 1/2\pi\tau_R , \tag{10.22}$$

and the normalized lineshape function $g(f)$ for a Lorentzian lineshape is given by

$$g(f) = \frac{\Delta f}{4\pi^2[(\Delta f/2)^2 + (f - f_0)^2]} . \tag{10.23}$$

Because of the finite duration of the atomic radiation process, the resulting "natural" linewidth Δf_N is

$$\Delta f_N = 1/2\pi\tau_R , \tag{10.24}$$

where τ_R is the radiative lifetime of the transition.

The natural linewidth is extremely small when compared to other line-broadening mechanisms, and ranges from about 10^8 Hz to less than 10^4 Hz. It can only be observed when the radiating atoms are at rest and do not interact with each other. Because Δf_N is so narrow, it is generally masked by other broadening mechanisms such as collision broadening or Doppler broadening.

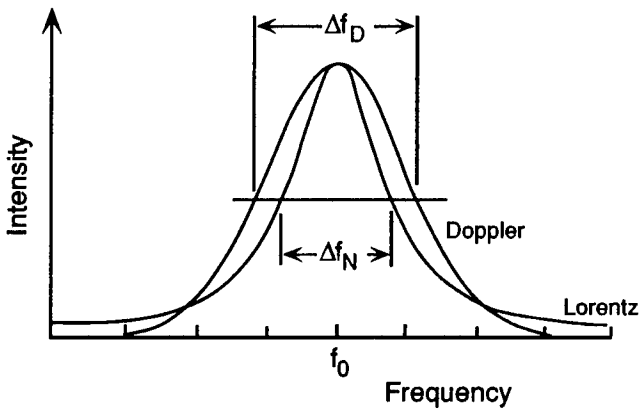


Fig. 10.5 Comparison of Gaussian and Lorentzian lineshapes of common area and peak value. The widths of the line profiles at one-half of the peak value are shown as Δf_D (for the Gaussian profile) and as Δf_N (for the Lorentzian profile).

10.2.2.2 Collision Broadening. As an atom bounces around in its container, it collides with other atoms. Because such collisions are generally elastic, no energy is transferred to or from a radiating atom. The only effect of a collision is a brief interruption of the emission (or absorption) process. The emission process is resumed after the collision without memory of the phase and plane of polarization of the emitted radiation before the encounter. Between collisions, the atom radiates at the fixed frequency f_0. The primary effect of collisions is the breaking up of the electromagnetic wave into smaller wavelets. This is shown schematically in Fig. 10.6.

Because of the frequent interruptions of the sinusoidal wave, the spectral content of the detected radiation is larger than that of the uninterrupted wave train. To calculate the linewidth, we must first calculate the frequency distribution of one wavelet of frequency f_0 and duration τ. Here τ is the elapsed time between two consecutive collisions.

Because collisions occur at random intervals, the wavelets are not all of the same duration. To find the irradiance spectrum of the entire train of wavelets, we must add the irradiance spectra of all the wavelets in the emitted electromagnetic wave. The result is

$$E(f) \propto \frac{E_0^2}{(1/T_2)^2 + 4\pi^2(f - f_0)^2}, \tag{10.25}$$

where

E_0 = amplitude of the electromagnetic wave
T_2 = average time between collisions for a single atom or molecule
f = frequency of electromagnetic wave
f_0 = frequency at line center.

Because Eq. (10.25) is of the same form as Eq. (10.21), we conclude that collisionally broadened spectral lines have a Lorentzian lineshape. The corresponding normalized lineshape function for collision broadening follows:

$$g(f) = \frac{1/T_2}{4\pi^3[(f_0 - f)^2 + (1/2\pi T_2)^2]}. \tag{10.26}$$

The full width at half height is $\Delta f_{\text{coll}} = 1/\pi T_2$. If we had assumed wavelets having an exponentially decreasing amplitude, we would have found that

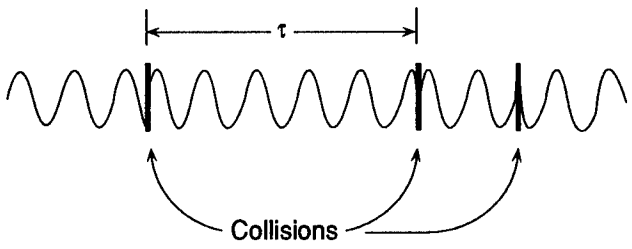


Fig. 10.6 Schematic illustration of the interruption of sinusoidal oscillation by randomly occurring collisions. The oscillation resumes after each collision with random phase.

$$\Delta f_{\text{coll}} = \frac{1}{2\pi\tau_R} + \frac{1}{\pi T_2} \ . \tag{10.27}$$

This result acknowledges that for collision broadening the complete linewidth is equal to the sum of collision plus lifetime broadening. In most actual situations the collision frequency $1/T_2$ is much larger than $1/\tau_R$.

For a single species gas the collision frequency is

$$T_2^{-1} = 4N\,\sigma_c\,(kT/\pi M)^{1/2} \ , \tag{10.28}$$

where

T = temperature
k = Boltzmann constant
N = atoms per unit volume
σ_c = collision cross section
M = atomic mass.

Because the concentration of atoms in a single species gas is related to the gas pressure p by

$$N \ \ [\text{atoms/cm}^3] = 9.65 \times 10^{18}\, p/T \ \ [\text{Torr/K}] \ , \tag{10.29}$$

collision broadening is seen to be directly proportional to the gas pressure. For a gas consisting of two different kinds of atoms, the collision frequency for collisions of a single atom of type a with atoms of type b is

$$\left(T_2^{-1}\right)_{ab} = 2\,N_b\,\sigma_{ab}\left[\frac{2kT}{\pi}\left(\frac{1}{M_a} + \frac{1}{M_b}\right)\right]^{1/2} \ , \tag{10.30}$$

where

N_b = number density of atoms of type b
M_a, M_b = atomic masses of a and b atoms
σ_{ab} = collision cross section between atom a and atom b.

Finally, for a mixture of several different gases, the total collision-broadened linewidth for atoms of any one type is obtained by summing the collision frequencies caused by every other type of atom present, including the collision frequency of the atom under consideration with atoms of its own kind. Thus,

$$\left(\frac{1}{T_2}\right)_a = \left(\frac{1}{T_2}\right)_{aa} + \left(\frac{1}{T_2}\right)_{ab} + \left(\frac{1}{T_2}\right)_{ac} \ . \tag{10.31}$$

10.2.2.3 Doppler Broadening. Doppler broadening refers to the broadening of spectral lines that occurs when radiating atoms or molecules do not all have the same velocity relative to the observer. The Doppler principle states that a stationary observer, viewing an atom that is moving with a line-of-sight

velocity v_z and radiating at the frequency f_0, will observe a frequency that differs by a small amount (the Doppler shift) from f_0. The observed frequency is

$$f = f_0[1 + (v_z/c)] \ . \tag{10.32}$$

Because atoms in a gas have random thermal motions, a stationary observer records a variety of Doppler-shifted frequencies. To be more specific, atoms (or molecules) of a luminous gas have a Maxwellian distribution of velocities and an observer would "see" a range of frequencies symmetrically distributed about the frequency of the atoms at rest. For a gas in thermal equilibrium, the fraction dN/N of atoms whose z component of velocity lies between v_z and $v_z + dv_z$ is given by

$$\frac{dN}{N} = \frac{\exp[-(v_z/u)^2]}{\sqrt{\pi}u}\, dv_z \ , \tag{10.33}$$

where u is the most probable speed. It is defined by

$$\frac{1}{2} Mu^2 = kT \ , \tag{10.34}$$

where M is the mass of the atom. If we assume that the irradiance at the frequency f is proportional to the number of radiating atoms having a velocity component v_z, then by combining Eqs. (10.32), (10.33), and (10.34), we obtain

$$\frac{E(f)\, df}{E} = \frac{\exp\left[-\left(\dfrac{f - f_0}{f_0}\right)^2 \dfrac{Mc^2}{2kT}\right]}{2\pi f_0 \left(\dfrac{2\pi kT}{Mc^2}\right)^{1/2}} df \ , \tag{10.35}$$

where E is the total integrated irradiance of the line. Notice that the spectral line has a Gaussian irradiance distribution. The spectral irradiance at line center is from Eq. (10.35);

$$E(f_0) = \frac{E}{2\pi f_0 \left(\dfrac{2\pi kT}{Mc^2}\right)^{1/2}} \ . \tag{10.36}$$

It is customary to define the half width $\Delta f_{1/2}$ as that frequency interval between $E(f_0)$ and $(1/2)\, E(f_0)$ and the full width as

$$\Delta f_D = 2\Delta f_{1/2} = 2f_0 \left(\frac{2kT \ln 2}{Mc^2}\right)^{1/2} \ , \tag{10.37}$$

where

f_0 = line-center frequency
k = Boltzmann constant

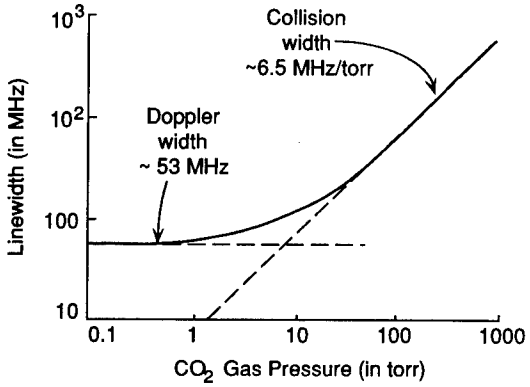


Fig. 10.7 Linewidth versus pressure for CO_2 gas. The curve shows the changeover from Doppler broadening, which dominates at pressures below ~10 Torr, to collision broadening at pressures above ~10 Torr (Ref. 2).

T = temperature
M = atomic mass of radiating atom or molecule
c = speed of light.

The full width between the half-intensity points is shown in Fig. 10.5, where the Gaussian and Lorentzian lineshapes are compared. The normalized lineshape function for Doppler broadening is

$$g(f) = \frac{2(\ln 2)^{1/2} \exp\{-[4(\ln 2)(f - f_0)^2/\Delta f_D^2]\}}{\sqrt{\pi}\Delta f_D} \tag{10.38}$$

Equation (10.37) shows that Doppler broadening is most pronounced for high-temperature gases made up of light atoms. As an example, consider the CO_2 transition with $\lambda = 10.6$ μm at 300 K. For this transition $f_0 = (c/\lambda) = 2.8 \times 10^{13}$ Hz, $M = (44 \text{ amu})(1.66 \times 10^{-27} \text{ kg/amu}) = 73 \times 10^{-27}$ kg, and $\Delta f_D \simeq 53$ MHz. Doppler broadening is the dominant broadening mechanism at CO_2 gas pressures less than about 10 Torr. At high pressures, however, the collision frequency becomes large enough so that collision broadening takes over as the dominant broadening effect. Once this occurs, the linewidth increases linearly with further increases in pressure. The changeover from Doppler broadening at low pressures to collision broadening at high pressures in pure CO_2 is shown in Fig. 10.7.

10.2.3 Homogeneous and Inhomogeneous Broadening

Natural as well as collision broadening of the radiation emitted by each excited atom in the gas has the same frequency distribution centered about f_0. Similarly, every unexcited atom has the same frequency response. This means that an electromagnetic wave of frequency f will produce the same response in each atom of the gas. For example, the likelihood of an induced transition is the same for all atoms in the sample. When this occurs, the spectral line is said to be homogeneously broadened. When a line is homogeneously broadened,

then every atom in the sample has the same resonant frequency and the same atomic lineshape and frequency response so that, if an electromagnetic field is applied to the transition, each atom has the same probability for a field interaction.

A second type, referred to as inhomogeneous broadening, is characterized by the fact that some atoms contribute radiation only to a narrow spectral region of the entire width of the line, while others contribute to a different spectral region of the line. Doppler broadening is an example of inhomogeneous broadening. The two types of broadening are illustrated schematically in Fig. 10.8.

The exact location of the radiating atom's frequency relative to the center of the spectral line is determined by the velocity of the atom relative to the spectrometer. Conversely, when a spectral line is Doppler broadened, a monochromatic wave of frequency $f' \simeq f_0$ interacts most strongly with those atoms in the sample that have the proper velocity component along the direction of wave propagation.

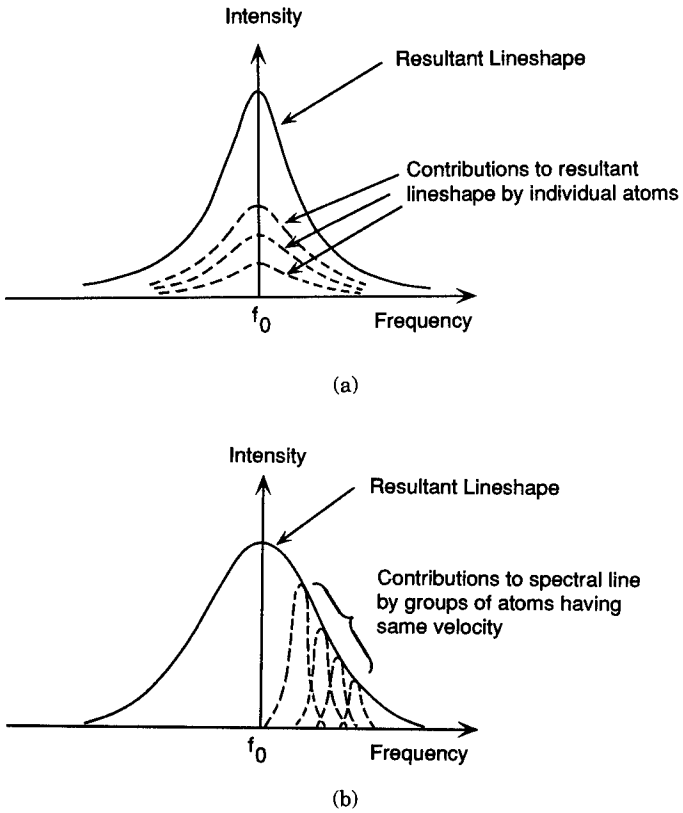


Fig. 10.8 (a) Homogeneously broadened spectral line. The radiation from each atom in the gas has the same center frequency and frequency distribution. (b) Inhomogeneously broadened spectral line. Because of Doppler shifts, different atoms or groups of atoms within the gas have slightly different resonant frequencies on the same transition. Atoms with transition frequencies located within the bandwidth Δf of a homogeneously broadened spectral packet (dashed curves) can be considered as indistinguishable. An inhomogeneously broadened gain curve is made up of many homogeneous spectral packets.

A monochromatic probe beam propagating through a tube containing an inhomogeneously broadened gain medium and having a frequency f' near the center of the spectral line will be amplified. However, the spectral line shape now exhibits a dent or "hole" at the frequency of the probe beam (see Fig. 10.9). This distortion of the line shape is generally referred to as "burning a hole" into the gain curve. The probe beam is amplified by energy that is extracted from that group of atoms whose shifted resonant frequency lies close to the probe beam frequency. The depth of the hole is proportional to the intensity of the probe beam and the width is roughly equal to two collision-broadened homogeneous linewidths.

To understand the effects of inhomogeneous broadening on laser operation, let us consider a standing wave in the resonator as consisting of two traveling waves moving in the positive and negative z directions along the resonator axis (see Fig. 10.10). According to the Doppler principle, an atom moving with a velocity v_z senses the frequency given by $f' = f_q\,[1 - (v_z/c)]$. The wave propagating to the right in Fig. 10.10(a) interacts most strongly with those atoms that, because of their motion, feel an electromagnetic field of frequency $f' = f_0$, where f_0 is the atoms' resonant frequency. This group of atoms is tuned to the oscillation frequency f_q if

$$f_0 = f' = f_q[1 - (v_z/c)] \ . \tag{10.39}$$

The necessary z component of the atoms' velocity is from Eq. (10.39), found to be

$$v_z = c(f_q - f_0)/f_q \ . \tag{10.40}$$

Atoms with this velocity have the correct Doppler shift to be in resonance with the wave traveling to the right. On the other hand, the wave traveling to the left interacts most strongly with those atoms whose velocity is

$$v_z = -c(f_q - f_0)/f_q \ . \tag{10.41}$$

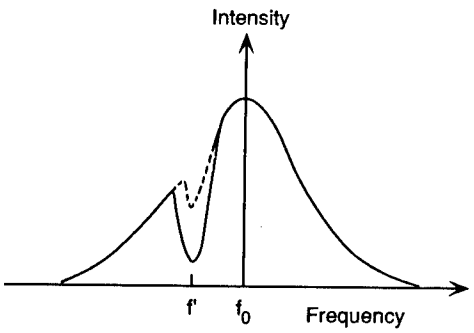


Fig. 10.9 Inhomogeneously broadened spectral line shape with a "hole." The hole is "burnt" into the line shape by a probe beam of frequency f'. The broken line shows a hole caused by a less intense probe beam.

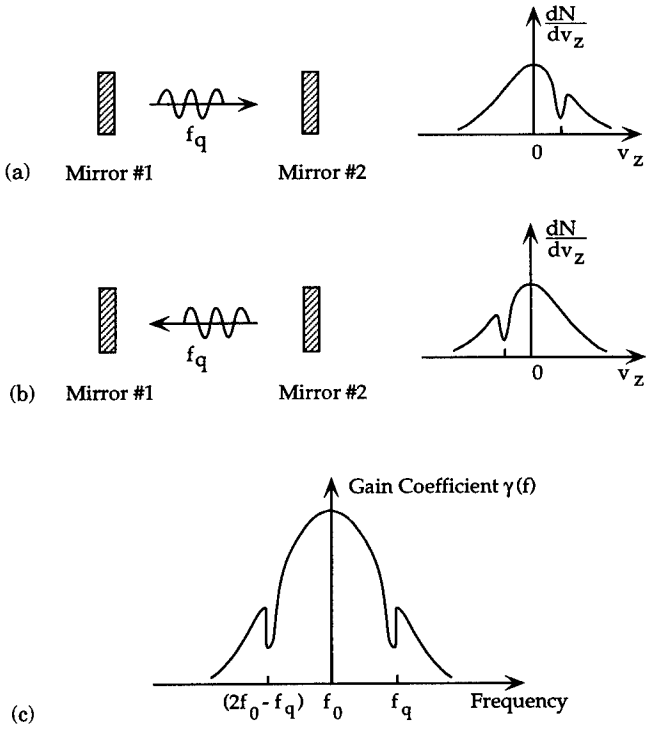


Fig. 10.10 (a) An electromagnetic wave of frequency f_q traveling to the right (positive z direction) burns a hole into the Maxwellian velocity distribution of the atoms in a gas laser. (b) Left traveling wave of frequency f_q burns a symmetrically located hole into the velocity distribution. (c) Frequency scan of Doppler broadened gain curve showing two symmetrically located holes.

From all this, we conclude that a laser oscillating at only one resonator frequency (and $f_q \neq f_0$) can extract energy from two groups of atoms. The two groups are shown in Figs. 10.10(a) and 10.10(b). A frequency scan of the gain by a probe signal would reveal two holes symmetrically located about line center. The two holes, which are shown in Fig. 10.10(c), are burned into the gain curve by the two traveling wave components of the standing wave.

The double hole burning leads to an interesting phenomenon that is observed when the cavity frequency f_q is tuned across an inhomogeneously (Doppler) broadened gain profile. A measurement of the laser's output power shows a slight dip of the output power when $f_q = f_0$. This decrease of output power is explained as follows: When the oscillating frequency is on either side of f_0, two holes are burned into the gain profile and the laser extracts energy from two groups of atoms. These two holes merge into one when $f_q = f_0$. Now the laser can extract energy from only one single group of atoms, namely, from those with $v_z = 0$. Because there are fewer atoms with this velocity component than there are atoms in the two velocity groups comprising symmetric holes somewhat removed from line center, the power output is less at $f_q = f_0$ than when f_q is slightly to one side of line center. The small, but measurable, decrease of output power, which was first predicted by Lamb and is referred to as the

Lamb dip, is used for frequency stabilization of certain gas lasers. Experimental measurements of the output power of a HeNe laser as a function of oscillation frequency are shown in Fig. 10.11. The curves show that the Lamb dip becomes more pronounced at higher excitation levels. A homogeneously broadened gain curve does not exhibit a Lamb dip.

The gain coefficient as given by Eq. (10.11) is proportional to the population difference $[N_2 - (g_2/g_1)N_1]$, where N_2 and N_1 are the total number density of atoms in the upper and lower laser levels, respectively. Equation (10.11) as written describes the gain coefficient for a homogeneously broadened gain curve. Because for inhomogeneously broadened gain curves not all atoms respond equally to a given frequency, the population difference in Eq. (10.11) must be replaced by

$$[N_2' - (g_2/g_1)N_1'] = \frac{[N_2 - (g_2/g_1)N_1]\,\Delta f_H}{\Delta f_I} , \tag{10.42}$$

where N_2' and N_1' are the level population densities of those indistinguishable atoms comprising one homogeneous packet of bandwidth Δf_H, and Δf_I is the bandwidth of the inhomogeneously broadened gain curve. This correction accounts for the fact that only those atoms whose transition frequencies are clustered within Δf_H can respond to a signal whose frequency lies within Δf_H. Also note that when the gain curve is inhomogeneously broadened, the line-shape function in Eq. (10.11) must reflect this fact.

The different properties of homogeneously and inhomogeneously broadened gain coefficients, especially their differing saturation characteristics, affect the behavior of laser oscillators in many ways. The Lamb dip is one example. Others are considered in Sec. 10.2.5.

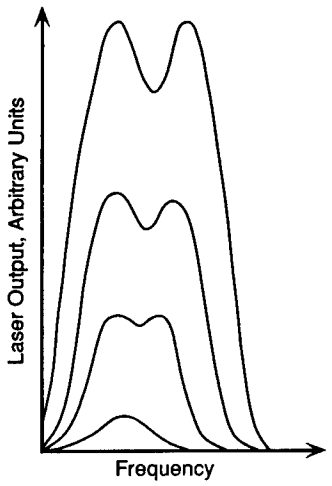


Fig. 10.11 Laser power (in arbitrary units) versus frequency of a HeNe laser operating at four different levels of excitation. The Lamb dip is observed near the center of the 1.15-μm-wavelength Ne^{20} transition.[3]

10.2.4 Threshold Condition for Oscillation

A traveling wave as it propagates back and forth between the two mirrors of the optical resonator is amplified as it passes through the gain medium and attenuated by optical components such as the windows of the tube containing the gain medium. Other energy losses include scattering, reflection, diffraction, mirror transmission, and absorption in the host medium. For well-designed lasers all losses, with the exception of mirror transmission, are negligible. However, for completeness we lump all loss mechanisms with the exception of mirror transmission together and designate $\alpha(\text{cm}^{-1})$ to be the total distributed loss coefficient.

The threshold condition for oscillation may be formulated by tracking a small segment of the optical wave of initial irradiance $E(z,t)$. After one complete round-trip through the resonator (shown in Fig. 10.23), the irradiance is

$$E\left(z,\, t + \frac{2L}{c}\right) = R_1 R_2 \exp[2(\gamma - \alpha)L]\, E(z,t) \ , \tag{10.43}$$

where R_1, R_2 are the mirror reflectivities, and $2L/c$ is the round-trip transit time. Clearly, if $R_1R_2 \exp[2(\gamma - \alpha)L] < 1$, laser oscillation cannot be sustained and the signal will decay. On the other hand, if $R_1R_2 \exp[2(\gamma - \alpha)L] > 1$, oscillation builds up until the irradiance is so large that the gain coefficient saturates and steady-state operation begins. When the laser operates at steady state, then in one round-trip through the resonator the energy gain of the wave segment is equal to all energy losses. This means that

$$R_1 R_2 \exp[2(\gamma - \alpha)L] = 1 \ . \tag{10.44}$$

Equation (10.44) is also a formulation of the threshold condition for the start of laser oscillation. The corresponding threshold gain coefficient is obtained from Eq. (10.45); it is

$$\gamma_{\text{th}} = (1/2L) \ln(1/R_1R_2) + \alpha \ . \tag{10.45}$$

Because for many lasers the loss coefficient α is much less than the first term in Eq. (10.45), and because the reflectivity of one mirror is generally chosen to be as close to unity as possible, the threshold gain coefficient may be written as

$$\gamma_{\text{th}} = -(1/2L) \ln(R) \simeq \frac{1 - R}{2L} \ , \qquad \text{for } R \simeq 1 \ . \tag{10.46}$$

The threshold population inversion for a laser medium with a homogeneously broadened, Lorentzian lineshape is obtained by combining Eqs. (10.11), (10.23) with $f = f_0$, and (10.46). Hence,

$$[N_2 - (g_2/g_1)N_1]_{\text{th}} = \frac{4\pi^3 \tau_R \, \Delta f_L}{L\lambda_0^2} [\ln(1/R_1R_2) + 2\alpha L] \ , \tag{10.47}$$

where Δf_L refers to the linewidth of the Lorentzian gain curve. The corresponding expression for a Gaussian gain curve is

$$[N_2 - (g_2/g_1)N_1]_{\text{th}} = \frac{4\pi^3 \tau_R \Delta f_G}{\lambda^2 L\sqrt{\pi \ln(2)}} [\ln (1/R_1R_2) + 2\alpha L] \ . \tag{10.48}$$

Equations (10.47) and (10.48) specify the minimum inversion necessary for laser oscillation for Lorentzian and Gaussian gain curves, respectively. This minimum inversion is not only a function of the laser medium, but also depends on the design characteristics of the optical resonator. For instance, high mirror reflectivities reduce the threshold inversion, whereas large absorptive losses in the laser medium raise the threshold inversion. The goal of a laser designer is to reduce the threshold inversion to as low a value as practical.

Finally, let us consider what happens when a laser designed to run in a cw fashion is turned on. Initially, as the pump power is turned up from zero, the population inversion and corresponding gain coefficient are too small to satisfy Eq. (10.44). Because the round-trip gain $E[z, t + (2L/c)]/E(z,t)$ is less than unity, laser oscillation cannot build up. As the pump power reaches the pumping value that produces the minimum population inversion for oscillation, the round-trip gain reaches unity and the laser begins to oscillate. If the pump power is raised to some higher value, the beam intensity increases accordingly. The population inversion, however, remains frozen at the threshold value (the reason for this is discussed in a later section), as does the gain coefficient. With the pump power stabilizing at a constant value above its threshold value, steady state prevails within the oscillator, and $E[z, t + (2L/c)] = E(z, t)$. In short, Eq. (10.44) describes both the threshold condition and the steady-state condition. Furthermore, the threshold values of the gain coefficient and the corresponding population inversion are equal to the respective steady-state values.

10.2.5 Single- and Multifrequency Oscillation

In Sec. 10.1.3 we found that the resonant frequencies of an optical resonator consisting of two plane mirrors separated by a distance L are given by

$$f_q = q\frac{c}{2nL} \ ; \quad q = 1, 2, 3, \text{etc.} \ , \tag{10.49}$$

where c is the speed of light in vacuum and n is the index of refraction of the material filling the space between the mirrors. For gas lasers, $n = 1$. The optical waves, whose frequencies satisfy Eq. (10.49) and propagate along the laser axis are called *axial modes*. For a typical gas laser with $L = 100$ cm, the free spectral range Δf_q [see Eq. (10.7)] is 150 MHz. Because the frequency range over which the gain coefficient exceeds the threshold value is often considerably larger than Δf_q, the output of the laser consists of either one or more resonant frequencies separated from each other by the frequency difference $c/2nL$. The exact nature of the spectrum of the beam depends on whether the gain curve is homogeneously or inhomogeneously broadened.

The conditions that determine the beam's spectrum when the gain curve is inhomogeneously broadened are shown in Fig. 10.12. We recall that such an inhomogeneously broadened line is composed of spectral packets of homogeneous lines [Fig. 10.8(b)]. As we saw earlier (Fig. 10.10) an optical wave of frequency $f_q \simeq f_0$ interacts only with those atoms that, by virtue of their velocity, are in resonance with the wave. The depletion of the population inversion among those atoms responsible for the packet burns holes in the gain curve. The holes are centered about f_q and $(2f_0 - f_q)$, as shown in Fig. 10.10(c). However, the population inversion among atoms producing packets at other frequencies is not depleted, and when the gain there exceeds the threshold gain for oscillation, a second axial mode bursts into oscillation. When the laser is initially turned on, the pump builds up the population inversion and the gain increases. Because the gain curve peaks at line center, the first axial mode to oscillate is the one nearest line center. In Fig. 10.12, this corresponds to the cavity mode with frequency f_3. Once oscillation takes place, the gain coefficient at f_3 saturates at the threshold value. With further increases in the pumping rate, the gain continues to rise at all frequencies. However, the gain coefficient at f_3 returns quickly to the threshold or, what amounts to the same thing, the steady-state value. When the gain coefficient for cavity modes at f_2

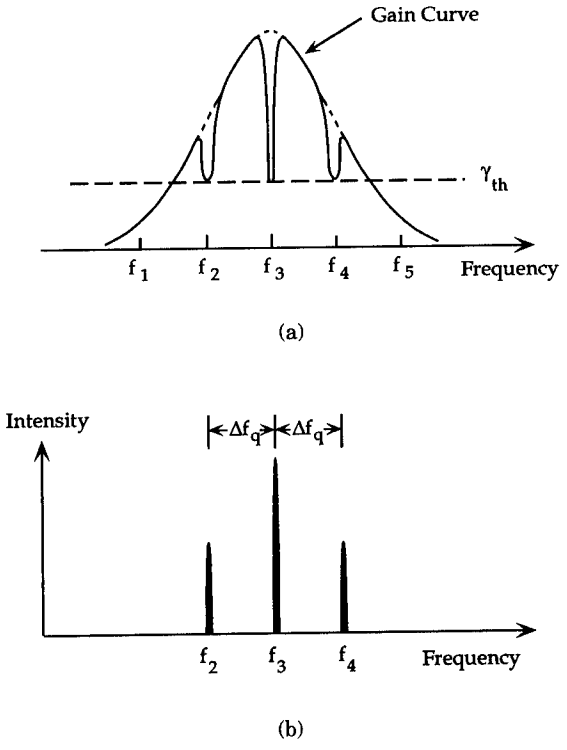


Fig. 10.12 (a) Three simultaneously oscillating axial modes of a laser with an inhomogeneously broadened gain curve. Each resonant frequency burns a separate hole into the gain curve. (b) The three frequencies of the laser beam as observed with a scanning optical-spectrum analyzer.

and f_4 reaches the threshold value, these modes begin to oscillate. A spectrum analysis of the beam would reveal three frequencies separated from each other by $\Delta f_q = c/2nL$ [see Fig. 10.12(b)]. In summary, when the gain curve is inhomogeneously broadened with a linewidth at γ_{th} broad enough to contain p axial modes, simultaneous oscillation at p frequencies can be expected.

Next, let us consider the effects of a homogeneously broadened gain curve on the spectral nature of the laser beam. When the pump is turned on, the gain rapidly increases until the gain coefficient of the axial cavity mode nearest line center reaches its threshold value. At this instance, the cavity mode will break into oscillation. Once this favored cavity mode reaches oscillation threshold, the gain coefficient cannot increase further, either at the oscillation frequency of the favored mode or at any other frequency. Further increases in the pumping rate only lead to a temporary increase in the gain coefficient and a more intense laser beam, but the gain coefficient always returns to the threshold value when the pumping rate levels off to a steady-state value. The oscillation of only one axial mode is attributed to the fact that, for homogeneously broadened gain curves, an applied electromagnetic field with a frequency anywhere within the width of the curve affects all atoms in the same way and the gain curve saturates uniformly across its full width. Therefore, for a fully homogeneous gain curve and cw operation, only the most preferred mode reaches oscillation threshold and the laser beam has only one frequency.

10.3 LASER OSCILLATION DYNAMICS

10.3.1 Buildup and Decay of the Laser Signal

The time rate of change of the number of photons n within the lasing volume defined by the resonator mirrors is

$$\frac{dn}{dt} = [\gamma - \alpha - (1/2L)\ln(1/R)]\,cn \ , \tag{10.50}$$

where

n = number of photons in the lasing volume
γ = gain coefficient
α = distributed absorption coefficient
L = distance between resonator mirrors
R = reflectivity of output mirror
c = speed of light
t = time.

For laser oscillation to begin, the gain coefficient γ must exceed γ_{th}. If at the onset of laser oscillation γ is a slowly varying function of the time, then an integration of Eq. (10.50) shows that the number of laser photons in the resonator increases exponentially according to

$$n(t) = n_0 \exp(t/\tau_b) \ , \tag{10.51}$$

where the characteristic buildup time is

$$\tau_b = \frac{1}{c[\gamma - \alpha - (1/2L)\ln(1/R)]} \quad (10.52)$$

Because of the mechanism of gain saturation, the number of photons builds up to a level such that the rate of stimulated transitions balances the rate of formation of the excited-state population in the laser medium. During the gain saturation process the number of photons increases, while the gain coefficient decreases until steady-state operation occurs. At that moment $\gamma = \gamma_{th} = \alpha + (1/2L)\ln(1/R)$, and the laser oscillator now emits a cw beam of constant power.

When the laser's power supply, or pump, is turned off, the gain coefficient vanishes and the time rate of change of the number of photons is found from Eq. (10.50) to be

$$\frac{dn}{dt} = -[\alpha + (1/2L)\ln(1/R)]\,cn \quad (10.53)$$

Integration yields

$$n(t) = n_{ss}\exp(-t/\tau_c') ,$$

where the exponential decay time constant τ_c' is

$$\tau_c' = \frac{1}{c[\alpha + (1/2L)\ln(1/R)]} \quad (10.54)$$

and n_{ss} is the steady-state number of photons of the axial mode at the instant pumping is terminated. The time constant τ_c', which is a measure of the time that a photon remains in the resonator before it is either absorbed or exits the resonator through the output mirror, is known as the photon lifetime. From Eq. (10.53) we find that the rate at which photons leave the resonator through the output mirror is

$$\left(\frac{dn}{dt}\right)_{out} = \frac{-n}{\tau_c} , \quad (10.55)$$

where

$$\tau_c = \frac{2L}{c\ln(1/R)} . \quad (10.56)$$

Equation (10.55) is an important expression because it allows us to calculate the laser's output power, which is equal to the photon energy times the rate (n/τ_c) at which laser photons stream through the output mirror. Thus,

$$P_{out} = \frac{nhf_q}{\tau_c} \quad (10.57)$$

and the output power is directly proportional to the total number of photons belonging to the axial mode of frequency f_q. How n depends on the various parameters of the laser medium and the pumping rate is an important issue we shall address shortly.

10.3.2 Pumping of Three- and Four-Level Systems

The pumping and lasing transitions of a typical laser medium actually involve a large number of energy levels. For example, Fig. 10.13(a) shows the relevant energy levels of a typical solid-state laser. The arrows pointing up indicate pumping transitions, while the arrows pointing down indicate various relaxation processes. The laser transition is identified with a thicker arrow. To simplify the analysis of such a multienergy level system, it is conventional to lump certain energy levels together and treat them as one level. In this way, the complicated set of energy levels may be approximated by the four-level model shown in Fig. 10.13(b). One important feature of this model is that the energy difference between the lower laser level and the ground state is large enough so that $(E_1 - E_0) \gg kT$ at the operating temperature T. This assumption permits us to neglect the thermal equilibrium population of level E_1. If the laser transition terminates at the ground level, or some other level E_1 whose separation from the ground state is small compared to kT, then the system can be approximated by the three-level model shown in Fig. 10.13(c). We will use these simplified, but fairly realistic, models to gain some understanding of how lasers are pumped and how the resulting population inversion depends on the excitation probability rate and the relaxation times of the various levels.

10.3.2.1 Three-Level Systems. In the three-level system level 3 represents all energy states lying above the upper laser level. The pumping process excites atoms from the ground level 1 to level 3. This is equivalent to pumping into any of the higher energy levels. The excitation process in optically pumped solid-state or liquid lasers involves incident light from either an external light source such as a flashlamp, or from a second laser operating at a suitable pump

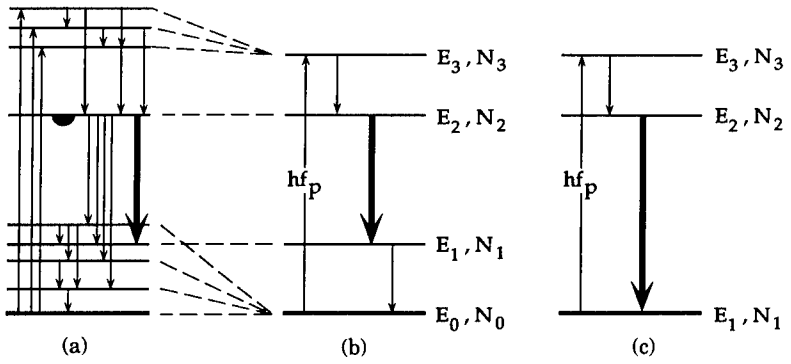


Fig. 10.13 Approximation of energy levels by idealized energy-level systems: (a) actual energy levels, (b) energy-level diagram of an idealized four-level laser, and (c) three-level laser.

wavelength. In gas lasers the pumping process may involve electron collisions or energy transfer between different kinds of colliding atoms. In any case the details of the excitation process are designated by the pump-induced transition rate. Once atoms have been excited to level 3, they relax (usually in a very short time) to level 2. If level 2 is quasi-stable, i.e., it has a very long relaxation time—sometimes indicated on energy-level diagrams by a black semicircle under the line representing the energy level, as in Fig. 10.13(a)—then a population inversion may occur between levels 2 and 1. The ruby laser (λ = 6943 Å), which was the first laser to be demonstrated, is an almost ideal example of a three-level laser system (see Fig. 10.14).

Let us calculate the population inversion in terms of the relaxation times τ and pump-induced transition rate R_p for a medium that is pumped, but not allowed to lase. The rate equation for level 3 is

$$\frac{d}{dt}N_3 = R_p\,(N_1 - N_3) - \frac{N_3}{\tau_3}\,, \tag{10.58}$$

where we use the convention that τ_n is the total relaxation time of level n caused by decays to *all* lower levels, while τ_{nm} is the relaxation time for a transition from level n to the lower level m. Thus,

$$\frac{1}{\tau_3} = \frac{1}{\tau_{32}} + \frac{1}{\tau_{31}}\,. \tag{10.59}$$

Note that τ_{nm} represents the measured relaxation time or actual lifetime of the $n \rightarrow m$ transition; τ_{nm} is a measure of the time interval over which the atom loses internal energy because of all damping processes.

The rate equation for level 2 (the upper laser level) in the absence of laser oscillation is

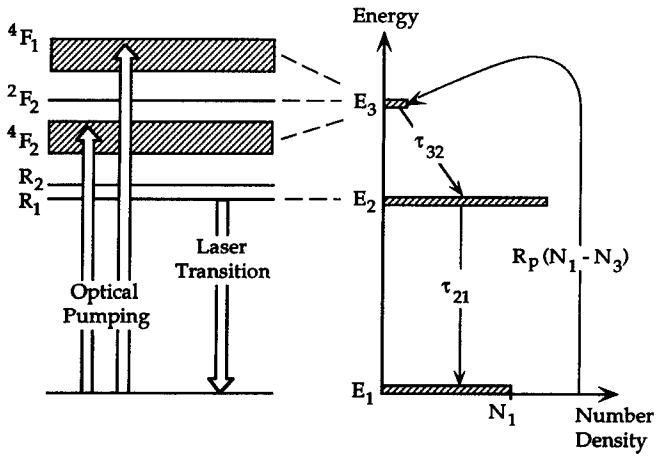


Fig. 10.14 Pumping and laser transition (λ = 6943 Å) of a ruby laser. The corresponding three-level energy system is shown on the right.

$$\frac{d}{dt} N_2 = \frac{N_3}{\tau_{32}} - \frac{N_2}{\tau_2} , \tag{10.60}$$

and the third equation simply states that

$$N_1 + N_2 + N_3 = N . \tag{10.61}$$

For steady-state operation, Eqs. (10.58), (10.60), and (10.61) yield the population difference between levels 2 and 1:

$$\frac{\Delta_{21}}{N} = \frac{(\tau_2 - \tau_{32})(R_p \tau_3/\tau_{32}) - 1}{(\tau_2 + 2\tau_{32})(R_p \tau_3/\tau_{32}) + 1} , \tag{10.62}$$

where $\Delta_{21} = N_2 - N_1$. Equation (10.62) shows that an inversion is possible if

$$\tau_2 > \tau_{32} \tag{10.63}$$

and

$$R_p \geq R_{p,\text{th}} = \frac{1}{\tau_2[1 - (\tau_{32}/\tau_2)](\tau_3/\tau_{32})} . \tag{10.64}$$

The latter condition states that the pump-induced transition rate R_p must exceed a threshold value before an inversion is possible—even if the first condition is met. Because of practical considerations it is desirable to keep $R_{p,\text{th}}$ as small as possible. Ideally, then, we should search for transitions with τ_2 very long and τ_{32} very short. For such ideal three-level systems, Eq. (10.62) reduces to

$$\left(\frac{\Delta_{21}}{N}\right)_{\text{3-level}} = \frac{R_p - (1/\tau_2)}{R_p + (1/\tau_2)} . \tag{10.65}$$

It is easy to show that at threshold $N_2 = N_1$. Thus, more than half the atoms in the laser medium must be pumped from the ground level into the upper laser level (level 2) before the laser can break into oscillation. This is a major disadvantage of the three-level system. The pump power required for a population inversion must exceed the threshold power

$$P_{\text{th}} = R_{p,\text{th}}\, hf_p\, (N/2) = \frac{Nhf_p}{2\tau_2} , \tag{10.66}$$

where hf_p is the energy difference between level 3 and the ground state.

10.3.2.2 Four-Level Systems. In the four-level system the pumping mechanism excites atoms from the ground state (level 0) to level 3, and the laser transition takes place between levels 2 and 1. Assuming steady-state operating conditions, the population inversion is given by

$$\frac{\Delta_{21}}{N} = \frac{(1 - \beta)\eta R_p \tau_2}{1 + [(1 + \beta) + 2(\tau_{32}/\tau_2)]\eta R_p \tau_2} , \tag{10.67}$$

where

$$\beta = \left(\frac{\tau_{10}}{\tau_{21}} + \frac{\tau_{10}}{\tau_2}\frac{\tau_{32}}{\tau_{31}}\right) , \tag{10.68}$$

$$\eta = \frac{(N_3/\tau_{32})}{(N_3/\tau_{32}) + (N_3/\tau_{31}) + (N_3/\tau_{30})} = \frac{\tau_3}{\tau_{32}} \leq 1 . \tag{10.69}$$

Note that if $\beta < 1$, then $N_1 < N_2$. In other words, as long as the lifetime of level 2 is longer than the lifetime of level 1 and $\tau_{32} \lesssim \tau_{31}$, then a population inversion on the $2 \rightarrow 1$ transition is virtually certain, even for vanishingly small pump transition rates. The pumping efficiency factor η measures the fraction of the total atoms excited to level 3 that decay from there to level 2, thereby becoming potentially useful for laser oscillation.

For an ideal four-level system with $\beta \simeq 0$, $\eta \simeq 1$, and $\tau_{32} \ll \tau_2$, Eq. (10.67) simplifies to

$$\left(\frac{\Delta_{21}}{N}\right)_{\text{4-level}} = \frac{R_p}{R_p + (1/\tau_2)} . \tag{10.70}$$

A comparison of the three- and four-level systems, assuming ideal conditions in each case, can be made by comparing the threshold pump powers P_{th} for both laser systems. The threshold pump power for the ideal four-level system is

$$(P_{\text{th}})_{\text{4-level}} \simeq h f_p R_{p,\text{th}} N_0 \simeq \frac{\Delta_{21,\text{th}}}{\tau_2} h f_p . \tag{10.71}$$

A comparison of Eqs. (10.66) and (10.71) shows that the pump power to reach the threshold condition in a three-level laser must exceed that of a four-level laser—all other functions being equal—by

$$\frac{(P_{\text{th}})_{\text{3-level}}}{(P_{\text{th}})_{\text{4-level}}} = \frac{N}{2(N_2 - N_1)_{\text{th}}} = \frac{N}{2\Delta_{21,\text{th}}} . \tag{10.72}$$

The number density of chromium ions in a ruby laser is $N \simeq 2 \times 10^{19}\ \text{cm}^{-3}$ and $\Delta_{21,\text{th}} \simeq 10^{16}\ \text{cm}^{-3}$ for the four-level Nd^{3+}:YAG laser. Thus, a unit volume of ruby rod requires $\sim$1000 times more pump power to reach threshold than a unit volume of Nd^{3+}:YAG rod. This example illustrates dramatically the general fact that four-level lasers are more efficient than three-level lasers.

10.3.3 Laser Rate Equations

The laser rate equations describe the transient and dynamical behavior both of the level populations and the number of laser photons in the resonator.

These equations are used to calculate the laser power output, the optimum output coupling of the resonator, the laser linewidth, and other important properties of lasers.

At the core of the laser rate equations is the generalized two-level system shown in Fig. 10.15. The use of this simplified energy level diagram, rather than the four-level system described earlier, is justified by noting that in general the level populations N_2 and N_1 are very small compared to the ground-state population N_0, so that during oscillation the latter is hardly affected, and $N_0 \simeq N$.

The photon rate equation that describes the gain and loss rates of photons contained in one axial mode is

$$\frac{dn}{dt} = Kn(t)\,\Delta(t)_T + KN_{2,T}(t) - \frac{n(t)}{\tau_c}\,, \tag{10.73}$$

where

$n(t)$ = number of photons within the resonator at time t
K = stimulated emission coupling coefficient
$\Delta(t)_T$ = $[N_2(t) - N_1(t)]V$
$N_2(t), N_1(t)$ = number density of atoms/molecules in levels 2 and 1, respectively, at time (t)
V = lasing volume
$N_{2,T}(t)$ = $N_2(t)V$
τ_c = photon lifetime [see Eq. (10.54)].

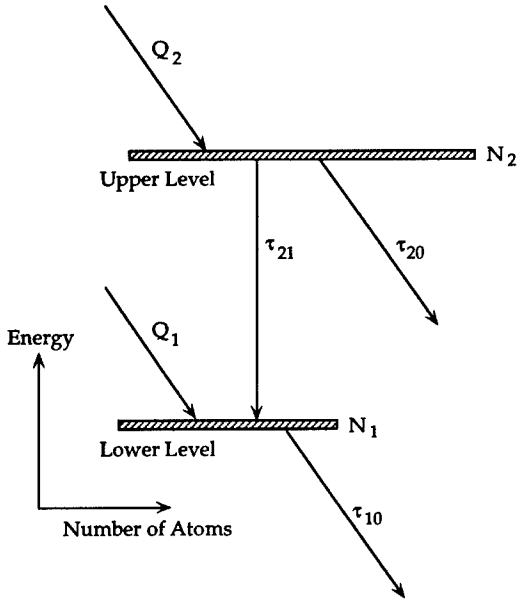


Fig. 10.15 Two-level atomic model for use in a laser rate equation analysis.

The first term in Eq. (10.73) is the time rate of change of photons caused by stimulated emission. The second term accounts for spontaneous emission of photons into the axial mode, and the last term describes the rate of decrease of the number of photons caused by cavity losses.

Note that the level populations $N_{2,T}$ and $N_{1,T}$ in the photon rate equation are functions of time t. Their time dependence is governed by

$$\frac{d}{dt}N_{2,T}(t) = -Kn(t)\,\Delta_T(t) - \frac{N_{2,T}(t)}{\tau_2} + Q_2 \,, \tag{10.74}$$

$$\frac{d}{dt}N_{1,T}(t) = Kn(t)\,\Delta_T(t) + \frac{N_{2,T}(t)}{\tau_{21}} - \frac{N_{1,T}(t)}{\tau_{10}} + Q_1 \,, \tag{10.75}$$

where Q_2 and Q_1 are the excitation rates of atoms "pumped" per second into levels 2 and 1, respectively. Pumping of level 1 causes a reduction of the gain coefficient and is detrimental to laser operation. In many lasers, and especially electric discharge pumped gas lasers, considerable pumping into level 1 is unavoidable; therefore a realistic analysis of such a system must take this into consideration.

Equations (10.74) and (10.75) describe the time rate of change of the total number of atoms in the upper and lower laser levels, respectively. The various relaxation terms depend on the characteristic transition lifetimes. For instance, $N_{1,T}(t)/\tau_{10}$ in Eq. (10.75) describes the number of spontaneous transitions per unit time from level 1 to some lower level (usually the ground state).

The rate equations in this section are only approximately correct because our use of photons does not permit the inclusion of phase information and because higher order nonlinear effects that may arise from intense applied fields have been neglected. However, because our interest here is only in first-order linear effects, we will see that these equations are nevertheless quite useful.

10.3.4 Steady-State Operation

Because of the product terms $n(t)N_{2,T}(t)$ and $n(t)N_{1,T}(t)$, the rate equations are nonlinearly coupled. Solutions are found easily only in the steady-state case (i.e., $d/dt = 0$), which implies cw operation. Fortunately, this case is of most interest for many laser applications. By setting the time derivative in Eq. (10.73) equal to zero, we obtain the steady-state photon population

$$n = \frac{N_{2,T}}{\dfrac{1}{K\tau_c} - \Delta_T} \,. \tag{10.76}$$

This result reveals that there is a critical or threshold value of the level population difference Δ_T given by

$$(\Delta_T)_{\text{th}} = (K\tau_c)^{-1} \,. \tag{10.77}$$

When Δ_T approaches this threshold value, the photon population of the resonator mode approaches infinity. This behavior is typical of all lasers.

The dependence of the below-threshold level populations $N_{1,T}$ and $N_{2,T}$ on the pumping rates Q_1 and Q_2 are found from Eqs. (10.74) and (10.75) by ignoring the stimulated emission term $Kn\Delta_T$. The results are

$$N_{1,T} = \left\{\frac{1}{[1 + (\tau_{21}/\tau_{20})]} + r\right\}\tau_{10}Q_2 \tag{10.78}$$

and

$$N_{2,T} = \frac{\tau_{21}Q_2}{[1 + (\tau_{21}/\tau_{20})]} , \tag{10.79}$$

with $r = Q_1/Q_2$. The below-threshold value for the energy-level population difference is

$$(N_{2,T} - N_{1,T})_{\text{below-th}} = \left(1 - \frac{\tau_{10}}{\tau_{21}} - \frac{r\tau_{10}}{\tau_2}\right)\tau_2 Q_2 . \tag{10.80}$$

The population difference is directly proportional to the pumping rate Q_2. Furthermore, for an ideal energy-level system $\tau_{10} \ll \tau_{21} \simeq \tau_2$, the population difference is also directly proportional to the total relaxation time τ_2 of the upper level. In fact, because Q_2 is always limited in value by power supplies, cooling requirements, or other material limitations, sufficiently large population inversions for laser oscillation can generally only be obtained when the upper laser level has a long radiative lifetime.

If we define the pumping rate Q_2 needed to bring Δ_T to its threshold value by $Q_{2,\text{th}}$, then it follows that

$$Q_{2,\text{th}} = \frac{1}{\tau_c\tau_2 K\left(1 - \dfrac{\tau_{10}}{\tau_{21}} - \dfrac{r\tau_{10}}{\tau_2}\right)} . \tag{10.81}$$

This result shows $Q_{2,\text{th}}$ depends critically on the ratios of the relaxation times as well as the pumping ratio r. In particular, the threshold pumping rate is smallest when $\tau_{10} \ll \tau_{21} \simeq \tau_2$ and $Q_1 \ll Q_2$. The restriction on Q_1 can be relaxed as long as τ_{10} is several orders of magnitude less than τ_2. When that is the case, Q_1 can be larger than Q_2, and an inversion is still possible because of the much more rapid decay of the lower level population. This important result explains why some lasers work even though no selective pumping of the upper laser level takes place.

In the above-threshold regime the laser oscillates at steady state with n photons in the resonator mode volume. Because n is always a very large number ($n \sim 10^8$ for a typical HeNe laser), the injection rate $KN_{2,T}$ of spontaneously emitted photons into the laser mode may be neglected. With this approximation (and $dn/dt = 0$), Eq. (10.73) yields the above-threshold population difference

$$(N_{2,T} - N_{1,T})_{\text{above-th}} = \frac{1}{K\tau_c} . \tag{10.82}$$

But from Eq. (10.77), $(1/K\tau_c) = (\Delta_T)_{\text{th}}$. Therefore, we conclude that for steady-state operation the above-threshold population difference is equal to the population difference at threshold and does not depend on the pumping rate.

As we have already mentioned in a previous section, for laser oscillation to build up from spontaneous emission, the gain coefficient must initially exceed the threshold gain coefficient. This means that the population difference must initially exceed the threshold population difference. A pumping rate $Q_2 > Q_{2,\text{th}}$ will accomplish this and cause the laser gain to exceed the resonator losses. As oscillation builds up and the laser photon number increases, the population difference decreases from its initial value until the rate of stimulated downward transitions just balances the pumping rate. The gain coefficient is now equal to its saturated value and the round-trip gain in the resonator is stabilized at exactly unity. The expression relating the saturated gain coefficient γ_s to the number of laser photons is

$$\gamma_s = \frac{\gamma_0}{1 + (n/n_s)} , \tag{10.83}$$

where $n_s = (1/K\tau_2)$ corresponds to the number of photons that cause the gain coefficient to drop to one-half of its initial or unsaturated value γ_0. This result shows how the gain coefficient saturates with increasing photons in the laser mode for a laser with a homogeneously broadened gain curve. For an inhomogeneously broadened gain curve it can be shown that the saturated gain coefficient is

$$(\gamma_s)_{\text{inhomo}} = \frac{\gamma_0}{[1 + (n/n_s)]^{1/2}} . \tag{10.84}$$

We conclude that for inhomogeneous broadening, gain saturation proceeds more slowly as the number of mode photons increases.

10.3.5 Output Power of a Laser

The output power of a laser is an important parameter because it determines what the laser can be used for. In this section, we use the laser rate equations to obtain a simple equation that relates the laser's output power to the parameters of a laser system.

Let us consider a laser oscillating in one single axial mode of frequency f and containing n photons. Then from Eq. (10.57) the laser's output power is given by

$$P = \frac{nhf}{\tau_c} . \tag{10.85}$$

Here τ_c, the photon lifetime determined only by the loss rate caused by mirror transmission, is given by Eq. (10.56). A more useful expression for the output

power can be obtained if the number of photons in Eq. (10.85) is expressed in terms of the pump power and other known parameters of the generalized two-level atomic system shown in Fig. 10.15.

To find an equation for the number of photons n in terms of Q_2 and the other laser parameters, we must solve the three rate equations for n for the above-threshold operation. The result is

$$n = \frac{\tau_c(\Delta_T)_{\text{th}}}{\left(1 + \dfrac{\tau_{10}}{\tau_{20}}\right)\tau_2}\left(\frac{Q_2}{Q_{2,\text{th}}} - 1\right) . \tag{10.86}$$

Several observations can now be made:

1. The photon population above threshold increases linearly with the pumping rate Q_2. Therefore, the output power is directly proportional to Q_2.
2. The photon population is largest when $\tau_{10} \ll \tau_{20}$. This means that the upper level of the laser transition should be metastable or long lived, and the lifetime of the lower level should be quite short.
3. The threshold pumping rate $Q_{2,\text{th}}$ is smallest when $\tau_{10} \ll \tau_{21}$, $\tau_{10} \ll \tau_2$, and the pumping ratio r is less than one.
4. For an ideal laser system, that is, one where all of the above ideal conditions are satisfied, the cw output power is

$$P = \frac{nhf}{\tau_c} = hf(Q_2 - Q_{2,\text{th}}) \; ; \qquad \text{for } Q_2 > Q_{2,\text{th}} \; . \tag{10.87}$$

Thus, the output power is directly proportional to the number of atoms excited to the upper laser level per unit time. Therefore, if the goal is to build high-energy (high cw power) lasers, it is important that the pumping processes and their efficiencies be understood. This knowledge must then be used to maximize Q_2 and minimize $Q_{2,\text{th}}$. When $Q_2 \gg Q_{2,\text{th}}$, the output power approaches the ideal limit $P = hfQ_2$. This means that every atom excited to the upper laser level by the pump adds one laser photon to the laser beam.

Each type of laser has its own special pumping mechanism. Excitations in gas lasers are governed by one or more of the following collision processes:

1. electron impact on atoms or molecules
2. atom-atom collisions
3. atom-molecule collisions
4. molecule-molecule collisions
5. chemical reactions.

For example, in the electric discharge laser, electrons collide with the lasing atoms or molecules and produce a population inversion. The efficiency of this process can be defined by

$$\eta = \frac{Q_2 hf_p}{iV} \; , \tag{10.88}$$

where

$$Q_2 = N_a n_e \sigma_e V_m (v_a + v_e) ,$$

and

hf_p = excitation energy to pump one atom/molecule
i = discharge current
V = discharge voltage
N_a = number density of lasing atoms/molecules
n_e = electron density
σ_e = excitation cross section
V_m = mode volume
v_a = average velocity of atoms = $(8kT/\pi M_a)^{1/2}$
v_e = average drift velocity of electrons.

The important parameter for the electron excitation process is the electron excitation cross section σ_e, which depends strongly on the electron energy. The main feature of σ_e is the threshold electron energy below which no excitation occurs. The cross section rises sharply to a peak and then falls off more or less rapidly thereafter. As a general rule, electron excitation is not a highly selective pumping process.

Equation (10.88) suggests that Q_2 is linearly proportional to the electrical input power from the power supply. In reality, this is not the case. As the electrical input power is increased by turning up the voltage, the electrons' energy distribution changes. This change will affect the excitation of the various energy levels. As the voltage is raised even more, the average electron energy will increase to the point where dissociation of the lasing molecule may occur. Because of these complications, Q_2 is generally not linearly proportional (except for short intervals) to the external pump power.

In fast-flow gas lasers, including the gas dynamic laser, the pumping rate depends on the velocity of the gas flow. Specifically, $Q_{2,\text{flow}} = N_2 uA$, where N_2 is the number density of atoms or molecules in the upper laser level, u is the velocity of gas flow, and A is the cross-sectional area of the gas flow.

In chemical lasers where the population inversion is a direct result of a chemical reaction, the pumping rate (for the simplest case) is $Q_{2,\text{chem}} = kN_{a,T}N_{b,T}$, where k is the reaction rate coefficient and $N_{a,T}$ and $N_{b,T}$ are the total number of reactant particles of species a and b. In this case, $Q_{2,\text{chem}}$ is linearly proportional to the product of the partial pressures of the reactants.

10.3.6 Removal of Waste Energy

Most lasers are very inefficient when it comes to converting input (excitation energy) into output (coherent radiation). One reason for this is that the energy given up by the atomic system in its downward transition from the lower laser level to the ground state does not contribute to the laser's output power. Consequently, a considerable amount of energy is wasted. This unused energy (or part of it) may appear as heat that, if not removed, will raise the temperature of the laser medium. According to the Boltzmann equation, this results in a larger population of the lower energy levels, which in turn reduces the pop-

ulation inversion and the gain. In most molecular gas lasers, the lower laser level is less than 1 eV from the ground state. Thus, to avoid thermal pumping of the lower laser level and a seriously reduced efficiency, the waste energy must be removed as quickly as possible.

It is easy to show that the output power is, in fact, limited by the waste energy removal rate of the cooling system (see Fig. 10.16).

For cw operation,

$$\frac{P}{P_{\text{waste}}} = \frac{Q_2 hf}{Q_2 hf_p - Q_2 hf} = \frac{f}{f_p - f}, \tag{10.89}$$

or

$$P = \frac{f}{f_p - f} P_{\text{waste}},$$

where P is the laser power in the output beam and P_{waste} is the waste energy generation rate. Under steady-state conditions, the waste energy removal rate must be equal to the waste energy generation rate. If this were not so, the temperature of the laser would be changing with time, and consequently the level populations, which means that steady state has not been achieved. With this interpretation then, it is clear that the laser output power P is directly proportional to the waste energy removal rate. This result is very important and leaves laser engineers with the problem of removing the waste energy as fast as it is produced.

Just how this is accomplished depends on the type of laser. In solid-state lasers waste energy is conducted to the surface of the rod where it is removed by a coolant. In the case of a gas laser, waste energy may be disposed of by the conduction of heat to the walls of the gas container. This process is characterized by a diffusion time τ_d, where

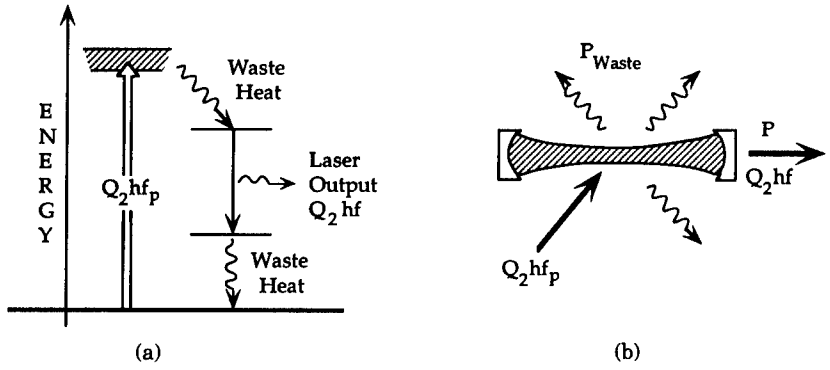


Fig. 10.16 Energy accounting in the laser cycle. In (a), on an appropriate energy level diagram for the laser cycle, the total power in, waste, and laser power out are shown. In (b), the same quantities are identified in a schematic diagram of the laser.

$$\tau_d = \frac{D^2}{lv} , \tag{10.90}$$

with

D = diameter of tube containing the gas
l = mean free path length
v = average (thermal) molecular speed.

In a convection or transverse flow-cooled laser, the waste energy is rejected in a time $\tau_f = D/u$, where u is the speed of the gas flow. Using Eqs. (10.89) and (10.90), the ratio of cw output power P_d from a diffusion-cooled laser to the cw output power P_f from an otherwise identical but transverse flow convection-cooled laser is

$$\frac{P_d}{P_f} = \frac{\tau_f}{\tau_d} . \tag{10.91}$$

Thus, for the same active volume, gas density, and operational efficiency, the ratio of the output powers of a stagnant (diffusion-cooled) laser to that of a flowing gas laser is equal to the ratio of the characteristic cooling times (τ_f/τ_d). In terms of system parameters the ratio is

$$\frac{P_d}{P_f} = \frac{l}{D(u/v)} . \tag{10.92}$$

Because the speed of sound in a gas is approximately equal to the mean molecular speed v, the ratio of the velocities in Eq. (10.92) can be replaced by the Mach number M. With this substitution, the output power of a transverse flow-cooled laser is $P_f = P_d(DM/l)$. For typical convectively cooled gas lasers the factor (DM/l) varies between 10^3 and 10^5. Therefore, by using high-speed flow to remove the waste energy more quickly, the cw output power can be increased by several orders of magnitude. The realization of this in the late 1960s more or less ended the research on cw diffusion-cooled CO_2 lasers. The emphasis was quickly switched to lasers using high gas flow rates and convective cooling. These second-generation lasers are capable of producing cw output powers in the 10^5- to 10^6-W range. However, the experimental problems that must be overcome are formidable and center on the provision of high-pressure, high-mass flow rates with adequate stable excitation under high-flow fluid dynamic conditions. The design of resonators and optical elements capable of withstanding these huge powers are a continuing challenge to laser engineers, even though considerable progress has already been made.

10.3.7 Optimum Laser Output Coupling

For a given pumping rate Q_2, the laser's output power is maximized when the value of the threshold pumping rate $Q_{2,\text{th}}$ is at its minimum, the smallest value of $Q_{2,\text{th}}$ being zero. Selection of an output mirror with a reflectivity R as close to unity as technically possible would make $Q_{2,\text{th}}$ a minimum. Unfortunately,

such a laser oscillator has an output power that is very small or zero, and all the available laser power would be dissipated by internal resonator losses, such as absorption, window reflection, and diffraction. On the other hand, if the reflectivity of the output mirror is made very low for the purpose of extracting a lot of power, the total resonator losses—which now also include mirror transmission—might exceed the round-trip gain, and the laser might not oscillate. Between these two extremes there is an optimum mirror reflectivity at which the output power is maximum. In this section, we obtain a simple expression for the optimum mirror reflectivity.

For simplicity, consider a laser oscillating at one single frequency f with a homogeneously broadened gain curve. The output power in terms of the reflectivity R of the output mirror is given by

$$P = \frac{n_s hf\, c \ln(1/R)}{\ln(1/R) + 2L\alpha} [\gamma_0 - \alpha - (1/2L) \ln(1/R)] , \tag{10.93}$$

where

n_s = number of photons when the gain coefficient is saturated
hf = energy of laser photon
α = distributed internal loss coefficient
L = distance between resonator mirrors
γ_0 = unsaturated gain coefficient
c = speed of light.

The solid curve in Fig. 10.17 shows how the output power P varies with the reflectivity R for constant values of the resonator's distributed internal loss coefficient α and the unsaturated gain coefficient γ_0. The curve shows that there exists an optimum reflectivity that leads to the maximum available output power. The total internal energy E_t within the resonator as a function of the mirror reflectivity is

$$E_t = nhf = \frac{hfn_s[\gamma_0 - \alpha + (1/2L) \ln R]}{\alpha - (1/2L) \ln R} . \tag{10.94}$$

The dashed curve in Fig. 10.17 is a plot of the laser energy stored in the resonator versus mirror reflectivity. Both curves show that the laser begins to oscillate when $R = R_{\text{min}}$. As the reflectivity is increased beyond R_{min}, both the internal energy and the output power increase. The internal energy continues to rise while the output power decreases once $R > R_{\text{opt}}$.

The optimum reflectivity is found by setting $dP/dR = 0$, and solving for $R \equiv R_{\text{opt}}$. The result is

$$R_{\text{opt}} = \exp[2L(\alpha - \sqrt{\alpha\gamma_0})] , \tag{10.95}$$

and the maximum output power with this optimized reflectivity is

$$P_{\text{max}} = n_s hfc\, (\sqrt{\gamma_0} - \sqrt{\alpha})^2 . \tag{10.96}$$

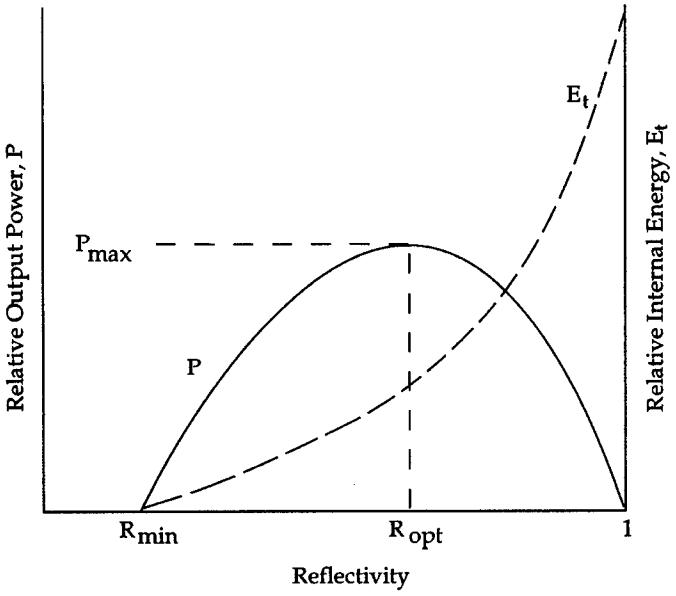


Fig. 10.17 Laser output power (solid curve) as a function of exit mirror reflectivity. The maximum output power P_{max} occurs when $R = R_{opt}$. The broken curve shows the dependence of the internally stored laser energy E_t on mirror reflectivity.

The appearance of the resonator's internal loss coefficient α in Eqs. (10.95) and (10.96) allows us to make two observations. First, as $\alpha \rightarrow 0$, the optimum reflectivity R_{opt} approaches unity, and the maximum output power is obtained with zero output coupling. The explanation for this seemingly contradictory result lies in the fact that the energy stored within the resonator, nhf, with n given by Eq. (10.83), approaches infinity as $R \rightarrow 1$ (because γ_s approaches zero), while the output power approaches the constant value $P = n_s hfc\gamma_0$. This case (i.e., $\alpha = 0$) is of academic interest only because, in reality, all laser resonators have some internal losses caused by less than perfect mirrors, scattering and absorption by imperfect windows, scattering within the gain medium, and even scattering by tiny particles of dust floating through the mode volume. However, and this brings us to the second observation, reducing the internal loss coefficient is important for obtaining maximum output power.

The expression for the optimum reflectivity contains the unsaturated gain coefficient γ_0. However, γ_0 is directly proportional to the pumping rate, which means that the optimum reflectivity is different for different values of the pumping rate. Therefore, a laser can be designed with an optimum reflectivity for only one specific pumping rate.

10.3.8 Dynamics of Laser Oscillation

In many lasers, and particularly in solid-state lasers, the laser output is not one continuous wave train, even when the pumping rate is constant, rather it consists of intensity fluctuations. A typical example of the output power fluc-

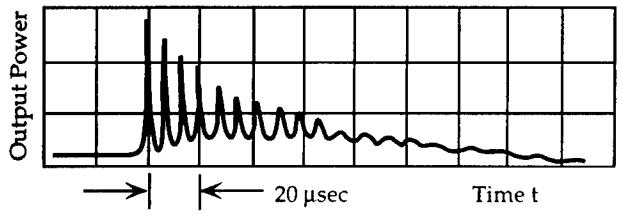


Fig. 10.18 Output power of a Nd:glass laser ($\lambda = 1.06$ μm) (Ref. 4).

tuations from a Nd:glass laser is shown in Fig. 10.18. This commonly observed transient phenomenon, which is known as spiking, can be described—at least approximately—with the rate equations.

Let us suppose that the pumping rate Q_2 is suddenly turned on from zero to a steady value $Q_2 \gg Q_{2,\text{th}}$. At the instant of turn-on, and for a short time thereafter, the upper level population N_2 increases and then passes the oscillation threshold value $N_{2,\text{th}}$. Recall that under steady-state conditions $N_2 = N_{2,\text{th}}$. However, under certain transient conditions, such as during initial start-up when $n \simeq 0$, N_2 can exceed its threshold value. The rate equation describing the initial buildup of spontaneous photons is

$$\frac{d}{dt}n(t) = KN_{2,T}(t) \ . \tag{10.97}$$

When one of these spontaneously created photons has produced an avalanche of about ten stimulated photons, then the initial growth of laser photons is governed by

$$\frac{d}{dt}n(t) = Kn(t)\,N_{2,T}(t) - \frac{n(t)}{\tau_c} \ . \tag{10.98}$$

Because $N_{2,T}(t)$ is, as a rule, a very large number (typically larger than $\sim 10^{14}$), and because at least during the early phase of buildup $KN_{2,T}(t) \gg (1/\tau_c)$, the number of photons will increase rapidly to a value considerably in excess of its steady-state value. With so many photons in the mode, the rate of depletion of the upper level population caused by stimulated emission may become considerably larger than the pumping rate Q_2. As a result, the population of the upper laser level may be depleted below the value at which oscillation can be sustained. Without further stimulated emissions, the number of photons within the resonator decreases now according to

$$\frac{dn(t)}{dt} = -\frac{n(t)}{\tau_c} \ , \tag{10.99}$$

and lasing can no longer be sustained. At this point, the cycle repeats itself, that is, the pumping mechanism builds up the population of level 2 until the round-trip gain exceeds unity and the laser breaks once more into oscillation. Several such cycles are shown in Fig. 10.19. From observations we know that

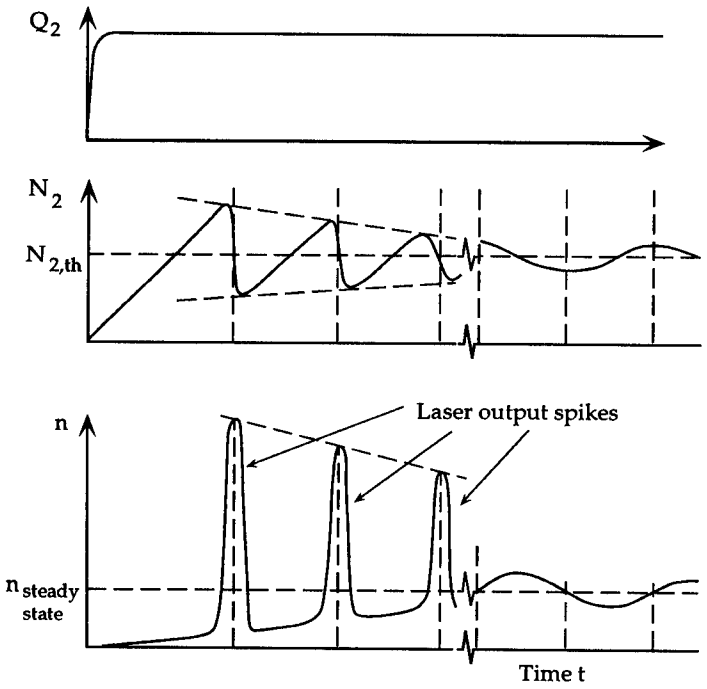


Fig. 10.19 Relaxation oscillations of an ideal laser oscillator when the pumping rate Q_2 is suddenly turned on at time $t = 0$. The resulting population density N_2 of level 2 and the resonator photon number n are shown as functions of time.

the intensity fluctuations are slowly damped with the spikes becoming successively smaller and with $N_{2,T}(t)$ and $n(t)$ approaching their steady-state values. It is possible to show that when the photon lifetime is

$$\tau_c > \frac{4(\chi - 1)}{\chi^2}\tau_2 , \tag{10.100}$$

where χ equals $Q_2/Q_{2,\text{th}}$ and τ_2 is the lifetime of the upper laser level, a small perturbation in the number of photons is rapidly damped and no spiking will be observed. When such lasers are turned on, both n and N_2 rise smoothly to their steady-state value. However, when

$$\tau_c < \frac{4(\chi - 1)}{\chi^2}\tau_2 , \tag{10.101}$$

fluctuations in the output power are observed. As an example, consider a Nd:glass laser with the following parameters:

$$\tau_2 = 1.6 \times 10^{-4}\ \text{s}$$

$$\tau_c \simeq 10^{-8}\ \text{s}$$

$$\chi \simeq 2.$$

Equation (10.101) indicates that power fluctuations will be observed.

A simple theory based on the rate equations developed in this section gives results that are in good agreement with the experimental behavior of a four-level system, such as the Nd:glass laser. In the ruby laser, which is an example of a three-level system, the appropriate rate equations differ somewhat, and measurements have shown the presence of strong and nearly undamped spiking in the laser output. In practice, the relaxation oscillations are a troublesome nuisance for most applications where repeatability and control of the intensity as a function of time are important. In the next section we describe a technique that makes it possible to eliminate the spiking phenomenon and, at the same time, shorten the time duration of oscillation and greatly increase the peak output power of the laser.

10.3.9 Q-Switching of Lasers

Q-switching is a method used to control both the time duration of laser oscillation and the pulse shape of the laser's output power. A typical Q-switched pulse is shown in Fig. 10.20. Q-switching can be considered as a type of stimulated single-spike behavior similar to that described in the previous section. It is generally accomplished by inserting an ultrafast optical shutter such as a Kerr cell, Pockels cell, or saturable absorber between the laser medium and one of the resonator mirrors. While the laser medium is pumped very strongly with a pulsed flashlamp, the optical shutter is closed. Toward the end of the pumping flash, when a very large population inversion has been built up in the laser medium, the shutter is suddenly opened. At this instant, the round-trip gain in the resonator is much greater than unity, and laser oscillation builds up much more rapidly than it would under normal operation. As a result, an avalanche of stimulated downward transitions occurs and the laser emits the available stored energy in a very short time. Typical pulses are between 10 and 75 ns in duration and have peak powers that vary from 10 MW to well over 1 GW (10^9 W).

The term Q-switching comes from the practice of describing resonators in terms of a quality factor Q and dates back to the early days of radio engineering. The Q-value is defined by the relation

$$Q = 2\pi f \frac{\text{average energy stored within resonator}}{\text{power dissipated by resonator}} . \tag{10.102}$$

The average energy stored is nhf and the power dissipated is nhf/τ_c. Therefore, $Q = 2\pi f\tau_c$. Typical values of Q for laser resonators are 10^8 to 10^9.

Because the evolution and decay of the pulse is completed in a time that is typically around 5×10^{-8} s, the effects of spontaneous relaxations of the upper and lower laser level populations as well as the pumping of these levels is negligible. With these simplifying assumptions, the photon rate equation becomes

$$\frac{dn}{dt} = n(t)\left[K(N_2 - N_1)_T - \frac{1}{\tau_c}\right] . \tag{10.103}$$

For convenience, we will from now on drop the subscript T. However, it must be remembered in what follows that the quantity $(N_2 - N_1)$ refers to the total population difference. Using Eq. (10.77) and measuring time in units of τ_c, we write Eq. (10.103) as

$$\frac{dn}{d\tau} = n(\tau)\left[\frac{(N_2 - N_1)}{(N_2 - N_1)_{\text{th}}} - 1\right] . \tag{10.104}$$

The first term on the right side gives the number of photons generated by stimulated emission per interval of normalized time $\tau = t/\tau_c$, and the second term accounts for the number of photons lost because of absorption and output coupling. As long as the first term is larger than the second term, the intensity of the Q-switched pulse increases, and when the terms are equal in value, the intensity is at its peak.

In a similar manner one obtains the rate equation for the total population:

$$\frac{d}{d\tau}(N_2 - N_1) = -2n(\tau)\left[\frac{(N_2 - N_1)}{(N_2 - N_1)_{\text{th}}}\right] . \tag{10.105}$$

Equations (10.104) and (10.105) are a pair of coupled differential equations. They can be readily solved by numerical methods. A typical result is shown in Fig. 10.20(a) where the normalized photon number $2n/(N_2 - N_1)_{\text{th}}$ is plotted

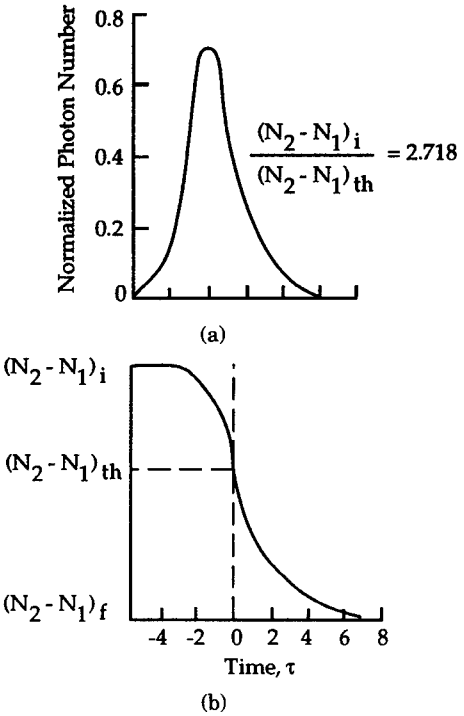


Fig. 10.20 Numerical solution of rate equation for Q-switching. Part (a) shows the normalized number of photons $[2n/(N_2 - N_1)_{\text{th}}]$ as a function of time. Time is measured in units of photon lifetime (after Ref. 5). Part (b) shows the corresponding population inversion during Q-switching.

as a function of normalized time. The corresponding population inversion is shown in Fig. 10.20(b).

As can be seen from Eq. (10.104), the photon buildup rate depends on the initial population inversion $(N_2 - N_1)_i$; the larger the ratio $(N_2 - N_1)_i/(N_2 - N_1)_{th}$, the shorter the rise time of the Q-switched pulse. The photon number, which is initially approximately equal to the number of spontaneously emitted photons within the mode volume, reaches its peak value when $dn/d\tau = 0$. From Eq. (10.104) we see that this occurs when the population inversion has plunged to its threshold value. Stimulated emission ceases altogether when $(N_2 - N_1) = 0$. From this moment on, the photon number decays exponentially with a decay time τ_c. Numerical solutions of the pair of coupled rate equations show that, for $(N_2 - N_1)_i \gg (N_2 - N_1)_{th}$, the rise time of the Q-switched pulse is short compared to τ_c, but the decay time of the pulse is very nearly equal to τ_c. For $(N_2 - N_1)_i$ just slightly larger than $(N_2 - N_1)_{th}$, the pulse duration is approximately four to six times the photon lifetime.

In addition to the pulse duration, the other quantities that are of principal interest are the total pulse energy and peak output power. The peak pulse power is given by

$$P_p = \frac{hf}{2\tau_c}\left\{(N_2 - N_1)_{th} \ln\frac{(N_2 - N_1)_{th}}{(N_2 - N_1)_i} + [(N_2 - N_1)_i - (N_2 - N_1)_{th}]\right\} . \quad (10.106)$$

For many Q-switched lasers, $(N_2 - N_1)_i \gg (N_2 - N_1)_{th}$, and the peak pulse power is then approximately given by

$$P_p \simeq \frac{(N_2 - N_1)_i}{2}\frac{hf}{\tau_c} . \quad (10.107)$$

The total radiative energy produced within the resonator is

$$E_T = \frac{1}{2}[(N_2 - N_1)_i - (N_2 - N_1)_f]\, hf . \quad (10.108)$$

Not all of this is useful output energy. Some of the total energy E_T will be lost because of scattering, diffraction, and absorption by the mirrors, as well as other optical elements within the resonator. Nevertheless, it is possible to show that for well-designed Q-switched lasers, and $(N_2 - N_1)_i \gg (N_2 - N_1)_f$, the energy of a Q-switched pulse is

$$E_p \simeq \frac{1}{2}(N_2 - N_1)_i\, hf . \quad (10.109)$$

Finally, the increase in peak output power when a laser with a known cw output power P_{cw} is Q-switched is

$$\frac{P_p}{P_{cw}} \simeq \frac{\tau_2}{2\tau_c} . \quad (10.110)$$

For CO_2 lasers $(\tau_2/2\tau_c) \sim 10^3$, while for ruby lasers $(\tau_2/2\tau_c) \sim 10^5$. Both of these examples illustrate the immense increases in peak output power when lasers are Q-switched. Because of the extremely high output powers and the reproducibility of the pulse shape, Q-switched lasers are used in applications that demand very high optical powers. For instance, certain nonlinear optical effects can only be observed with Q-switched pulses. Other uses of Q-switched pulses include drilling, plasma heating and diagnostics, ranging, remote sensing, initiation of chemical reactions, and the study of material properties, to name just a few.

10.4 OPTICAL RESONATORS AND GAUSSIAN BEAMS

10.4.1 Introduction

For a laser to function properly it is necessary to include a suitable feedback element in addition to the gain medium. This optical feedback element that sustains the lasing process by directing stimulated photons back and forth through the gain medium consists of a pair of precisely aligned plane or spherical mirrors centered on the optical axis of the gain medium.

The geometry of the mirrors and their separation determine the configuration of the electromagnetic field within the resonator. The stationary electromagnetic field configuration that satisfies both Maxwell's equations and the boundary conditions is a mode of the optical resonator. A typical resonator can support many transverse electromagnetic modes (TEM). By suppressing the gain of the higher order modes, the laser can be made to lase in a single fundamental mode, the TEM_{00} mode.

Depending on the distance between the mirrors and their radii of curvature, any given laser resonator is either stable or unstable. In the geometrical optics sense, this stability (or lack of it) is associated with the boundedness of the transverse displacement of a ray trajectory between the mirrors. Specifically, paraxial ray bundles propagating back and forth between the mirrors of a stable resonator are repeatedly refocused and the optical energy remains in the resonator. In unstable resonators, the ray bundles leave the resonator after only a few traversals. Consequently, a sizable fraction of the optical energy escapes from unstable resonators by spilling past the edges of the mirrors. In a stable resonator, the output beam is generally obtained by making one of the mirrors partially transmitting, and spillage of the beam past the mirrors is an undesirable loss. In contrast, the output beam from an unstable resonator is the radiation that spills past the edge of the mirror. This is why the near-field output beam from an unstable resonator usually has an annular shape.

Stable resonators are used in virtually all low-power lasers, and their modes are characterized by slender, threadlike beams with Hermite-Gaussian intensity patterns. Because of their small mode volume, they are not suitable for high-energy lasers. In contrast, unstable optical resonators have non-Gaussian modes whose intensity distributions and phase fronts are extremely difficult to find analytically. Nonetheless, the unstable resonator's characteristics of large mode volume, good transverse mode selection, and convenience of output

coupling with all-reflective optics have made this resonator type a prime candidate for high-energy lasers.

10.4.2 Modes of Stable Optical Resonators

To a casual observer the laser appears to emit a threadlike beam with a well-defined irradiance pattern. Mapping of the beam irradiance will generally show that the beam consists of a single bright spot with a transverse intensity distribution that is approximately Gaussian in every cross section of the beam. Such a Gaussian irradiance distribution is shown in Fig. 10.21. A beam with this profile is emitted when the laser oscillates in the fundamental TEM_{00}, which is also known as the lowest order mode. Oscillation of higher order modes causes the beam to break up into an array of sub-beams. This is most often observed when the laser tube or rod diameter is at least four times larger than the e^{-2} radius w of the fundamental mode.

Each resonator mode, aside from satisfying the scalar wave equation, must meet two requirements. First, the total phase shift of a wave front in propagating from one mirror to the other mirror and back must equal $q2\pi$; where q, the axial (or longitudinal) mode index, is the number of half wavelengths along the resonator axis. Second, the transverse field distribution within the resonator must reproduce itself after repeated reflections from the mirrors. Each self-reproducing transverse field configuration is referred to as a transverse mode. Thus, there are two types of modes. The axial modes determine the laser oscillation frequency, whereas the transverse modes determine the intensity pattern, beam divergence, and maximum output power. A few of the lower order transverse modes are illustrated in Fig. 10.22.

The second requirement can only be met if the mirror curvatures are exactly matched to the beam phase fronts, and if the mirror diameters $2a_1$ and $2a_2$ are much larger than the beam radii $w(z_1)$ and $w(z_2)$. Only then will the mirrors reflect the transverse field in such a way that it can reproduce itself.

In terms of rectangular coordinates with the resonator orientation shown in Fig. 10.23, the normalized complex field distribution of the TEM_{mn} mode is

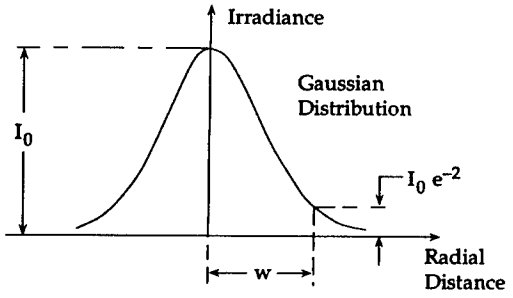


Fig. 10.21 Gaussian irradiance distribution of the TEM_{00} mode. The spot size (or radius) w is the distance from the beam center to the point where the irradiance is I_0e^{-2}.

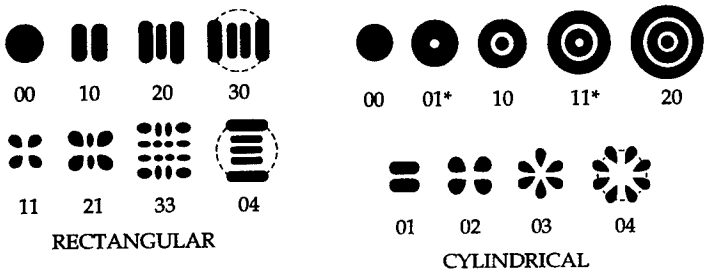


Fig. 10.22 Various lower order cylindrical and rectangular transverse mode patterns. The dotted circles indicate the beam's radius. The TEM_{01^*} (doughnut mode) is a linear combination of the TEM_{10} and TEM_{01} rectangular modes. The asterisk indicates a mode that arises from the linear superposition of two like modes, one rotated 90 deg about the z axis relative to the other.

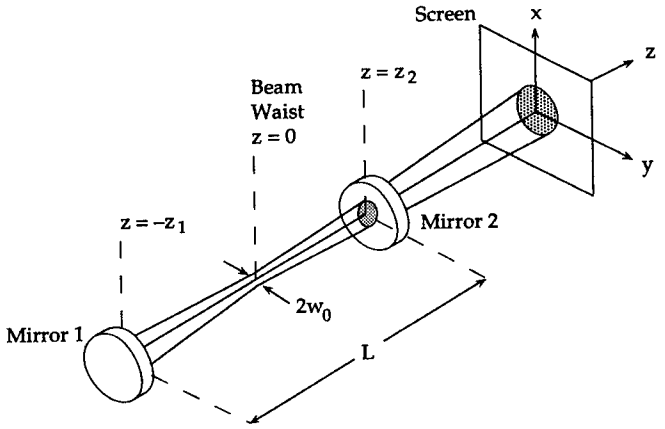


Fig. 10.23 Stable resonator and its orientation with respect to a rectangular coordinate system. The z axis coincides with the resonator axis, and x and y are distances measured in the plane normal to the resonator axis.

$$\begin{aligned}\tilde{U}_{mn}(x,y,z) = &\left(\frac{2}{2^{m+n}m!n!\,\pi}\right)^{1/2}\frac{1}{w(z)}H_m\left[\frac{2x}{w(z)}\right] \\ &\times H_n\left[\frac{2y}{w(z)}\right]\exp\left[-\frac{x^2+y^2}{w^2(z)}\right] \\ &\times \exp\left[-\frac{ik}{2}\frac{(x^2+y^2)}{\mathcal{R}(z)}\right] \\ &\times \exp\{-i[kz-(m+n+1)\phi(z)]\}\ .\end{aligned} \tag{10.111}$$

The transverse mode indices m and n may take on the values 0, 1, 2, ... The various parameters of Eq. (10.111) are determined by the following formulas:

$$H_n(x) = (-1)^n \exp(x^2) \frac{\partial^n}{\partial x^n} \exp(-x^2) \tag{10.112}$$

$$w(z) = w_0 [1 + (z/z_R)^2]^{1/2} \tag{10.113}$$

$$\mathcal{R}(z) = z[1 + (z_R/z)^2] \tag{10.114}$$

$$\phi(z) = \tan^{-1} (z/z_R) \tag{10.115}$$

$$k = 2\pi/\lambda \ .$$

The quantity $z_R = (1/2)kw_0^2$ is known as the Rayleigh range. It is the distance from $z = 0$ to $z = z_R$, where the beam radius has increased by $(2)^{1/2}$ over the smallest radius w_0 at $z = 0$. The parameters H_m and H_n are the Hermite polynomials of order m and n, respectively. The integers are sometimes known as the node numbers because the TEM_{mn} mode has m nodes along the x direction and n nodes along the y direction. The characteristic transverse dimension of the beam at any position z is given by Eq. (10.113), where $w(z)$ is the e^{-2} radius.

The radius of curvature $\mathcal{R}(z)$ of the wave front is determined by Eq. (10.114). It varies from $\mathcal{R} = \infty$ at $z = 0$ where the phase front is plane to a minimum value of $\mathcal{R}_{\min} = 2z_R$ at $z = z_R$; and when $z \gg z_R$, $\mathcal{R}$ approaches z asymptotically. This means that at sufficiently large distances from the beam waist the wave front's center of curvature is located essentially at the waist. Finally, $\phi(z)$ as defined by Eq. (10.115) is a phase shift relative to an ideal uniform plane wave whose free-space phase shift along the z axis is simply kz for propagation over the distance z.

The parameters H_m, H_n, $w(z)$, $\mathcal{R}(z)$, and $\phi(z)$ are important because they determine the characteristics of the laser beam. For this reason we will discuss them separately below.

10.4.3 Transverse Modes

The transverse aspects of laser modes can be more readily visualized by multiplying Eq. (10.111) by its complex conjugate $\tilde{U}^*_{mn}(x,y,z)$. The intensity pattern is then given by

$$|\tilde{U}_{mn}(x,y,z)|^2 = \frac{2}{2^{m+n}} \frac{1}{m!n!\pi} \frac{1}{w^2(z)} H_m^2\left[\frac{\sqrt{2}x}{w(z)}\right] H_n^2\left[\frac{\sqrt{2}y}{w(z)}\right]$$

$$\times \exp[-2(x^2 + y^2)/w^2(z)] \ . \tag{10.116}$$

This result, together with Eq. (10.112), describes the beam profile of any given transverse mode. For example, according to Eq. (10.112) the first four Hermite polynomials are $H_0(x) = 1$, $H_1(x) = 2x$, $H_2(x) = 4x^2 - 2$, and $H_3(x) = 8x^3 - 12x$. The corresponding beam profiles of these four lowest order modes are illustrated in Fig. 10.24.

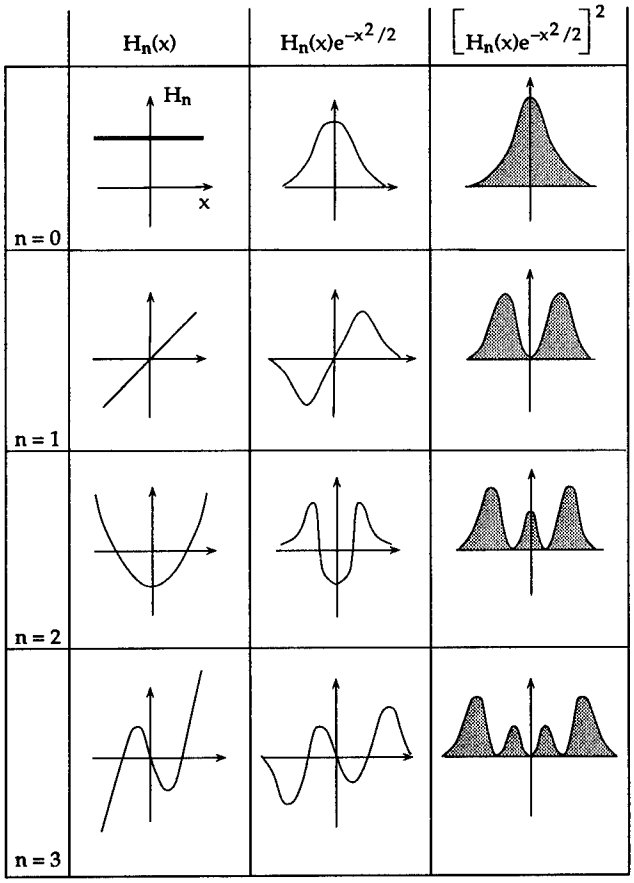


Fig. 10.24 Determination of the beam profiles of the four lowest order Hermite-Gaussian modes. The Hermite polynomials are plotted in the left column, the corresponding field distributions are shown in the center column, and the intensity profiles are in the right column.

For the TEM_{00} mode the fraction of the total power P_0 transmitted through a circular aperture of radius a is given by

$$\frac{P(a)}{P_0} = 1 - \exp[-2a^2/w^2(z)] \ . \tag{10.117}$$

Therefore, an aperture whose radius a is equal to the spot size $w(z)$ will transmit 86.5% of the total beam power, and when $a = 1.5\ w(z)$, essentially all of the beam power is transmitted. The transmitted fractional beam power as a function of aperture size for several TEM_{mn} modes is shown in Fig. 10.25. This figure shows that a higher order mode extends farther out in the transverse direction and has more of its energy at a greater distance from the axis than does any lower order mode. This fact is often used to suppress the oscillation of high-order modes by using an adjustable iris within the resonator.

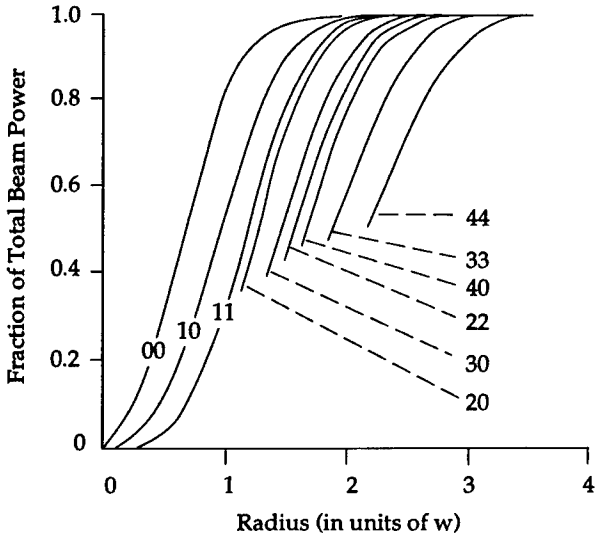


Fig. 10.25 Fraction of total beam power for some lower order rectangular modes within a circular spot. The radius is in units of the fundamental mode (TEM_{00}) spot size $w(z)$. Because of symmetry, the TEM_{mn} and TEM_{nm} values are represented by the same curve.[6]

The Hermite-Gaussian modes can also be expressed in terms of the cylindrical coordinates (r,θ,z). These modes are distinguished by their mode numbers p and l. The transverse mode numbers p and l measure the nodes in the r and θ coordinates, respectively. Some low-order cylindrical mode patterns are illustrated in Fig. 10.22. These modes are not commonly observed in practice unless the resonator is carefully adjusted to be axisymmetric.

10.4.4 Stability Condition and Diffraction Losses

For a self-reproducing mode to exist within a resonator, the phase fronts of the beam must match the mirror surfaces. This is the same as demanding that the radius of curvature of the phase front $\mathcal{R}$ be equal to the radius of the curvature R of the mirror. Thus,

$$R_1 = -\mathcal{R}(z_1) = z_1 + (z_R^2/z_1) \ , \tag{10.118}$$

$$R_2 = \mathcal{R}(z_2) = z_2 + (z_R^2/z_2) \ , \tag{10.119}$$

where

$$|z_1| + |z_2| = L \ , \tag{10.120}$$

and we have adopted the following sign convention. The radius of curvature is defined as positive if the mirror is concave with respect to the interior of the resonator, and negative if the mirror is convex. According to this convention each mirror in Fig. 10.26 has a positive radius of curvature. For waves, $\mathcal{R}$ is

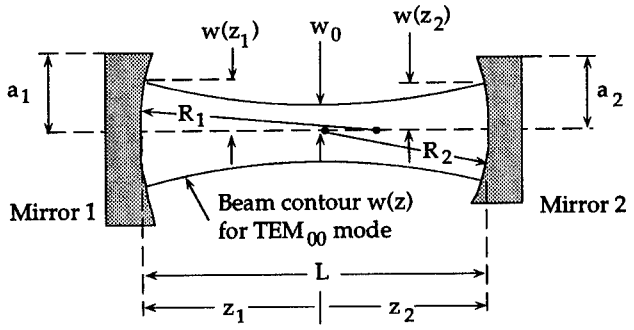


Fig. 10.26 Cross-sectional view of a typical stable resonator. The phase fronts of a self-reproducing field must match the mirror surfaces.

taken as positive if the center of curvature is to the left of the wave front, and negative if it is to the right.

Equations (10.118), (10.119), and (10.120) must now be solved for z_1, z_2, and z_R in terms of the resonator parameters R_1, R_2, and L. The results are

$$z_1 = -g_2(1 - g_1)G \tag{10.121}$$

$$z_2 = +g_1(1 - g_2)G \ , \tag{10.122}$$

$$z_R^2 = (kw_0^2/2)^2 = g_1g_2(1 - g_1g_2)G^2 \ , \tag{10.123}$$

where

$$g_1 = 1 - \frac{L}{R_1} \ , \qquad g_2 = 1 - \frac{L}{R_2} \tag{10.124}$$

and

$$G = \frac{L}{(g_1 + g_2 - 2g_1g_2)} \ . \tag{10.125}$$

The parameters g_1 and g_2 occur repeatedly in resonator analysis and are generally referred to as the g parameters. Equations (10.121) and (10.122) locate the waist and Eq. (10.123) determines the e^{-2} beam radius w_0 at the waist. The beam dimensions of even greater interest are the radii $w(z_1)$ and $w(z_2)$ at the mirrors of the resonator. In terms of the resonator g parameters they are

$$w(z_1) = \left[\left(\frac{L\lambda}{\pi}\right)^2 \frac{g_1g_2}{g_1^2(1 - g_1g_2)}\right]^{1/4} \ , \tag{10.126}$$

$$w(z_2) = \left[\left(\frac{L\lambda}{\pi}\right)^2 \frac{g_1g_2}{g_2^2(1 - g_1g_2)}\right]^{1/4} \ . \tag{10.127}$$

These two equations show that the beam radii at the resonator mirrors are finite only if $0 < g_1 g_2 < 1$. This inequality is the well-known resonator stability condition that is expressed graphically by the stability diagram of Fig. 10.27. The g_1, g_2 plane is divided into stable and unstable regions and any given optical resonator corresponds to a point on the plane. The stable region, darkened in Fig. 10.27, is defined by the condition

$$0 \leqslant g_1 g_2 \leqslant 1 \ . \tag{10.128}$$

It is separated from the unstable region by the hyperbola $g_1 g_2 = 1$ and by the lines $g_1 = 0$ and $g_2 = 0$.

When a resonator structure is such that its g parameters lie within the shaded region, the transverse mode dimension is approximately given by $(L\lambda/\pi)^{1/2}$, which for typical lasers, oscillating in the visible region, is of the order of 1 mm or less. Because the mirror diameters, or any other aperture in the laser structure, are then readily made to be at least three to four times larger than the largest spot size within the resonator, the diffraction losses of the system are very small. On the other hand, when the resonator g parameters lie on the boundary of the stable region or in the unstable region, the transverse mode dimension exceeds all practical mirror diameters, and the diffraction losses of the resonator are several orders of magnitude larger.

It is possible to obtain a rough estimate of the diffraction losses by analyzing a resonator with plane circular mirrors of identical diameters (see Fig. 10.28). A plane wave with a uniform field propagates from mirror 1 to mirror 2. Because of diffraction the wave will have spread when it arrives at mirror 2. The amount of spreading is given by the angle $\beta \simeq \lambda/2a$. The fractional power δ that spills over the rim of mirror 2 is then

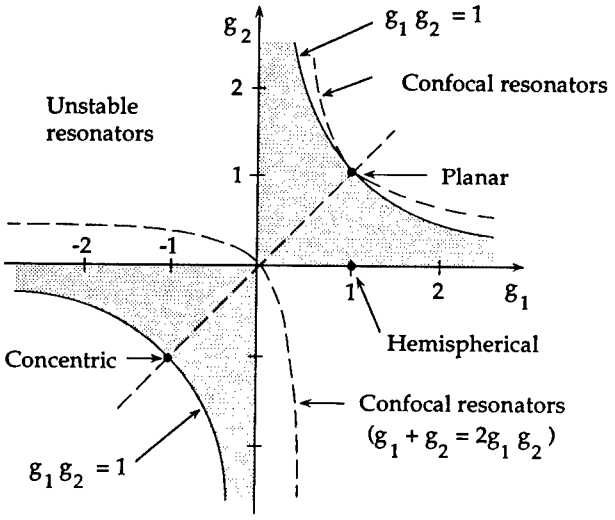


Fig. 10.27 The stability diagram. Stable resonators lie in the shaded region. All symmetric resonators lie on the dashed line with the 45-deg slope. The location of some of the more common resonators are also shown.

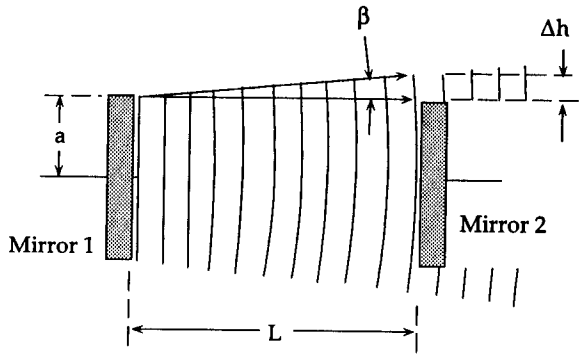


Fig. 10.28 Cross-sectional view of resonator used to calculate the diffraction loss of a uniform plane wave.

$$\delta = \frac{\text{area of annulus}}{\text{area of mirror}} \simeq \frac{2\Delta h}{a} , \tag{10.129}$$

where $\Delta h = \beta L \simeq (\lambda L/2a)$ is the width of the annulus. Hence,

$$\delta \simeq \lambda L/a^2 = N^{-1} . \tag{10.130}$$

The dimensionless resonator parameter N in Eq. (10.130) is known as the Fresnel number, which is of considerable importance in resonator analysis. Fresnel numbers of typical laser resonators vary somewhere between ~ 50 for HeNe lasers to ~ 100 for ruby or Nd:glass lasers. The fractional power lost by diffraction per one-way pass is quite small for most common lasers. The results of detailed computer calculations by Fox and Li[7] of the diffraction losses for two resonator configurations are shown in Fig. 10.29. Equation (10.130) predicts a power loss that is larger than the more accurate values calculated by Fox and Li[7] even for the highest loss resonator with plane circular mirrors. The reason for this discrepancy is due to our earlier assumption of a uniform field. With such a field, the fractional power spilling over the mirror rim is larger than that of a Gaussian field distribution where the energy is mostly concentrated at the center region of the mirror. In this connection we should also note that for a given Fresnel number, the amplitude of higher order transverse modes is larger at the edge than it is for the TEM_{00} mode. The spillover loss should therefore be higher for higher order modes. This is indeed confirmed by the loss curves in Fig. 10.29 labeled TEM_{10} and TEM_{20}.

10.4.5 Frequencies of Stable Resonator Modes

One of the conditions imposed on a resonator mode is that the round-trip phase shift must be equal to an integer multiple of 2π. The axial phase of the TEM_{mn} mode is from Eq. (10.111), given by

$$\Phi(z) = kz - (m + n + 1)\,\phi(z) , \tag{10.131}$$

and the one-way axial phase shift is

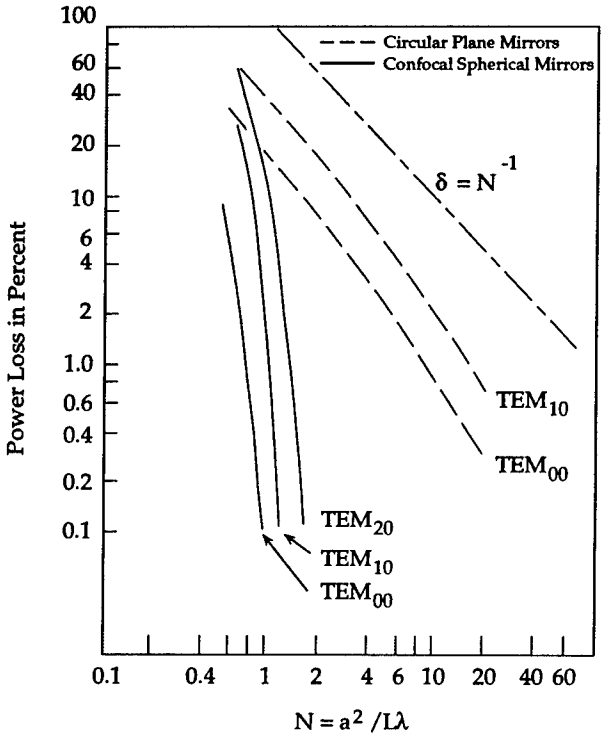


Fig. 10.29 Power loss per one-way pass in percent as a function of the mirror size in units of $a^2/L\lambda$ for two resonator configurations. The diffraction loss δ for a uniform plane wave is also shown.[7]

$$\Phi(z_2) - \Phi(z_1) = \frac{2\pi L}{c} f_{mnq} - (m + n + 1)\,[\phi(z_2) - \phi(z_1)] = q\pi \ . \tag{10.132}$$

Therefore, the resonant frequency of the q'th axial mode in the mn'th transverse mode pattern is

$$f_{mnq} = \left(q + (m + n + 1)\frac{\cos^{-1}\sqrt{g_1 g_2}}{\pi}\right)\frac{c}{2L} \ , \tag{10.133}$$

where

$$\phi(z_2) - \phi(z_1) = \cos^{-1}\sqrt{g_1 g_2} \ . \tag{10.134}$$

The resonant frequencies of several lower order modes of a near planar ($R \gg L$) cavity are shown in Fig. 10.30(a). The axial mode frequency spacing is given by

$$\Delta f_q = |f_{mn(q+1)} - f_{mnq}| = \frac{c}{2L} \ , \tag{10.135}$$

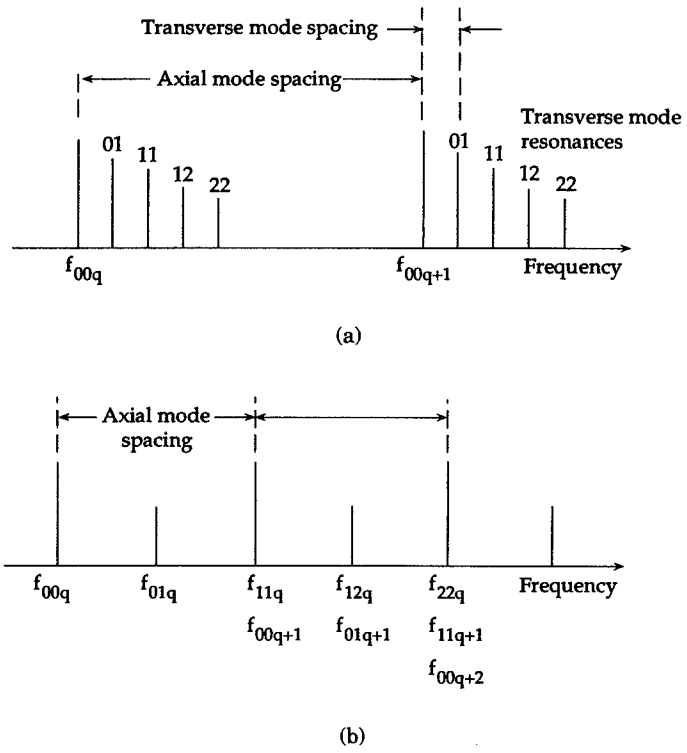


Fig. 10.30 Resonant frequencies of some lower order modes. (a) For the near-planar resonator, the transverse mode spacing is much less than the axial mode spacing. (b) The confocal case represents the extreme situation where the axial mode spacing is two times the transverse mode spacing.

and the transverse mode frequency spacing is

$$\Delta f_{mnq} = f_{(m+1)nq} - f_{mnq} = (\cos^{-1}\sqrt{g_1 g_2})\frac{c}{2L\pi} . \tag{10.136}$$

Thus, for the planar resonator with $g_1 = g_2 = 1$, $\Delta f_{mnq} = 0$. If the spacing between the mirrors is held constant but the radius of curvature is gradually decreased, starting from $R = \infty$, the frequency difference between two neighboring transverse modes (for q = constant) increases. This means that the groupings of the higher order mode frequencies in Fig. 10.30(a) expand toward the right, while the axial mode intervals remain the same. When $R_1 = R_2 = L$ (confocal resonator), then $g_1 = g_2 = 0$ and $\cos^{-1}(g_1 g_2)^{1/2} = \pi/2$, and the frequency of the TEM_{mnq} mode becomes

$$f_{mnq} = (1/2)\,(2q + m + n + 1)\frac{c}{2L} . \tag{10.137}$$

Because many combinations of q, m, and n can give the same value for $(2q + m + n + 1)$, a number of different transverse and longitudinal modes can

oscillate at the same frequency. For example, the TEM_{00q} mode oscillates at the same frequency as the $TEM_{11(q-1)}$. Other examples are shown in Fig. 10.30(b). As the radii of curvature of the mirrors approach the value $L/2$ (concentric resonator), the expanded groupings of the higher order mode frequencies now contract toward the left, and the mode spectrum for the near-concentric resonator begins to resemble the near-planar case with which we started.

In summary, the complete specification of any given mode of stable optical resonators is TEM_{mnq}, where $m,n = 0, 1, 2, 3, \ldots$ refer to the intensity nulls in the x,y directions and specify transverse modes, and $q \simeq 10^5 - 10^7$ gives the number of half wavelengths in the axial direction. In general, the frequency of any given mode depends on m, n, and q, as well as the mirror separation and radii of curvature of both resonator mirrors.

10.4.6 Beam Spreading

The amount of spreading of the beam as it propagates along the z direction is obtained from Eq. (10.113). By definition, $w(z)$ is the radial distance at which the beam irradiance has decreased to $(1/e^2)$ of the on-axis value for the TEM_{00} mode. The beam contracts to a minimum diameter $2w_0$ at $z = 0$. At this point on the z axis is the location of the beam waist. The beam contour $w(z)$, which expands with distance z from the waist, is plotted in Fig. 10.31, where for the sake of illustration, the transverse dimension is considerably exaggerated.

At large distances from the waist ($z \gg \pi w_0^2/\lambda$) the beam diverges linearly with distance. The far-field divergence angle θ (i.e., the full angle) for the TEM_{00} mode is by definition

$$\theta_{00} = \lim_{z\to\infty} [2w(z)/z] = \frac{1.27\,\lambda}{d_{00}}\,, \tag{10.138}$$

where $d_{00} = 2w_0$ is the beam waist diameter of the TEM_{00} mode. The divergence angle of a circular aperture of diameter d_{00} that is illuminated by a plane wave of uniform intensity is

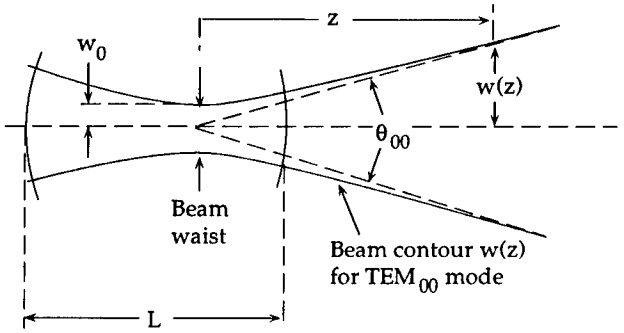


Fig. 10.31 Beam contour $w(z)$ for the TEM_{00} mode. The angle θ_{00} is the beam divergence angle. For clarity the transverse dimensions are considerably exaggerated.

$$\theta = \frac{2.44\,\lambda}{d_{00}} \,. \tag{10.139}$$

A comparison of these two equations reveals the interesting result that for beams with the same diameters, those beams with a Gaussian amplitude distribution diverge less than those with a uniform amplitude.

Because the ratios of different mode diameters are constant at any transverse plane inside or outside of the resonator, the beam divergence for a higher order TEM_{mn} mode is

$$\begin{aligned} \theta_{mn} &= \lim_{z\to\infty} \frac{2w_{mn}(z)}{z} \\ &= C_{mn}\,\theta_{00} \,, \end{aligned} \tag{10.140}$$

where $w_{mn}(z) = C_{mn}w(z)$, and C_{mn} is a constant to be determined.

The spot size of higher order modes may be defined as the transverse distance from the middle of the beam to the point where the irradiance is e^{-2} of the irradiance of the outermost peak of that mode. For the Laguerre-Gaussian (circularly symmetric) modes, the spot size $w_{pl} = C_{pl}w(z)$ is then simply the radius at which the irradiance is e^{-2} of the irradiance of the outermost peak of that mode. This is illustrated in Fig. 10.32 where the irradiance distribution of the three lowest order circularly symmetric modes is plotted. The values of C_{00}, C_{01^*}, C_{10}, C_{11}, C_{20}, and C_{21}, are 1.0, 1.5, 1.9, 2.1, 2.4, and 2.6, respectively.[8] For the three lowest order rectangular modes $C_{00} = 1$, $C_{10} = C_{01} = 1.5$, and $C_{20} = C_{02} = 1.9$. Defining the spot size and the divergence angle for modes of rectangular symmetry in this manner may not be adequate for those modes whose pattern is far from square. A more practical definition might be in terms of a radius that encircles 86% of the beam power.

In general, $C_{m+1,n+1} > C_{mn}$ (or $C_{p+1,l+1} > C_{pl}$) and the higher the mode order the larger the far-field beam divergence angle. The beam divergence

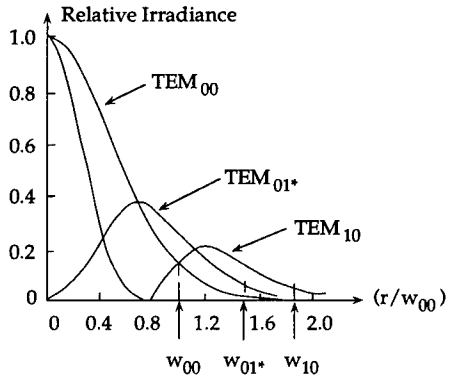


Fig. 10.32 Radial irradiance distribution of the three lowest order circularly symmetric modes. The radial distance is in units of the fundamental (TEM_{00}) spot size $w(z) = w_{00}$.

angle is one of the most important beam parameters because it determines the diameter of the focal spot.

10.4.7 Beam Transformation by a Lens

Many applications of laser beams require an extremely high-energy flux over a relatively small area. This is accomplished by focusing the beam. The focused spot is then simply the beam waist produced by either a positive lens or a concave mirror. The effects of a lens on the propagation of a Gaussian spherical beam are schematically illustrated in Fig. 10.33. The dimensions d_2 and w_2 are given by

$$d_2 - f = \frac{f^2 (d_1 - f)}{(d_1 - f)^2 + (\pi w_1^2/\lambda)^2} \tag{10.141}$$

and

$$w_2^2 = \frac{w_1^2}{\left(1 - \dfrac{d_1}{f}\right)^2 + (\pi w_1^2/f\lambda)^2} . \tag{10.142}$$

From Eq. (10.141) we see that the minimum spot size is generally not located at the geometrical focal plane of the lens; only when $d_1 \to \infty$ or $d_1 = f$ is the transformed beam waist located at the focal plane.

The diameter of the focal spot $2w_2$ is seen from Eq. (10.142) to be proportional to the focal length f (when $d_1 \simeq f$) and inversely proportional to both the distance d_1 and the spot size w_1. The entrance aperture of the lens fixes, of course, an upper limit on the value of w_1. Because for most practical applications

$$\left(1 - \frac{d_1}{f}\right) < (\pi w_1^2/f\lambda) ,$$

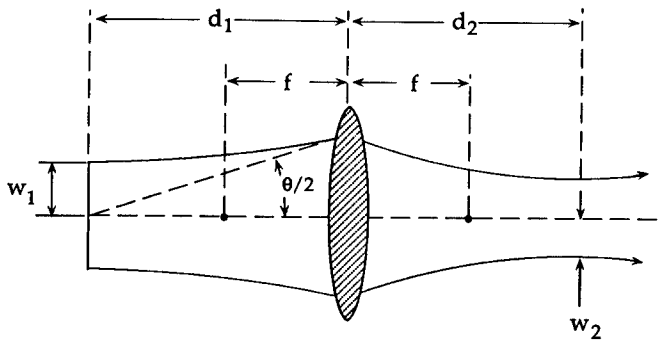


Fig. 10.33 Transformation of a Gaussian spherical beam by a thin lens. The diameter of the lens must be larger than $\sim 3w_1$.

Eq. (10.142) can be simplified so that

$$w_2^2 \cong f^2 (\lambda/\pi w_1)^2$$
$$= \left(\frac{f\theta_{00}}{2}\right)^2 , \tag{10.143}$$

where θ_{00} is the far-field beam divergence angle defined by Eq. (10.138).

Equation (10.143) gives an important and often-used relationship between the diameter of the focal spot, the focal length of the lens, and the divergence angle of the beam.

10.4.8 Unstable Resonators

A major goal in the design of high-energy laser resonators is the attainment of a large mode volume so that more of the excited laser medium can contribute energy to the radiation field. Unfortunately, stable resonators have a relatively small mode volume.

For high-energy lasers operating in the IR region of the spectrum, stable resonators also present a materials problem. The output beam from stable resonators is generally obtained by making the reflective surface of one of the resonator mirrors partially transmitting. This is most often done by depositing multiple layers of a dielectric substance on a substrate of quartz or similar material whose surface has been optically ground to the desired curvature. To avoid beam attenuation and possible thermal fracturing, the absorption coefficient of the substrate at the laser wavelength must be essentially zero. IR window materials that can meet this requirement are generally not available for use in high-energy lasers operating in the 3- to 10-μm spectral region.

Because of these shortcomings, stable optical resonators are generally not suitable for high-energy lasers. The unstable resonator on the other hand is not mode volume or material limited. This type of resonator can provide a large mode volume, a collimated output beam, excellent transverse mode control, variable output coupling, and can be constructed of all-reflective metallic mirrors that are cooled easily. For a detailed treatment of unstable resonators see Refs. 9, 10, and 11.

10.5 TYPES OF LASERS

10.5.1 Solid-State Lasers

The term *solid-state laser* applies to those lasers where the gain medium is a glass or crystalline material doped with an impurity ion; it does not include semiconductor lasers, which are described separately. The lasers considered here are optically pumped over a fairly broad band by means of a flashlamp or some other source of intense radiation. The impurity ion typically belongs to the transition metals or rare earth elements.

Besides being optically pumped, solid-state lasers have a few other properties in common. The bulk of the gain medium is the host that does not directly participate in lasing. The lasing ions are present in low concentration, generally 1% or less. The lasing transition is between states that belong to the

inner unfilled shells. These transitions are not strongly influenced by the host's background field and usually have no electric dipole moment. Therefore, the spontaneous decay time of the upper laser level is in the millisecond rather than in the nanosecond range. The wavelength of the emitted radiation, which is characteristic of the individual ion modified somewhat by the host, is typically in the red or near infrared. The laser material is usually in the form of a rod. Much of the pump energy is absorbed by the host and is converted to heat. The resulting temperature gradient within the rod changes the index of refraction of the rod and causes the rod to act similar to a variable focal-length lens. To minimize this effect, solid-state lasers are usually liquid cooled and operated close to around design power levels so the focal length remains relatively fixed.

The laser rod can be damaged by the beam both at the end surfaces (especially if they are not clean) and internally via self-focusing. This is a nonlinear process in which high-beam powers increase the refractive index, which in turn causes the beam to collapse into a thin threadlike filament whose irradiance may exceed the rod's damage threshold. Unlike gas and liquid lasers, solid-state lasers are not self-healing.

All pulsed solid-state lasers exhibit spiking in the output beam. The output consists of many irregular spikes of about 10^{-6} s in duration. While spiking is detrimental to most scientific applications where shot-to-shot reproducibility is required, it is usually desired for materials processing where a burst of power can aid in cutting or drilling. Continuous-wave-operated lasers and Q-switched lasers do not exhibit spiking; however, they may have a fluctuating output caused by mode locking, relaxation oscillations, or power-supply fluctuations.

Construction of cw operating solid-state lasers is more difficult because of the need to provide a sufficiently powerful pump radiation source that can operate continuously. Therefore, to attain cw operation it is necessary to increase the efficiency of the excitation process and to improve the cooling of the laser rod and arc lamps.

Solid-state lasers are most useful for the generation of powerful radiation pulses lasting for a millisecond to fractions of a nanosecond, containing energy ranging from a few millijoules to hundreds of joules per pulse. To increase the energy per pulse, amplifier stages are used. An amplifier stage consists of a laser rod and one or more flashlamps. The most common solid-state lasers along with performance characteristics are listed in Table 10.2.

10.5.2 Gas Lasers

The most common form of excitation in a gas laser is an electric discharge. The electrons in the discharge collide with and excite atoms or molecules that then participate in the lasing process or transfer their excitation energy to another species of laser active atoms or molecules. The lasing gas or gas mixture is usually confined to a glass or quartz tube. The discharge may be powered by a rf voltage, with electrodes placed on the tube externally, or alternatively by direct or low-frequency alternating current that is maintained between internal electrodes.

The ends of the glass tube are sealed with either the resonator mirrors or with optical windows oriented at the Brewster angle to eliminate reflections

Table 10.2 Solid-State Lasers

Laser	Active Medium	Wavelength (μm)	Output	Wall-plug Efficiency	Cooling	Typical Applications
Ruby	Chromium ions in pink ruby (Al_2O_3)	0.694	Pulsed with 0.01 to 4 pulses per second (pps)	0.1 to 0.5%, flashlamp pumped	Refrigerated water	Ranging Holography Materials working
Nd:YAG	Neodymium ions in yttrium aluminum garnet	1.064 and 1.3	cw or pulsed with 0.01 to 5×10^4 pps	0.1 to 3% if flashlamp pumped. 5 to 8% if diode pumped.	Refrigerated water	Materials working Target designation and ranging Medicine Inspection
Nd:glass	Neodymium ions in glass	1.06	Pulsed with 0.1 to 2 pps	1 to 5%, flashlamp pumped	Refrigerated water	Fusion research Materials working
Ti: sapphire	Titanium ions in sapphire (Al_2O_3)	Tunable from 0.665 to 1.13 with maximum output energy between 0.75 and 0.85 μm	cw and pulsed, up to 30 pps	0.01% (if pumped by argon ion laser) to 0.1% (if pumped with frequency-doubled Nd:YAG laser)	Refrigerated water	Semiconductor characterization Isotope separation Biomedical research
Alexandrite	Chromium ions in $BeAl_2O_4$ crystal (Chrysoberyl)	Tunable from 0.71 to 0.82 with peak at 0.755 μm	cw and pulsed, up to 50 pps	0.25%, flashlamp pumped	Air or water cooling	Medicine Lidar Laser spectroscopy
Nd:YLF	Neodymium-doped yttrium lithium fluoride	1.05 and 1.3	cw and pulsed	0.3% GaAs, diode pumped	Air	Fiber optic communication
Er:glass	Erbium-doped glass	1.54	Pulsed up to 5 pps	0.2%, flashlamp pumped	Water	Primarily used in "eyesafe" range finders
Er:YAG	Erbium-doped YAG	2.94	Pulsed at 25 pps	1.5%, flashlamp pumped	Refrigerated water	Medicine Biomedical research
Co:MgF_2	Cobalt-doped magnesium fluoride crystal	Tunable from 1.75 to 2.50	Pulsed at 10 pps	Pumped by 1.3-μm Nd:YAG laser	Water cooling	Remote sensing Medicine
Ho:YAG	Holmium-doped YAG	2.1	cw and pulsed at up to 20 pps	1.3%, flashlamp pumped	Refrigerated water	Medicine
F-center	Electrons trapped in alkali halides	Tunable 1.45 to 1.75, 2.3 to 3.45	cw and pulsed at up to 10^8 pps	$\leq$ 5% of pump laser	Liquid nitrogen	Molecular spectroscopy

from the window surfaces. In either case, the mirrors must be carefully aligned with the axis of the tube and with each other. The Brewster windows are reflection-free for only one polarization direction, and the laser will operate only in that polarization.

The achievement of a population inversion depends on the excitation rate and the decay rates of all levels involved in the cascading process. The processes contributing to decay of a single level are radiative processes and collisions with electrons, other atoms, and molecules, as well as with the walls of the tube. The decay rates depend on the gas composition, pressure, and dimensions of the tube. Because of the complexity and interrelationship of these phenomena, lasing can be achieved only in rather exceptional circumstances.

In addition to electric discharges, some gas lasers are pumped by electron beams, chemical reactions, or gas dynamic expansions of a heated gas mixture. Regardless of the excitation method used, the spectral width of the energy levels is considerably smaller than that in solids. For gases at the low pressures generally used in lasers, the linewidths are Doppler broadened and of the order of a few gigahertz or less. Because of the narrow spectral widths, optical pumping is not practical for gas lasers.

In what follows we briefly describe the operation and properties of a few representative examples of gas lasers that are likely to be of most interest. Some of the more frequently used gas lasers are listed in Table 10.3.

The best known gas laser is the HeNe laser with emissions in the green to red regions. The excitation process begins with collisions between discharge electrons and helium atoms. The collisions excite the helium into two metastable levels. The lifetimes of these levels are long because there exists no lower lying levels in helium to which transitions can occur. The excited helium eventually collides with unexcited neon, and because neon has two energy states that correspond closely to the helium metastables, the excitation energy is transferred from the helium to the neon during a collision. Whether an inversion will actually occur depends on the relative abundance of the helium and neon atoms in the mixture, on the excitation and decay rates that are determined by the electron energy distribution and the gas pressure, as well as the inside diameter of the discharge tube, which affects the collision frequency of excited atoms with the tube wall.

The HeNe laser is one example of neutral atom lasers that, aside from He, include the noble gases Ne, Kr, Ar, and Xe and metal vapor atoms such as Pb, Cu, Au, Ca, Sr, and Mn. With the notable exception of the HeNe laser, all neutral noble gas lasers oscillate in the IR (1 to 10 μm). The metal vapor lasers generally oscillate in the visible. Because they are self-terminating, they operate in the pulsed mode only.

In discharges with very large current densities it is possible to obtain lasing from the ions of noble gases. These lasers typically operate in the visible and UV regions. The most notable example is the argon ion (Ar^+) laser. Because both the upper and lower laser levels actually consist of many sublevels, the argon ion laser can produce a multicolored output. The excitation process involves numerous energetic electron collisions that move the argon ion through many energy states that are not coupled effectively to the upper laser level. The overall laser efficiency is therefore quite small, typically less than 0.1%. Because commercially available Ar^+ lasers have cw output powers of up to

20 W, active cooling removes the large amounts of wasted energy. Water cooling is used for the larger lasers, while forced-air cooling is adequate for the low-power lasers (<1 W).

A large family of gas lasers operates on transitions between the vibrational-rotational energy levels of molecules that are in the electronic ground state. The relatively small energy differences between the levels involved in these laser transitions produces radiation in the mid to far IR (3 to 300 μm). By far the most important example is the CO_2 laser, which generally uses a mixture of CO_2 (the lasing molecule), N_2, and He. The nitrogen aids in the excitation process, and the helium plays an important function in the de-excitation of the lower levels. This laser is tunable from about 9 to 11 μm. However, the most prominent emission occurs at 10.6 μm. The laser is tuned by using a diffraction grating in place of one of the cavity mirrors. Rotation of the grating causes the laser emission to jump from one rotational line to another. From a design point of view, the CO_2 laser is truly versatile in its construction. Depending on the desired output power and specific application, this laser can be separated into seven categories: (1) sealed-off tubes (between 1- and 100-W output power), (2) slow axial flow of the gases (up to about 100-W output power), (3) fast axial flow of the gases (up to about 1000-W output power), (4) waveguide lasers (less than 30-W output power), (5) transverse-flow lasers (a few kilowatts output power), (6) transversely excited atmospheric pressure (TEA) lasers (pulsed output with 10 to 50 J/l and pulse durations of several microseconds), and (7) gas-dynamic lasers (with hundreds of kilowatts of output power). Even though these lasers differ from each other in their construction and operating parameters, their operating (wall plug) efficiency can be as high as 20%. This high efficiency is due to a theoretical efficiency of $\sim 40\%$ and the selective pumping by the vibrationally excited nitrogen.

Another family of molecular lasers is the excimer[b] laser. These lasers operate primarily in the UV. The relevant laser transitions are between different electronic states with the ground state being repulsive. Therefore, after undergoing the laser transition, the molecule immediately dissociates and the lower laser level is always empty. As a rule these lasers are only operated in the pulsed mode with pulse durations in the tens of nanosecond range. The excited levels in which the molecules are formed through collisions have radiative lifetimes of 10^{-9} to 10^{-10} s. As a result, intense pumping is required to establish a population inversion. Pumping is accomplished with electron beams or pulsed transverse electric discharges with pump durations of 10 to 20 ns. Preionization is usually achieved with a row of UV sparks. The gas mixture generally consists of a rare gas atom (such as Ar, Kr, or Xe) and a halogen (such as fluorine or chlorine). Specific examples are ArF ($\lambda = 193$ nm), KrF ($\lambda = 248$ nm), XeCl ($\lambda = 309$ nm), and XeF ($\lambda = 351$ nm).

The excitation mechanism in chemical lasers is a chemical reaction between gaseous elements, and generally involves either an associative ($A + B \rightarrow AB$) or a dissociative ($XYZ \rightarrow X + YZ$) exothermal chemical reaction. The most notable examples of an associative reaction are the HF and DF lasers with multiple-wavelength lasing between about 2.5 to 4.5 μm. For the dissociative

[b]The word *excimer* is a contraction of the words *excited dimer*.

Table 10.3 Gas Lasers

Laser	Active Medium	Wavelength (μm)	Output	Wall-plug Efficiency	Output Power or Energy	Typical Applications
HeNe	Ne atoms in HeNe gas mixture in sealed tube	0.543, 0.594, 0.604, 0.633, 1.15, 1.52, 3.39	cw	≤0.1%	0.1 to 50 mW at 0.633 μm; ≤15 mW at 1.15 or 3.39 μm, ≈1 mW at other lines	Alignment Price code scanning Construction Holography Reprographics Measurement
HeCd	Ionized cadmium vapor mixed with He in sealed tube	0.325 or 0.442	cw	0.01 to 0.1%	1.5 to 10 mW at 0.325 μm, 2 to 50 mW at 0.442 μm	Microlithography Reprographics Spectroscopy Medicine Recording
Argon	Ionized argon in sealed tube	Several lines between 0.35 and 0.528, main lines 0.488 and 0.514	cw	0.01 to 0.1%	2 mW to 20 W	Laser light shows Recording Spectroscopy Dye laser pumping Medicine
Krypton	Ionized krypton in sealed tube	Several lines between 0.350 and 0.8; main line at 0.647	cw	≤0.05%	5 mW to 6 W	Multicolor light shows and displays Dyelaser pumping
ArF (excimer)	Gas mixture of argon and fluorine	0.193	Pulsed, 5 to 25 ns, at 1 to 10^3 pulses per second (pps)	≤1%	Up to 0.5 J/pulse	Research and development Spectroscopy Photochemistry Lithography
KrF (excimer)	Gas mixture of krypton and fluorine	0.248	Pulsed, 5 to 50 ns, 1 to 500 Hz	≤2%	Up to 1 J/pulse	Same as ArF laser

Table 10.3 *(continued)*

Laser	Active Medium	Wavelength (μm)	Output	Wall-plug Efficiency	Output Power or Energy	Typical Applications
XeCl (excimer)	Gas mixture of xenon and chlorine	0.308	Pulsed, 1 to 80 ns, 1 to 500 Hz	≤2.5%	Up to 1.5 J/pulse	Research and development Spectroscopy Photochemistry Lithography Dye pumping
XeF (excimer)	Gas mixture containing xenon and fluorine	0.351	Pulsed, 1 to 30 ns, 1 to 500 Hz	≤2%	Up to 0.5 J/pulse	Same as XeCl laser
HF (chemical)	Gas mixture containing H_2 and F_2	Several lines between 2.6 and 3	cw and pulsed, 50 to 200 ns, at 0.5 to 20 Hz	≤1%	Up to 150 W cw or 2 to 600 mJ/pulse	Atmospheric research Other research and development
DF (chemical)	Gas mixture containing deuterium and fluorine	Several lines between 3.6 and 4	cw and pulsed, 50 to 200 ns, 0.5 to 20 Hz	≤1%	Up to 100 W cw or 2 to 600 mJ/pulse	Same as HF laser
CO_2	Flowing or sealed gas mixture of CO_2, N_2, and He	9 to 11 or 10.6	cw and pulsed with wide range of pulse durations and pulse rates	Up to 15%	Up to 15 kW and 150 J/pulse	Research and development Surgery Materials working Photochemistry Laser radar Remote sensing

reaction, the most notable example is the atomic iodine laser, in which excited atomic iodine is produced by dissociating CH_3I, CF_3I, or C_3F_7I with UV photons from a flashlamp. The atomic iodine lases at 1.315 μm.

Chemical lasers are somewhat unique because they convert chemical energy directly into laser radiation, which they can do without the use of an external electric power supply. The reaction begins when the reagents are mixed in the gain region. In general, however, chemical lasers require some form of electrical input to initiate the chemical reaction. One initiating technique involves a high-energy electron beam. Another technique makes use of an arc jet heater to provide free fluorine atoms by thermal dissociation of molecular fluorine (F_2) or sulfurhexafluoride (SF_6). The atomic fluorine gas is next cooled by being expanded through a row of closely spaced supersonic nozzles and then mixed with the fuel, which can be either H_2 or D_2. The fuel diffuses into the jet of fluorine atoms and undergoes the reaction:

$$H_2 + F \rightarrow HF^* + H$$

or

$$D_2 + F \rightarrow DF^* + D \ .$$

The excited HF* or DF* flows next through a resonator whose optical axis is transverse to the flow direction.

Instead of using an electric arc jet heater, the dissociation energy can also be provided by the combustion of $H_2 + F_2$ for a DF laser, or $D_2 + F_2$ for an HF laser. In these cases no external electric power is required; the laser operates completely on chemical energy. This type of chemical laser is capable of producing very large amounts of output power.

10.5.3 Liquid Lasers

The use of a liquid gain medium provides some advantages over both solid-state and gas lasers. For example, because fluids can be readily circulated the cooling is more effective when compared to solid-state lasers. Also, liquids are self-healing, whereas in solid-state lasers the gain medium can develop cracks, bubbles, or other defects that quickly degrade performance. Also, while gas lasers share many advantages of liquid lasers, their performance is limited because of the much lower concentration of active atoms. Finally, the cost of a solid-state laser increases rapidly with the size of the rod, which is generally limited by the method of fabrication. No such limitation exists for liquid lasers. On the negative side, liquids usually have a much larger coefficient of expansion than do solids. This property must be taken into consideration in the design of liquid lasers by not confining the liquid in a fixed glass or quartz container.

Of the several types of liquid lasers, the organic-dye laser is by far the most common. The active medium is an organic fluorescent dye dissolved in a liquid solvent such as ethyl alcohol, methyl alcohol, or water. The dye consists of a class of organic molecules that absorb strongly in the UV or visible part of the spectrum and fluoresce intensely in the visible or near IR. The laser dyes

typically belong to one of the following types: (1) polymethine dyes lase in the red or near IR (0.7 to 1.5 μm); (2) xanthene dyes lase in the visible (0.5 to 0.7 μm)—a common example is the rhodamine 6G dye; (3) coumarin dyes oscillate in the blue-green (0.4 to 0.5 μm); and (4) scintillator dyes lase in the UV (below 0.4 μm).

The uninitiated may well be overwhelmed by the many choices of dyes. Hundreds of dyes have been found to be suitable for this purpose and new dyes are being constantly developed. Most dyes can be used with several solvents and a range of concentrations. To tune across the entire visible part of the spectrum may require the use of more than one dozen different dyes. A single dye can typically be tuned over a 40-nm range.

To tune the laser within the gain bandwidth of the dye, a tuning element such as a diffraction grating, dispersing prism, tunable etalon, or birefringent filter may be used. A single tuning element results in a beam linewidth of about 1 nm. Stacking two or more such tuning elements narrows the linewidth even more. When a grazing incident diffraction grating is used, tuning is achieved by rotating the resonator mirror closest to the grating. Grazing incidence increases the resolving power of the grating, which reduces the linewidth to as little as ~0.01 nm. Another approach is to use an intraresonator beam expander, which magnifies the thin laser beam so that a larger fraction of the grating is illuminated. This also improves the resolving power of the grating and results in spectral narrowing of the laser beam.

Dye lasers operate in either a pulsed or cw mode. The two types of lasers are so different that little overlap exists in either the technology or the applications. The first dye laser operated in 1966 and was pumped by a flashlamp. Because the dye molecules have a very short lifetime, the pumping pulse must be short as well. To reach threshold for cw operation, the pump beam must be focused on a small volume of a dye flowing at high speed. The dye is in the shape of a planar, free-flowing (no dye-cell windows) jet dissolved in a viscous solvent such as ethylene glycol. The dye jet is at the Brewster angle relative to the resonator axis and is about 200 μm thick (in the direction of the resonator axis).

The large fluorescent gain bandwidth (typically 20 to 50 nm) also permits the generation of picosecond (10^{-12} s) and subpicosecond pulses. This operation mode requires that the pump laser be mode locked first, then the dye laser resonator is adjusted to the same length as the pump laser so the two operate synchronously.

Most of the commercially available dye lasers are pumped by another laser. The gain coefficient for a dye laser can take on values as high as 10^3 cm^{-1}. Typical values for a ruby laser are 0.1 cm^{-1}. Only semiconductor lasers have gain coefficients ($\approx 10^2$ cm^{-1}) approaching that of a dye laser. Some frequently used lasers for UV excitation of the dye are the nitrogen laser and excimer lasers such as KrF and XeF, while for dyes that oscillate at wavelengths longer than approximately 0.55 to 0.6 μm, the second harmonic of a Q-switched Nd:YAG laser is generally used. Other excitation sources are argon-ion and krypton-ion lasers. The conversion efficiency from pump-laser to dye-laser output can be as high as ~40%.

Because of their wavelength tunability from the UV to near IR, and the capability to generate very short pulses, organic dye lasers are frequently used

in scientific research as a tunable source of radiation for high-resolution spectroscopy, or as short-pulse lasers (down to 0.1 ps) for high-resolution time-domain spectroscopy.

10.5.4 Semiconductor Lasers

Semiconductor lasers are unique for their small size, high efficiency, and suitability for economical mass production. Output wavelengths of commercially available devices range from 0.63 to 1.58 μm with cw output powers of milliwatts to several tens of watts for arrays. The availability of these lasers has created many new applications in such areas as fiber optic communication, compact disk players, optical memory drives, laser printers, and laser-pumped solid-state lasers.

Semiconductor lasers are quite different from the types of lasers that have been described thus far. This difference arises from the band structure of energy levels. Each energy band contains a huge number of closely spaced energy states. The gaps between the bands are forbidden energy ranges. At a very low temperature the lowest energy bands are completely filled with electrons. The highest filled band is the valence band, and the lowest empty band is the conduction band. These two bands are illustrated in Fig. 10.34(a) where the range of forbidden energy states between the valence band and the conduction band is the energy gap E_g. The excitation, or pumping, process transfers electrons from the valence band to the conduction band. This could happen, for example, when photons of energy $hf \geq E_g$ are absorbed in the semiconductor. The electron vacancies that occur in the valence band are called holes. The electrons that have been pumped into the conduction band will, after about 10^{-11} to 10^{-13} s, drop to the lowest levels in that band, and any electrons near the top of the valence band will have dropped to the lowest unoccupied levels, thereby leaving the top of the valence band full of holes [see Fig. 10.34(b)]. Eventually the electrons in the conduction band recombine with the holes, emitting photons in the process. This radiation is called recombination radiation. Under the right circumstances, such as when the semiconductor is inside an optical resonator and the gain threshold condition is exceeded, the process of stimulated emission of recombination radiation can produce laser activity.

When small amounts of certain elements (donors) are added to a semiconductor, excess electrons are introduced into the semiconductor. The energy

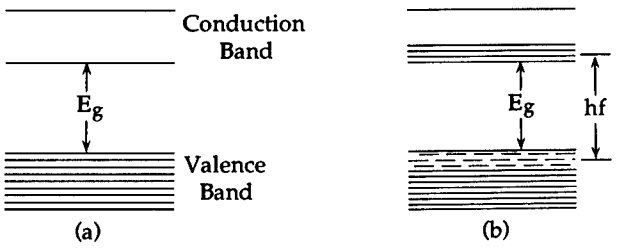


Fig. 10.34 (a) The valence band, forbidden energy gap E_g, and the conduction band of a semiconductor. (b) Optical pumping excites electrons from the filled valence band to the conduction band, leaving holes behind.

levels of these electrons are close to the bottom of the conduction band. The semiconductor is now called an n-type semiconductor. Small amounts of other elements (acceptors) can create holes in the valence band. The resulting material is known as a p-type semiconductor. When a p-type and an n-type semiconductor are joined, a p-n junction is formed. Recombination radiation and stimulated emission of recombination radiation (laser emission) can be obtained at the junction region if an electric current is sent through the semiconductor in such a way that electrons are injected into the n type and holes into the p type. This process is known as carrier injection, and the laser is an injection (or diode) laser. The other types of semiconductor lasers are optically pumped devices where an external light source produces excess carriers and electron-beam pumped lasers where high-energy electrons produce the excess carriers. These two types of semiconductor lasers are used for experimental purposes in the study of semiconductors, which are not suitable for the fabrication of injection lasers. In the remainder of this section we limit our description to the injection laser because it is the most practical of the semiconductor lasers.

The structure of a p-n junction laser is shown in Fig. 10.35. It is evident that these lasers are unique because of their small size. They are fabricated from a single crystal wafer of an n-type semiconductor such as, for example, GaAs containing 10^{17} to 10^{18} donors (Te or Se) per cubic centimeter. An acceptor element, such as zinc, is then diffused into the top layer of the wafer to a depth of 10 to 100 μm until the acceptor concentration exceeds the donor concentration by about one order of magnitude. In this way the top of the wafer is p type and the rest of the wafer n type, with a transition region (or junction) between them. The thickness of the junction from which the laser beam emerges is a few micrometers. To obtain feedback for lasing, the end faces are cut or cleaved perpendicular to the junction and parallel to each other. The index of refraction of the semiconductor is usually sufficiently high (e.g., $n = 3.6$ for GaAs) that there is no need to increase the reflectivity of the end faces with reflective coatings. However, if output is desired from one end, or if mirrors of higher reflectivity are necessary to reduce the gain threshold, the reflectivity can be increased by coating one or both facets with reflective materials. The optical output power can be increased by increasing the current flowing through

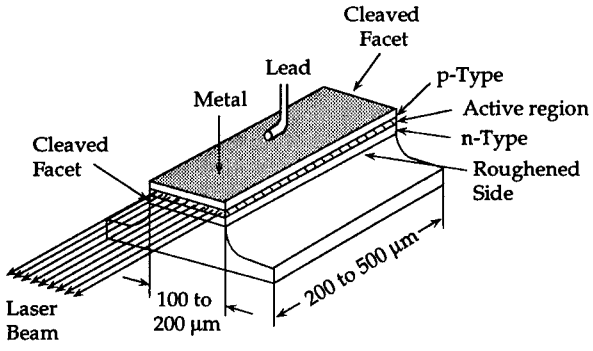


Fig. 10.35 Typical construction of a homojunction diode laser made of gallium arsenide (GaAs).

the junction. Unfortunately, higher currents also produce more heat, which can produce heat-induced damage.

Losses in the active region are due to diffraction and absorption of laser photons. Because the active region is only about a few micrometers thick, diffraction spreading causes the laser beam to extend into the p and n regions where it is strongly attenuated. To overcome these losses the room temperature current density necessary to reach threshold is approximately 10^5 A/cm^2 for GaAs. These high current densities prevent cw operation. Because the threshold current density decreases rapidly with reduced operating temperature, it is possible to achieve cw operation only at cryogenic temperatures. This is a serious disadvantage of the simple diode laser described thus far, and has limited its potential for practical applications.

The basic p-n junction of the diode laser described previously is called a homojunction because only one material is used in the junction. The junction region has a slightly higher index of refraction than the adjoining regions. This produces a light pipe effect that aids in trapping the laser beam in the junction region. In the homojunction, however, this index difference is small, and much light is still lost.

It is possible to improve the beam trapping and reduce the threshold current by using different semiconductor materials to control the index of refraction and the width of the junction. For example, the threshold current density for room temperature operation can be reduced by about two orders of magnitude by using heterojunctions. A schematic diagram of a double-heterojunction device is shown in Fig. 10.36. The two junctions are $Al_{0.3}Ga_{0.7}As(p)$-GaAs and GaAs-$Al_{0.3}G_{0.7}As(n)$. The active region is a thin layer of GaAs (0.1 to 0.3 μm).

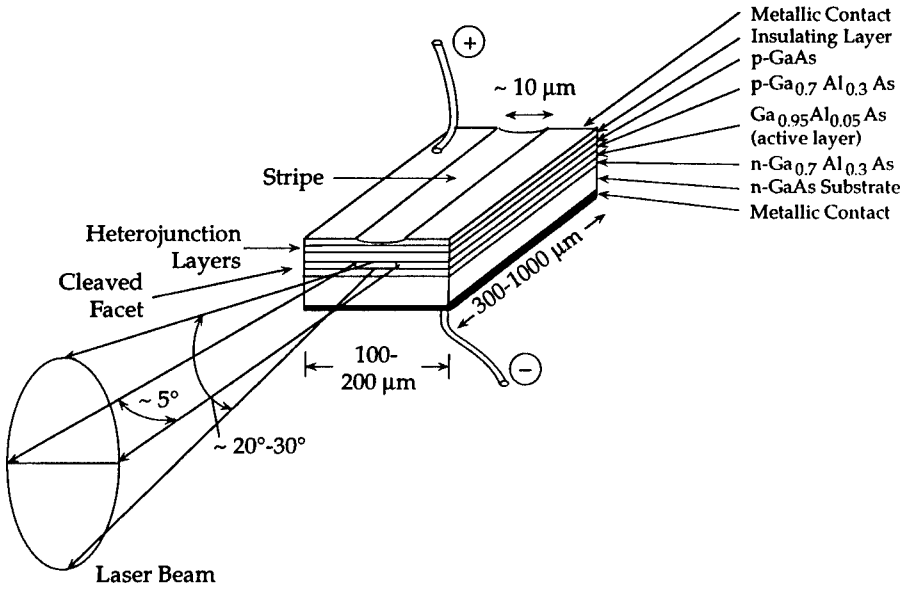


Fig. 10.36 Structural details of a stripe geometry double-heterojunction semiconductor laser.

Because of the greatly reduced threshold current density, cw operation at room temperature is possible. In the device shown in Fig. 10.36, the current from the positive electrode is confined to a narrow strip that is 5 to 10 μm wide. The benefits of this are a reduced threshold current (typically about 50 mA) and a reduced gain volume that allows laser emission in the fundamental transverse mode only. The diffraction-limited beam has an elliptical cross section with a beam divergence angle of approximately 20 to 30 deg in the plane orthogonal to the junction and ~5 deg in the plane parallel to the junction. Optical systems are available that transform this highly elliptical beam to a beam with a circular cross section.

Epitaxial techniques are used to grow the various layers of a heterojunction on a single-crystal substrate. The laser wavelength—and photon energy—is limited to a narrow range determined by the bandgap energy of the material forming the active layer. In commercial diode lasers that operate at room temperature, the substrate is either InP or GaAs and the active and confining layers are III-V alloys that are lattice matched to the substrate. The performance characteristics of the three principal types of commercially available diode lasers are listed in Table 10.4. For each type of laser, different emission wavelengths in the specified range can be obtained by changing the alloy composition of the active layer.

In summary, progress in epitaxial growth techniques and the use of quantum-well active regions, thin (~100 Å thick) layers sandwiched between other layers of materials with wider energy gaps and smaller indexes of refraction, have in recent years produced dramatic improvements in the performance of semiconductor lasers. Threshold currents as low as 50 A/cm^2 and wall-plug efficiencies exceeding 50% are now commonplace for AlGaAs devices. Because a single semiconductor laser element can produce only a limited amount of power (typically less than 1 W cw), arrays of lasers are needed for high-power applications such as laser radar. As a result, a large portion of the semiconductor research activity in the 1980s was devoted to the development of both

Table 10.4 Performance Data of the Three Types of Commercially Available Diode Lasers

Type	Wavelength Range (μm)	Operational Mode	Output Power
$(Al_xGa_{1-x})_{0.5}In_{0.5}P/(Al_yGa_{1-y})_{0.5}I_{0.5}P$ on GaAs[a]	0.63 to 0.67	cw and pulsed, with pulse duration of <1 ns and repetition rate of 2 × 10^5 Hz	0.5 to 10 mW cw
$Al_xGa_{1-x}As/Al_yGa_{1-y}As$ on GaAs	0.75 to 0.91	cw and pulsed, with pulse duration of <1 ns to 0.2 ms	1 to 10^3 mW cw
$In_{1-x}Ga_xAs_yP_{1-y}$/InP on InP	1.06 to 1.58	cw and pulsed, with pulse duration of <1 ns to 2 μs	<1 to 17 mW

[a]By convention, the active layer is listed first and the confining layers second.

Table 10.5 Typical Performance Parameters of a Semiconductor Array

Wavelength range	805 to 820 nm
Beam bandwidth	4 to 8 nm (FWHM)
Pulses per second	50 Hz
Pulse duration	200 μs
Peak power	45 to 60 W/bar
Power density	1500 W/cm^2 for 30 bar/cm array
Lifetime	$>10^9$ shots

incoherent as well as coherent diode laser arrays. Because of more rapid progress with the incoherent arrays, they are now primarily being used as pump sources for solid-state lasers such as Nd:YAG. Typical performance characteristics for an incoherent array are listed in Table 10.5. The achievement of coherence among elements of large high-power multielement arrays is much more difficult. Such coherent combining could produce near-diffraction-limited beams at high-power levels. This, together with the possibility of regulating the phase to control beam direction, will greatly expand the scope of both military and civilian applications of semiconductor lasers.

Acknowledgment

I am fortunate to have benefited from the assistance rendered by numerous people. In particular, I would like to thank Leno S. Pedrotti for his many contributions to an earlier draft of this manuscript.

References

1. V. B. Chebotaiev, "Super-high resolution spectroscopy," in *Laser Handbook,* M. Bass and M. L. Stitch, Eds., Vol. 5, pp. 289–404, North-Holland, Amsterdam (1985).
2. A. E. Siegman, *An Introduction to Lasers and Masers,* p. 122, McGraw-Hill, New York (1971).
3. A. Szoke and A. Javan, *Physical Review Letters* **10,** 521 (1963).
4. L. F. Johnson, *Lasers,* A. K. Levine, Ed., p. 174, Marcel Dekker, New York (1966).
5. W. G. Wagner and B. A. Lengyel, *Journal of Applied Physics* **34,** 2042 (1963).
6. J. S. Kruger, *Electro-Optical System Design,* p. 12 (Sep. 1972).
7. A. G. Fox and T. Li, *Bell System Technical Journal* **40,** 453 (March 1961).
8. R. J. Freiberg and A. S. Halsted, "Transverse modes in gas lasers," *Laser Focus,* p. 21 (1968).
9. A. E. Siegman, *Lasers,* University Science Books, Mill Valley, CA (1986).
10. W. H. Steier, "Unstable resonators," in *Laser Handbook,* M. L. Stitch, Ed., Vol. 3, pp. 3–39, North-Holland, Amsterdam (1979).
11. H. Weichel, *Selected Papers on Laser Design,* Vol. MS-29, SPIE Optical Engineering Press, Bellingham, WA (1991).

Bibliography

Anderson, J. D., *Gasdynamic Lasers: An Introduction,* Academic Press, New York (1976).

Basov, N. G., et al., *Chemical Lasers,* Springer-Verlag, New York (1990).

Bekefi, G., Ed., *Principles of Laser Plasmas,* Wiley-Interscience, New York (1976).

Brau, C. A., *Free-Electron Lasers,* Academic Press, San Diego, CA (1990).

Bridges, W. G., "Atomic and ionic gas lasers," in *Methods of Experimental Physics,* C. L. Tang, Ed., Vol. **15,** pp. 33–151, Academic Press, New York (1979).

Budgor, A. B., et al., Eds., *Tunable Solid-State Lasers II,* Vol. 52 in Springer Series in Optical Sciences, Springer-Verlag, Berlin (1986).

Danielmeyer, J. G., "Progress in Nd:YAG lasers," in *Lasers,* A. K. Levine and A. J. DeMaria, Eds., Vol. **4,** Chap. 1, Marcel Dekker, New York (1976).

Das, P., *Lasers and Optical Engineering,* Springer-Verlag, New York (1991).

Duarte, F. J., and L. W. Hillmann, Eds., *Dye Laser Principles with Applications,* Academic Press, New York (1990).

Dunn, D. H. and J. N. Ross, "The argon ion laser," in *Progress in Quantum Electronics,* J. H. Sanders and S. Stenholm, Eds., Vol. **4,** pp. 233–270, Pergamon Press, London (1977).

Evthuhov, V., and J. K. Neeland, "Pulsed ruby lasers," in *Lasers,* A. K. Levine, Ed., Vol. 1, Chap. 1, Marcel Dekker, New York (1966).

Ewing, J. J., "Excimer lasers," in *Laser Handbook,* M. L. Stitch, Ed., Vol. **3,** pp. 135–197, North-Holland, Amsterdam (1979).

Findlay, D., and D. W. Goodwin, "The neodymium in YAG laser," in *Advances in Quantum Electronics,* D. W. Goodwin, Ed., Vol. **1,** pp. 77–128, Academic Press, New York (1970).

Francon, M., *Laser Speckle and Applications in Optics,* Academic Press, New York (1979).

Hecht, J., *The Laser Guidebook,* McGraw-Hill, New York (1986).

Koechner, W., *Solid State Laser Engineering,* Springer, New York (1976).

Kogelnik, H., "Propagation of laser beams," in *Applied Optics and Optical Engineering,* R. Shannon and J. C. Wynant, Eds., Vol. II, pp. 156–190, Academic Press, New York (1979).

Losev, S. A., *Gasdynamic Laser,* Springer-Verlag, Berlin (1981).

Luchini, P., and H. Motz, *Undulators and Free-Electron Lasers, International Series of Monographs on Physics* **79,** Oxford University Press, New York (1990).

Mallow, A., and L. Chabot, *Laser Safety Handbook,* Van Nostrand, New York (1978).

Marshal, T. G., *Free-Electron Lasers,* MacMillan, New York (1985).

Measures, R. M., *Laser Remote Sensing,* Wiley-Interscience, New York (1984).

Mollenauer, L. F., "Color center lasers," in *Laser Handbook,* M. L. Stitch and M. Bass, Eds., Vol. **4,** pp. 143–228, North-Holland, Amsterdam (1985).

O'Shea, D. C., W. R. Callen, and W. T. Rhodes, *Introduction to Lasers and Their Applications,* Addison-Wesley, Reading, MA (1978).

Pedrotti, F. L., and L. S. Pedrotti, *Introduction to Optics,* Chap. 7, Prentice-Hall, Englewood Cliffs, NJ (1987).

Pike, E. R., Ed., *High-Power Gas Lasers,* The Institute of Physics, Bristol and London (1975).

Rhodes, C. K., et al., Eds., *Excimer Lasers—1983,* American Institute of Physics, Woodbury, NY (1983).

Ross, D., *Lasers, Light Amplifiers, and Oscillators,* Academic Press, New York (1969).

Sargent, M., M. O. Scully, and W. E. Lamb, *Laser Physics,* Addison-Wesley, London (1974).

Schäfer, F. P., Ed., *Dye Lasers,* 2nd ed., Springer-Verlag, Berlin (1977).

Siegman, A. E., *An Introduction to Lasers and Masers,* McGraw-Hill, New York (1971).

Siegman, A. E., *Lasers,* University Science Books, Mill Valley, CA (1986).

Sliney, D., and M. Wolbarcht, *Safety with Lasers and Other Optical Sources,* Plenum Press, New York (1980).

Steier, W. H., "Unstable resonators," in *Laser Handbook,* M. L. Stitch, Ed., Vol. **3,** pp. 3–39, North-Holland, Amsterdam (1979).

Stitch, M. L., Ed., *Laser Handbook,* North-Holland, Amsterdam (1979).

Svelto, O., *Principles of Lasers,* 3rd ed., Plenum Press, New York and London (1989).

Tarasov, L. V., *Laser Physics,* Mir Publications, Moscow, in English (1983).

Thompson, G. H. B., *Physics of Semiconductor Laser Devices,* John Wiley & Sons, New York (1980).

Ultee, C. J., "Chemical and gas-dynamic lasers," in *Laser Handbook,* M. L. Stitch, Ed., Vol. **3,** pp. 199–287, North-Holland, Amsterdam (1979).

Verdeyen, J. T., *Laser Electronics,* Prentice-Hall, Englewood Cliffs, NJ (1981).

Walling, J. C., et al., "Tunable alexandrite lasers," *IEEE J. Quantum Electron.* **QE-16,** 1302–1315 (1980).

Webb, C. E., "Metal vapour lasers: Recent advances and applications," in *Gas Flow and Chemical Lasers, No. 15, Springer Proceedings in Physics,* S. Rosenwark, Ed., pp. 481–494, Springer-Verlag, Berlin (1987).

Weber, M. J., Ed., *CRC Handbook of Laser Science and Technology,* Vol. 1 (1982) to Vol. 5 (1987), CRC Press, Boca Raton, FL (1982).

Weichel, H., *Laser Beam Propagation in the Atmosphere,* Vol. TT3, SPIE Optical Engineering Press, Bellingham, WA (1990).

Weichel, H., Ed. *Selected Papers on Laser Design,* Vol. MS-29, SPIE Optical Engineering Press, Bellingham, WA (1991).

Willet, C. S., *An Introduction to Gas Lasers: Population Inversion Mechanisms,* Pergamon Press, Oxford (1974).

Wood, II, O. R., "High pressure pulsed molecular lasers," *Proc. IEEE* **62,** 355–397 (1974).

Yariv, A., *Introduction to Optical Electronics,* Holt, Rinehart and Winston, New York (1971).

Yariv, A., *Optical Electronics,* 3rd ed., Holt, Rinehart and Winston, New York (1985).

Young, M., *Optics and Lasers,* 2nd ed., Springer-Verlag, New York (1984).

Zuev, V. E., *Laser Beams in the Atmosphere,* Consultants Bureau, New York (1982).

Index