ENCYCLOPEDIA OF
MODERN OPTICS
SECOND EDITION

EDITORS IN CHIEF

Bob D. Guenther

Duke University, Durham, NC, United States

Duncan G. Steel

University of Michigan, Ann Arbor, MI, United States

VOLUME 1

Fiber Optics ■ Coherence ■ Diffraction ■ History of Lasers ■ Geometrical Optics ■ Optical Modulation ■ Polarization ■ Photonic Crystals ■ Communications ■ Lasers and Amplifiers ■ TeraHertz ■ Quantum Cavity Physics ■ Coherent Control

ELSEVIER

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

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN 978-0-12-809283-5

Publisher: Oliver Walter
Acquisition Editor: Ruth Ireland
Content Project Manager: Sean Simms
Associate Content Project Manager: Marise Willis
Designer: Greg Harris

Printed and bound in the United Kingdom

EDITORIAL BOARD

CONTENTS OF VOLUME 1

VOLUME 1

LIST OF CONTRIBUTORS FOR VOLUME 1

JM Bennett
Michelson Laboratory, China Lake, CA, USA

G Benoit
Massachusetts Institute of Technology, Cambridge, MA, USA

Marie-Anne Bizouard
Paris-Saclay University, Orsay, France

P Bollond
JDS Uniphase Corporation, Ewing, NJ, USA

Malcolm S Carroll
Sandia National Laboratory, Albuquerque, NM, United States

Nelson Christensen
Carleton College, Northfield, MN, United States and University of Côte d'Azur, Nice, France

Michael J Connelly
University of Limerick, Limerick, Ireland

LR Dalton
University of Washington, Seattle, WA, USA

MFS Ferreira
University of Aveiro, Aveiro, Portugal

Y Fink
Massachusetts Institute of Technology, Cambridge, MA, USA

G Galliano
Telecom Italia Lab, Torino, Italy

AK Ghatak
Indian Institute of Technology, New Delhi, India

HM Gibbs
University of Arizona, Tucson, AZ, USA

A Girard
EXFO, Quebec, Canada

M Gottlieb
Carnegie Mellon University, Pittsburgh, PA, USA

Klaus Grobe
ADVA Optical Networking SE, Martinsried, Germany

BD Guenther
Duke University, Durham, NC, USA

S Hart
Massachusetts Institute of Technology, Cambridge, MA, USA

D Hewak
University of Southampton, Southampton, UK

W Hoyer
Philipps-University, Marburg, Germany

Qing Hu
MIT, Cambridge, MA, United States

DFV James
Los Alamos National Laboratory, Los Alamos, NM, USA

JD Joannopoulos
Massachusetts Institute of Technology, Cambridge, MA, USA

SG Johnson
Massachusetts Institute of Technology, Cambridge, MA, USA

Stéphane Kéna-Cohen
Polytechnique Montréal, Department of Engineering Physics, Montreal, QC, Canada

HC Kandpal
National Physical Laboratory, New Delhi, India

G Khitrova
University of Arizona, Tucson, AZ, USA

M Kira
Philipps-University, Marburg, Germany

Zbigniew Kisiel
Institute of Physics of the Polish Academy of Sciences, Warsaw, Poland

SW Koch
Philipps-University, Marburg, Germany

Tristan Kremp
OFS Fitel, LLC, Somerset, NJ, United States

W Kuang
University of Southern California, Los Angeles, CA, USA

P Kumar
Northwestern University, Evanston, IL, USA

PG Kwiat
University of Illinois at Urbana-Champaign, Urbana, IL, USA

Thaddeus D Ladd
HRL Laboratories, LLC, Malibu, CA, United States

O Leclerc
Alcatel Research & Innovation, Marcoussis, France

Kang Liu
University of Rochester, Rochester, NY, United States

Alan Mickelson
University of Colorado at Boulder, Boulder, CO, United States

G Millot
Université de Bourgogne, Dijon, France

A Nussbaum
University of Minnesota, Minneapolis, MN, USA

J O'Brien
University of Southern California, Los Angeles, CA, USA

T-C Poon
Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Antoni Rogalski
Military University of Technology, Warsaw, Poland

A Sharma
Indian Institute of Technology, New Delhi, India

JE Sharping
Cornell University, Ithaca, NY, USA

JE Sipe
University of Toronto, Toronto, ON, Canada

D Suhre
Carnegie Mellon University, Pittsburgh, PA, USA

D Suino
Telecom Italia Lab, Torino, Italy

P Tchofo-Dinda
Université de Bourgogne, Dijon, France

L Thévenaz
École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

K Thyagarajan
Indian Institute of Technology, New Delhi, India

CH Townes
University of California, Berkeley, CA, USA

J Turunen
University of Joensuu, Joensuu, Finland

T Vallius
University of Joensuu, Joensuu, Finland

HM van Driel
University of Toronto, Toronto, ON, Canada

H Walther
University of Munich and Max-Planck-Institute for Quantum Optics, Garching, Germany

Paul S Westbrook
OFS Fitel, LLC, Somerset, NJ, United States

Benjamin S Williams
University of California, Los Angeles, CA, United States

RS Windeler
OFS Laboratories, Murray Hill, NJ, USA

Xi-Cheng Zhang
University of Rochester, Rochester, NY, United States; ITMO University, Saint-Petersburg, Russia; and Capital Normal University, Beijing, China

CONTENTS OF ALL VOLUMES

VOLUME 1

VOLUME 2

VOLUME 3

VOLUME 4

VOLUME 5

EDITORS IN CHIEF

Bob D. Guenther received his undergraduate degree from Baylor University and his graduate degrees in Physics from University of Missouri. He has had research experience in condensed matter and optical physics. For 9 years he was active in research management as a Senior Executive in the Army, responsible for the physics research sponsored by the Army. After retiring from the government he held the position of Interim Director of the Free Electron Laser Laboratory and helped establish Duke's Fitzpatrick Center for Photonics and Communication Systems and served as Executive Director of the Center until his retirement. In a continuation of the retirement process, he moved to Applied Quantum Technology and has just retired from that company. He is author of the textbook, *Modern Optics*, second edition. He is now composing an elementary book in optics.

Duncan G. Steel is the Robert J. Hiller Professor of Electrical and Computer Engineering, as well as Professor of Physics and Biophysics. Prior to joining the faculty of the University of Michigan in 1985, he was a senior research scientist at Hughes Aircraft Company at the Hughes Research Laboratories. At the University of Michigan, he was Area Chair and Director of the Optical Sciences Laboratory for 20 years until 2007 when he took over the position of Chair of the Biophysics Research Division during its transition to an academic unit. As an educator and teacher, he has chaired or cochaired over 60 doctoral committees. His research includes coherent optical studies of semiconductors and their application to quantum information. His work also included 30 years of studies on age-related modifications in proteins where he exploited numerous optical techniques including single molecule studies in neurons in his studies on Alzheimer's Disease. He was a Guggenheim Scholar and received the 2010 Isakson Prize from the American Physical Society.

INTRODUCTION

This is the second, updated edition of the *Encyclopedia of Modern Optics*. There are 197 entries, many of them new or updated, reflecting the enormous progress in the optical sciences and technology and the ever-expanding impact since the publication of the first edition. Some of the new topics are:

- Nano-photonics and Plasmonics
- Quantum Optics
- Quantum Information
- Optical Interconnects
- Photonic Crystals and Their Applications
- High Efficiency LED's
- Displays
- Transformation Optics
- Fiber Lasers
- Terahertz
- Multidimensional Spectroscopy
- Organic Optoelectronics
- Gravitational Wave Detectors
- Meta Materials and Plasmonics

Selection of article topics and recruiting authors for those topics in this edition has been the work of the topical editors listed in the prologue. They were able to solicit contributions from internationally recognized leaders in their field.

The entries of the encyclopedia are arranged by subject as best as possible, a task made difficult because the field is now highly interdisciplinary and there are many subjects that have an impact in many different areas. We have added references to other articles so that readers can obtain a deeper understanding of the material or understand how a specific discussion on basic science may impact an application or technology.

Dispersion

L Thévenaz, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Nomenclature

D_λ Chromatic dispersion [ps nm^{-1} km^{-1}]
E Electric field [V m^{-1}]
N Group index
V_g Group velocity [m s^{-1}]
D_ν Group Velocity Dispersion GVD [s^2 m^{-1}]
α Linear attenuation coefficient [m^{-1}]
χ Medium susceptibility
V Normalized frequency
ν Optical frequency [s^{-1}]
β Propagation constant [m^{-1}]
τ_D Propagation delay [s]
P Polarization density field [A s m^{-2}]
n Refractive index
c_0 Vacuum light velocity [m s^{-1}]
λ Wavelength [m]

Introduction

Dispersion designates the property of a medium to propagate the different spectral components of a wave with a different velocity. It may originate from the natural dependence of the refractive index of a dense material to the wavelength of light, and one of the most evident and magnificent manifestations of dispersion is the appearance of a rainbow in the sky. In this case dispersion gives rise to a different refraction angle for the different spectral components of the white light, resulting in an angular dispersion of the sunlight spectrum and explicating the etymology of the term dispersion. Despite this spectacular effect, dispersion is mostly seen as an impairment in many applications, in particular for optical signal transmission. In this case, dispersion causes a rearrangement of the signal spectral components in the time domain, resulting in temporal spreading of the information and eventually in severe distortion.

There is a trend to designate all phenomena resulting in a temporal spreading of information as a type of dispersion, namely the polarization dispersion in optical fibers that should be properly named as polarization mode delay (PMD). In this article chromatic dispersion will be solely addressed, that designates the dependence of the propagation velocity on wavelength and that corresponds to the strict etymology of the term.

Effect of Dispersion on an Optical Signal

An optical signal may always be considered as a sum of monochromatic waves through a normal Fourier expansion. Each of these Fourier components propagates with a different phase velocity if the optical medium is dispersive, since the refractive index n depends on the optical frequency ν. This phenomenon is linear and causal and gives rise to a signal distortion that may be properly described by a transfer function in the frequency domain.

Let $n(\nu)$ be the frequency-dependent refractive index of the propagation medium. When the optical wave propagates in a waveguiding structure the effective wavenumber is properly described by a propagation constant β, that reads:

$$\beta(\nu) = \frac{2\pi\nu}{c_0} n(\nu) \tag{1}$$

where c_0 is the vacuum light velocity. The propagation constant corresponds to an eigenvalue of the wave equation in the guiding structure and normally takes a different value for each solution or propagation mode. For free-space propagation this constant is simply equal to the wavenumber of the corresponding plane wave.

The complex electrical field $E(z,t)$ of a signal propagating in the z direction may be properly described by the following expression:

$$\begin{aligned} E(z,t) &= A(z,t)\, e^{i(2\pi\nu_0 t - \beta_0 z)} \\ &\text{with} \quad \nu_0 : \text{the central optical frequency} \\ &\text{and} \quad \beta_0 = \beta(\nu_0) \end{aligned} \tag{2}$$

where $A(z,t)$ represents the complex envelope of the signal, supposed to be slowly varying compared to the carrier term oscillating with the frequency ν_0. Consequently, the signal will spread over a narrow band around the central frequency ν_0 and the propagation constant β can be conveniently approximated by a limited expansion to the second order:

$$\beta(\nu) = \beta_0 + \frac{d\beta}{d\nu}\Big|_{\nu=\nu_0}(\nu-\nu_0) + \frac{d^2\beta}{d\nu^2}\Big|_{\nu=\nu_0}(\nu-\nu_0)^2 \tag{3}$$

For a known signal $A(0, t)$ at the input of the propagation medium, the problem consists in determining the signal envelope $A(z, t)$ after propagation over a distance z. The linearity and the causality of the system make possible a description using a transfer function $H_z(v)$ such as:

$$\tilde{A}(z, v) = H_z(v)\tilde{A}(0, v) \tag{4}$$

where $\tilde{A}(z, v)$ is the Fourier transform of $A(z, t)$.

To make the transfer function $H_z(v)$ explicit, let us assume that the signal corresponds to an arbitrary harmonic function:

$$A(0, t) = A_0\, \mathrm{e}^{\mathrm{i}2\pi f t} \tag{5}$$

Since this function is arbitrary and the signal may always be expanded as a sum of harmonic functions through a Fourier expansion, there is no loss of generality. The envelope identified as a harmonic function actually corresponds to a monochromatic wave of optical frequency $v = v_0 + f$ as defined by **Eq. (2)**. Such a monochromatic wave will experience the following phase shift through propagation:

$$\begin{aligned} E(z, t) &= A_0 \mathrm{e}^{\mathrm{i}[2\pi(v_0+f)t-\beta(v_0+f)z]} \\ &= A_0\, \mathrm{e}^{\mathrm{i}2\pi f t} \mathrm{e}^{\mathrm{i}[2\pi v_0 t-\beta(v_0+f)z]} \end{aligned} \tag{6}$$

On the other hand, an equivalent expression may be found using the linear system described by **Eqs. (2)** and **(4)**:

$$\begin{aligned} E(z, t) &= A(z, t)\, \mathrm{e}^{\mathrm{i}(2\pi v_0 t-\beta_0 z)} \\ &= A_z\, \mathrm{e}^{\mathrm{i}2\pi f t} \mathrm{e}^{\mathrm{i}(2\pi v_0 t-\beta_0 z)} \end{aligned} \tag{7}$$

Since **Eqs. (6)** and **(7)** must represent the same quantities and using the definition in **Eq. (4)**, a simple comparison shows that the transfer function must take the following form:

$$H_z(v) = \mathrm{e}^{-\mathrm{i}[\beta(v)z-\beta_0 z]} \tag{8}$$

The transfer function takes a more analytical form using the approximation in **Eq. (3)**:

$$H_z(v) = \mathrm{e}^{-\mathrm{i}2\pi(v-v_0)\tau_\mathrm{D}} \mathrm{e}^{-\mathrm{i}\pi D_v (v-v_0)^2 z} \tag{9}$$

In the transfer function interpretation the first term represents a delay term. It means that the signal is delayed after propagation by the quantity:

$$\tau_\mathrm{D} = \frac{1}{2\pi}\frac{\mathrm{d}\beta}{\mathrm{d}v_0} z = \frac{z}{V_\mathrm{g}} \tag{10}$$

where V_g represents the signal group velocity. This term therefore brings no distortion for the signal and thus states that the signal is replicated at the distance z with a delay τ_D.

The second term is the distortion term which is similar in form to a diffusion process and normally results in time spreading of the signal. In the case of a light pulse it will gradually broaden while propagating along the fiber, like a hot spot on a plate gradually spreading as a result of heat diffusion. The effect of this distortion is proportional to the distance z and to the coefficient D_v, named group velocity dispersion (GVD):

$$D_v = \frac{1}{2\pi}\frac{\mathrm{d}^2\beta}{\mathrm{d}v^2} = \frac{\mathrm{d}}{\mathrm{d}v}\left(\frac{1}{V_\mathrm{g}}\right) \tag{11}$$

It is important to point out that the GVD may be either positive (normal) or negative (anomalous) and the distortion term in the transfer function in **Eq. (9)** may be exactly cancelled by propagating in a medium with D_v of opposite sign. It means that the distortion resulting from chromatic dispersion is reversible and this is widely used in optical links through the insertion of dispersion compensators. These are elements made of specially designed fibers or fiber Bragg gratings showing an enhanced GVD coefficient, with a sign opposite to the GVD in the fiber.

From the transfer function in **Eq. (9)** it is possible to calculate the impulse response of the dispersive medium:

$$h_z(t) = \frac{1}{\sqrt{i|D_v|z}} \mathrm{e}^{\mathrm{i}\pi\frac{(t-\tau_D)^2}{D_v z}} \tag{12}$$

so that the distortion of the signal may be calculated by a simple convolution of the impulse response with the signal envelope in the time domain.

The effect of dispersion on the signal can be more easily interpreted by evaluating the dispersive propagation of a Gaussian pulse. In this particular case the calculation of the resulting envelope can be carried out analytically. If the signal envelope takes the following Gaussian distribution at the origin:

$$A(0, t) = A_0 \mathrm{e}^{-\frac{t^2}{\tau_0^2}} \tag{13}$$

with τ_0 the $1/e$ half-width of the pulse, the envelope at distance z is obtained by convoluting the initial envelope with the impulse response $h_z(t)$:

$$\begin{aligned} A(z,t) &= h_z(t) \otimes A(0,t) \\ &= A_0 \sqrt{\frac{iz_0}{z+iz_0}} e^{i\pi \frac{(t-\tau_D)^2}{D_v(z+iz_0)}} \end{aligned} \tag{14}$$

where

$$z_0 = -\frac{\pi \tau_0^2}{D_v} \tag{15}$$

represents the typical dispersion length, that is the distance necessary to make the dispersion effect noticeable.

The actual pulse spreading resulting from dispersion can be evaluated by calculating the intensity of the envelope at distance z:

$$|A(z,t)|^2 = A_0 \frac{\tau_0}{\tau(z)} e^{-\frac{2(t-\tau_D)^2}{\tau^2(z)}} \tag{16}$$

that is still a Gaussian distribution centered about the propagation delay time τ_D, with $1/e^2$ half-width:

$$\tau(z) = \tau_0 \sqrt{1 + (z/z_0)^2} \tag{17}$$

The variation of the pulse width $\tau(z)$ is presented in **Fig. 1** and clearly shows that the pulse spreading starts to be nonnegligible from the distance $z = z_0$. This gives a physical interpretation for the dispersion length z_0. It must be pointed out that there is a direct formal similarity between the pulse broadening of a Gaussian pulse in a dispersive medium and the spreading of a free-space Gaussian beam as a result of diffraction. Asymptotically for distances $z \gg z_0$, the pulsewidth increases linearly:

$$\tau(z) \simeq |D_v| \frac{z}{\pi \tau_0} \tag{18}$$

It must be pointed out that the width increases proportionally to the dispersion D_v, but also inversely proportionally to the initial width τ_0. This results from the larger spectral width corresponding to a narrower pulsewidth, giving rise to a stronger dispersive effect.

The chromatic dispersion does not modify the spectrum of the transmitted light, as any linear effect. This can be straightforwardly demonstrated by evaluating the intensity spectrum of the signal envelope at any distance z, using the **Eqs. (4)** and **(8)**:

$$\begin{aligned} |\tilde{A}(z,\nu)|^2 &= |H_z(\nu)\tilde{A}(0,\nu)|^2 = |e^{-i[\beta(\nu)z - \beta_0 z]}|^2 |\tilde{A}(0,\nu)|^2 \\ &= |\tilde{A}(0,\nu)|^2 \end{aligned} \tag{19}$$

It means that the pulse characteristics in the time and frequency domains are no longer Fourier-transform limited, since after the broadening due to dispersion, the spectrum should normally spread over a narrower spectral width. This feature results from a rearrangement of the spectral components within the pulse. This can be highlighted by evaluating the distribution of instantaneous frequency through the pulse. The instantaneous frequency ω_i is defined as the time-derivative of the wave phase factor $\phi(t)$ and is uniformly equal to the optical carrier pulsation $\omega_i = 2\pi\nu_0$ for the initial pulse, as can be deduced from the phase factor at $z = 0$ by combining **Eqs. (2)** and **(13)**. This constant instantaneous frequency means that all spectral components are uniformly present within the pulse at the origin.

After propagation through the dispersive medium the phase factor $\phi(t)$ can be evaluated by combining **Eqs. (2)** and **(14)** and evaluating the argument of the resulting expression. The instantaneous frequency ω_i is obtained after a simple time derivative

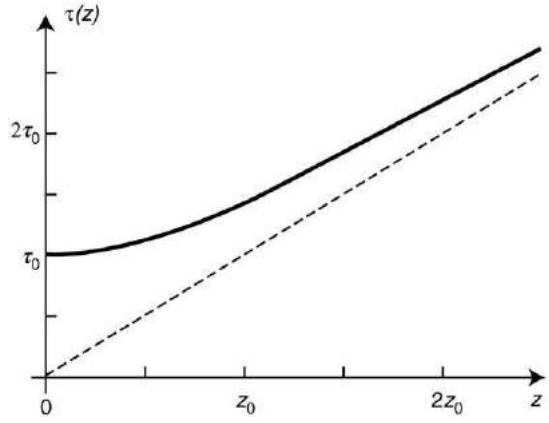


Fig. 1 Variation of the $1/e^2$ half-width of a Gaussian pulse, showing the pulse spreading effect of dispersion. The dashed line shows the asymptotic linear spreading for large propagation distance.

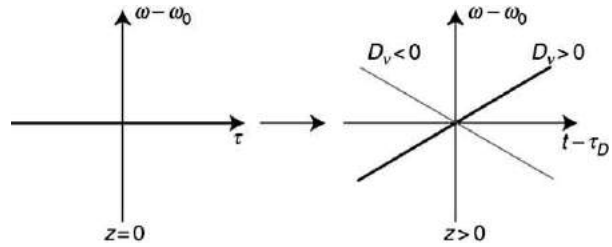


Fig. 2 Distribution of the instantaneous frequency through a Gaussian pulse. At the origin the distribution is uniform (left) and the dispersion induces a frequency chirp that depends on the sign of the GVD coefficient D_v.

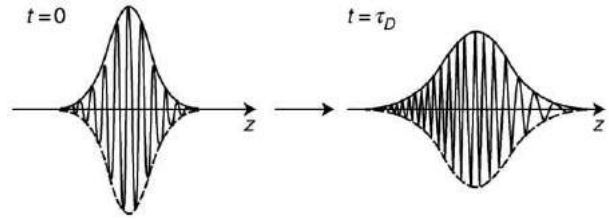


Fig. 3 The dispersion results in a pulse broadening together with a frequency chirp, here for a normal GVD, that can be seen like a frequency modulation.

of $\phi(t)$ and reads:

$$\omega_{\mathrm{i}}(t) = \omega_0 + \frac{2\pi z}{D_v(z^2 + z_0^2)}(t - \tau_{\mathrm{D}}) \tag{20}$$

For $z>0$ the instantaneous frequency varies linearly over the pulse, giving rise to a frequency chirp. The frequency components are re-arranged in the pulse, so that the lower frequency components are in the leading edge of the pulse for a normal dispersion $D_v>0$ and the higher frequencies in the trailing edge. For an anomalous dispersion $D_v<0$, the arrangement is opposite, as can be seen in **Fig. 2**. The effect of this frequency chirp can be visualized in **Fig. 3**, showing that the effect of dispersion is equivalent to a frequency modulation over the pulse. The chirp is maximal at position z_0 and the pulsewidth takes its minimal value τ_0 when the chirp is zero. It is evident with this description that the pulse broadening may be entirely compensated through propagation in a medium of opposite group velocity dispersion; this is equivalent to reversing the time direction in **Fig. 3**. Moreover a pre-chirped pulse can be compressed to its Fourier-transform limited value after propagation in a medium with the proper dispersion sign. This feature is widely used for pulse compression after pre-chirping through propagation in a medium subject to optical Kerr effect.

The above description of the propagation of a Gaussian pulse implicitly states that the light source is perfectly coherent. In the early stages of optical communications it was not at all the case, since most sources were either light emitting diodes or multimode lasers. In this case the spectral extent of the signal was much larger than actually required for the strict need of modulation. Each spectral component may thus be considered as independently propagating the signal and the total optical wave can be identified to light merged from discrete sources emitting simultaneously the same signal at a different optical frequency. The group velocity dispersion will cause a delay $\delta\tau$ between different spectral components separated by a frequency interval $\delta\nu$ that can be simply evaluated by a first-order approximation:

$$\delta\tau = \frac{\mathrm{d}\tau_{\mathrm{D}}}{\mathrm{d}\nu}\delta\nu = \frac{\mathrm{d}}{\mathrm{d}\nu}\left(\frac{z}{V_{\mathrm{g}}}\right)\delta\nu = D_v z\delta\nu \tag{21}$$

where **Eqs. (10)** and **(11)** have been used. The delay $\delta\tau$ is thus proportional to the GVD coefficient D_v, to the propagation distance z and the frequency separation $\delta\nu$. This description can be extended to a continuous frequency distribution with a spectral width σ_ν, resulting to the following temporal broadening σ_τ:

$$\sigma_\tau = |D_v|\sigma_\nu z \tag{22}$$

Traditionally the spectral characteristics of a source are given in units of wavelength and the GVD coefficient is expressed in optical fibers accordingly. Following the same description as above, the temporal broadening σ_τ, for a spectral width σ_λ in wavelength units, reads:

$$\sigma_\tau = |D_\lambda|\sigma_\lambda z \tag{23}$$

Since equal spectral widths must give equal broadening the value of the GVD in units of wavelength can be deduced from the natural definition in frequency units:

$$D_\lambda = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{1}{V_{\mathrm{g}}}\right) = \frac{\mathrm{d}}{\mathrm{d}\nu}\left(\frac{1}{V_{\mathrm{g}}}\right)\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} = -\frac{c_0}{\lambda^2}D_v \tag{24}$$

It must be pointed out that the coefficient D_λ takes a sign opposite to D_ν; in other words, a normal dispersion corresponds to a negative GVD coefficient D_λ. It is usually expressed in units of picoseconds of temporal broadening, per nanometer of spectral width and per kilometer of propagating distance, or ps/nm km. For example, a pulse showing a spectral width of 1 nm propagating through a 100 km fiber having a dispersion D_λ of $+10$ ps/nm km, will experience, according to **Eq. (23)**, a pulse broadening σ_τ of $10 \times 1 \times 100 = 1000$ ps or 1 ns.

Material Group Velocity Dispersion

Any dense material shows a variation of its index of refraction n as a function of the optical frequency ν. This natural property is called material dispersion and is the dominant contribution in weakly guiding structures such as standard optical fibers. This natural dependence results from the noninstantaneous response of the medium to the presence of the electric field of the optical wave. In other words, the polarization field $P(t)$ corresponding to the material response will vary with some delay or inertia to the change of the incident electric field $E(t)$. This delay between cause and effect generates a memory-type response of the medium that may be described using a time-dependent medium susceptibility $\chi(t)$. The relation between medium polarization at time t and incident field results from the weighted superposition of the effects of $E(t')$ at all previous times $t' < t$. This takes the form of the following convolution:

$$P(t) = \varepsilon_0 \int_{-\infty}^{+\infty} \chi(t - t')E(t')\mathrm{d}t' \tag{25}$$

Through application of a simple Fourier transform this relation reads in the frequency domain as:

$$P(\nu) = \varepsilon_0 \chi(\nu)E(\nu) \tag{26}$$

showing clearly that the noninstantaneous response of the medium results in a frequency-dependent refractive index using the standard relationship with the susceptibility χ:

$$n(\nu) = \sqrt{1 + \chi(\nu)} \quad \text{where} \quad \chi(\nu) = \mathrm{FT}\{\chi(t)\} \tag{27}$$

This means that a beautiful natural phenomenon such as a rainbow, originates on a microscopic scale from the sluggishness of the medium molecules to react to the presence of light. For signal propagation, it results in a distortion of the signal and in most cases in a pulse spreading, but the microscopic causes are in essence identical. This noninstantaneous response is tightly related to the molecules' vibrations that also give rise to light absorption. For this reason it is convenient to express the propagation in an absorptive medium by adding an imaginary part to the susceptibility $\chi(\nu)$:

$$\chi(\nu) = \chi'(\nu) + \mathrm{i}\chi''(\nu) \tag{28}$$

so that the refractive index $n(\nu)$ and the absorption coefficient $\alpha(\nu)$ reads in a weakly absorbing medium:

$$\begin{aligned} n(\nu) &= \sqrt{1 + \chi'(\nu)} \\ \alpha(\nu) &= -\frac{2\pi\chi''(\nu)}{\lambda n(\nu)} \end{aligned} \tag{29}$$

Since the response of the medium, given by the time-dependent susceptibility $\chi(t)$ in **Eq. (25)**, is real and causal, the real and imaginary part of the susceptibility in **Eq. (28)** are not entirely independent and are related by the famous Kramers–Kronig relations:

$$\begin{aligned} \chi'(\nu) &= \frac{2}{\pi}\int_0^\infty \frac{s\chi''(\nu)}{s^2 - \nu^2}\mathrm{d}s \\ \chi''(\nu) &= \frac{2}{\pi}\int_0^\infty \frac{\nu\chi'(\nu)}{\nu^2 - s^2}\mathrm{d}s \end{aligned} \tag{30}$$

Absorption and dispersion act in an interdependent way on the propagating optical wave and knowing either the absorption or the dispersion spectrum is theoretically sufficient to determine the entire optical response of the medium. The interdependence between absorption and dispersion is typically illustrated in **Fig. 4**. The natural tendency is to observe a growing refractive index for increasing frequencies in low absorption regions. In this case, the dispersion is called normal and this is the most observed situation in transparency regions of a material, which obviously offer the largest interest for optical propagation. In an absorption line the tendency is opposite, a diminishing index for increasing wavelength, and such a response is called anomalous dispersion. The dispersion considered here is the phase velocity dispersion, represented by the slope of $n(\nu)$, that must not be mistaken with the group velocity dispersion (GVD) that only matters for signal distortion. The difference between these two quantities is clarified below.

To demonstrate that a frequency-dependent refractive index $n(\nu)$ gives rise to a group velocity dispersion and thus a signal distortion we use **Eqs. (1)** and **(10)**, so the group velocity V_g can be expressed:

$$V_g = \frac{c_0}{N} \quad \text{with} \quad N = n + \nu\frac{\mathrm{d}n}{\mathrm{d}\nu} = n - \lambda\frac{\mathrm{d}n}{\mathrm{d}\lambda} \tag{31}$$

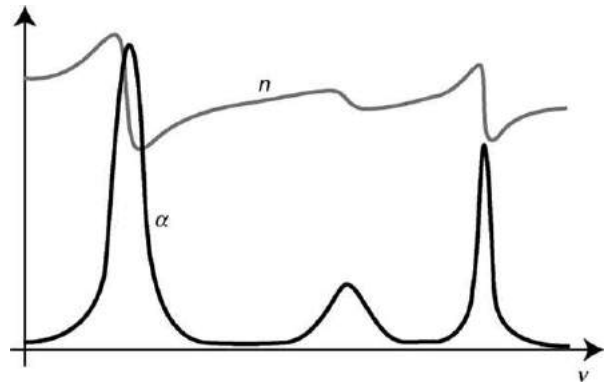


Fig. 4 Typical optical response of a transparent medium, showing the spectral interdependence between the absorption coefficient α and the index of refraction n.

N is called group velocity index and differs from the phase refractive index n only if n shows a spectral dependence. In a region of normal dispersion $(\mathrm{d}n/\mathrm{d}\nu)>0$, the group index is larger than the phase index and this is the situation observed in the great majority of transparent materials. From **Eq. (31)** and using **Eq. (24)** the GVD coefficient D_λ can be expressed as a function of the refractive index $n(\lambda)$:

$$D_\lambda = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{1}{V_\mathrm{g}}\right) = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{N}{c_0}\right) = -\frac{\lambda}{c_0}\frac{\mathrm{d}^2 n}{\mathrm{d}\lambda^2} \tag{32}$$

The GVD coefficient is proportional to the second derivative of the refractive index n with respect to wavelength and is therefore minimal close to a point of inflexion of $n(\lambda)$. As can be seen in **Fig. 4** such a point of inflexion is always present where absorption is minimal, at the largest spectral distance from two absorption lines. It means that the conditions of low-group velocity dispersion and high transparency are normally fulfilled in the same spectral region of an optical dielectric material. In this region the phase velocity dispersion is normal at any wavelength, but the group velocity is first normal for shorter wavelengths, then is zero at a definite wavelength corresponding to the point of inflexion of $n(\lambda)$, and finally becomes anomalous for longer wavelengths. The situation in which normal phase dispersion and anomalous group dispersion are observed simultaneously is in no way exceptional.

In pure silica the zero GVD wavelength is at 1273 nm, but is subject to be moderately shifted to larger wavelengths in optical fibers, as a result of the presence of doping species to raise the index in the fiber guiding core. This shift normally never exceeds 10 nm using standard dopings; larger shifts are observed resulting from waveguide dispersion and this aspect will be addressed in the next section. The zero GVD wavelength does not strictly correspond to the minimum attenuation in silica fibers, because the dominant source of loss is Rayleigh scattering in this spectral region and not molecular absorption. This scattering results from fluctuations of the medium density as observed in any amorphous materials such as vitreous silica and is therefore a collective effect of many molecules that does not impinge on the microscopic susceptibility $\chi(t)$. It has therefore no influence on the material dispersion characteristics and this explains the reason for the minimal attenuation wavelength at 1550 nm, mismatching and quite distant from the zero material GVD at 1273 nm.

Material GVD in amorphous SiO_2 can be accurately described using a three-term Sellmeier expansion of the refractive index:

$$n(\lambda) = \sqrt{1 + \sum_{j=1}^{3} \frac{C_j \lambda^2}{\lambda^2 - \lambda_j^2}} \tag{33}$$

and performing twice the wavelength derivative. The coefficients C_j and λ_j are found in most reference handbooks and result in the GVD spectrum shown in **Fig. 5**. Such a spectrum explains the absence of interest for propagation in the visible region through optical fibers, the dispersion being very important in this spectral region. It also explains the large development of optical fibers in the 1300 nm region as a consequence of the minimal material dispersion there. It must be pointed out that it is quite easy to set up propagation in an anomalous GVD regime in optical fibers, since this regime is observed in the lowest attenuation spectral region. Anomalous dispersion makes possible interesting propagation features when combined with a third-order nonlinearity as the optical Kerr effect, namely soliton propagation and efficient spectral broadening through modulation instability.

Waveguide Group Velocity Dispersion

Solutions of the wave equation in an optical dielectric waveguide such as an optical fiber are discrete and limited. These solutions, called modes, are characterized by an unchanged field distribution along the waveguides and by a uniform propagation constant β over the wavefront. This last feature is particularly important if one recall that the field extends over regions presenting different

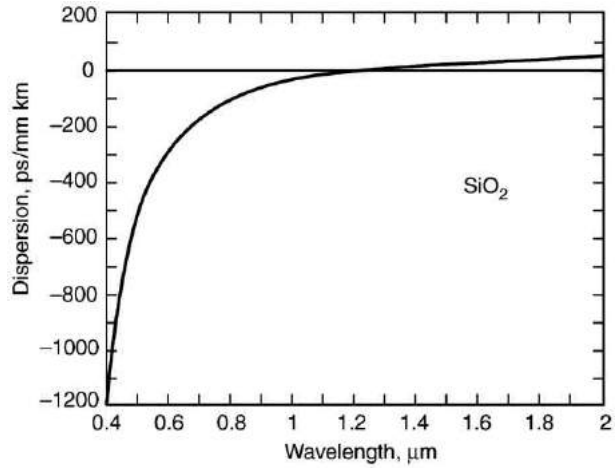


Fig. 5 Material dispersion of pure silica. The visible region (0.4–0.7 μm) shows a strong normal GVD that decreases when moving into the infrared and eventually vanishes at 1273 nm. In the minimum attenuation window (1550 nm) the material GVD is anomalous.

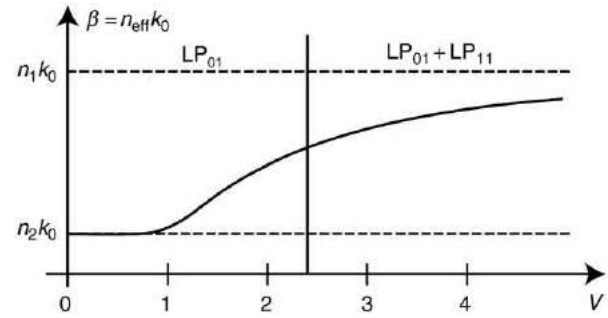


Fig. 6 Propagation constant β as a function of the normalized frequency V in a step-index optical fiber. The single mode region is in the range $0 < V < 2.405$.

refractive indices in a dielectric waveguide. For a given mode, the propagation constant β defines an effective refractive index n_{eff} for the propagation by similarity to **Eq. (1)**:

$$\beta = \frac{2\pi v}{c_0} n_{\text{eff}} \tag{34}$$

The value of this effective refractive index n_{eff} is always bound by the value of the core index n_1 and of the cladding index n_2, so that $n_2 < n_{\text{eff}} < n_1$. For a given mode, the propagation constant β, and so the effective refractive index n_{eff}, only depend on a quantity called normalized frequency V that essentially scales the light frequency to the waveguide optical parameters:

$$V = \frac{2\pi}{\lambda} a \sqrt{n_1^2 - n_2^2} \tag{35}$$

where a is the core radius. **Fig. 6** shows the dependence of the propagation constant β of the fundamental mode LP_{01} on the normalized frequency V. The variation is nonnegligible in the single-mode region and gives rise to a chromatic dispersion, since V depends on the wavelength λ, even in the fictitious case of dispersion-free refractive indices in the core and the cladding materials. This type of chromatic dispersion is called waveguide dispersion.

To find an expression for the waveguide dispersion in function of the guiding properties, let us define another normalized parameter, the normalized phase constant b, such as:

$$\beta = \frac{2\pi}{\lambda} \sqrt{n_2^2 + b(n_1^2 - n_2^2)} \tag{36}$$

The parameter b takes values in the interval $0 < b < 1$, is equal to 0 when $n_{\text{eff}} = n_2$ at the mode cutoff, and is equal to 1 when $n_{\text{eff}} = n_1$. This latter situation is never observed and is only asymptotic for very large normalized frequencies V. Solving the wave equation provides the dispersion relation $b(V)$ and it is important to point out that this relation between normalized quantities depends only on the shape of the refractive index profile. Step-index, triangular or multiple-clad index profiles will result in different $b(V)$ relations, independently of the actual values of the refractive indices n_1 and n_2 and of the core radius a.

From the definitions in **Eqs. (10)** and **(35)** and in the fictitious case of an absence of material dispersion, the propagation delay per unit length reads:

$$\frac{1}{V_g} = \frac{1}{2\pi}\frac{d\beta}{dv} = \frac{1}{2\pi}\frac{d\beta}{dV}\frac{dV}{dv} = \frac{\lambda}{2\pi}V\frac{d\beta}{dV} \tag{37}$$

so that the waveguide group velocity dispersion can be expressed using the relation in **Eq. (24)**:

$$\begin{aligned} D_\lambda^w &= \frac{d}{d\lambda}\left(\frac{1}{V_g}\right) = \frac{d}{dv}\left(\frac{1}{V_g}\right)\frac{dv}{d\lambda} \\ &= -\frac{v}{\lambda}\frac{d}{dv}\left(\frac{1}{V_g}\right) = -\frac{1}{2\pi c_0}V^2\frac{d^2\beta}{dV^2} \end{aligned} \tag{38}$$

The calculation of the combined effect of material and waveguide dispersions results in very long expressions in which it is difficult to highlight the relative effect of each contribution. Nevertheless, by making the assumption of weak guidance:

$$\frac{n_1 - n_2}{n_2} \simeq \frac{N_1 - N_2}{N_2} \ll 1 \tag{39}$$

where N_1 and N_2 are the group indices in the core and the cladding, respectively, defined in **Eq. (31)**, the complete expression can be drastically simplified to obtain for the delay per unit length:

$$\frac{1}{V_g} = \frac{1}{c_0}\left[N_2 + (N_1 - N_2)\frac{d(bV)}{dV}\right] \tag{40}$$

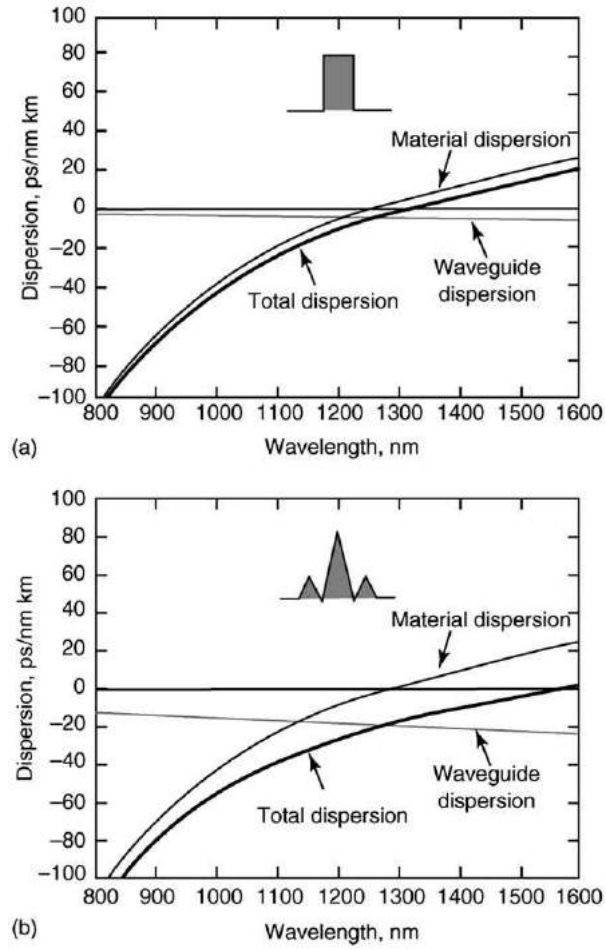


Fig. 7 Material, waveguide and total group velocity dispersions for: (a) step-index fiber; (b) triangular profile fiber. The waveguide dispersion can be significantly enhanced by changing the shape of the index profile, making possible a shift of the zero GVD to the minimum attenuation window.

and for the total dispersion:

$$D_\lambda = D_2 + (D_1 - D_2)\frac{\mathrm{d}(bV)}{\mathrm{d}V} - (N_1 - N_2)\frac{N_2}{n_1}\frac{1}{\lambda c_0}V\frac{\mathrm{d}^2(bV)}{\mathrm{d}V^2} \tag{41}$$

where

$$D_1 = -\frac{\lambda}{c_0}\frac{\mathrm{d}^2 n_1}{\mathrm{d}\lambda^2} \quad \text{and} \quad D_2 = -\frac{\lambda}{c_0}\frac{\mathrm{d}^2 n_2}{\mathrm{d}\lambda^2} \tag{42}$$

are the material GVD in the core and the cladding, respectively.

The first two terms in **Eq. (41)** represent the contribution of material dispersion weighted by the relative importance of core and cladding materials for the propagating mode. In optical fibers, the difference between D_1 and D_2 is small, so that this contribution can be often well approximated by D_2, independently of any guiding effects.

The last term represents the waveguide dispersion and is scaled by 2 factors:

- *The core-cladding index difference* $n_1 - n_2 \simeq N_1 - N_2$. The waveguide dispersion will be significantly enhanced by increasing the index difference between core and cladding.
- *The shape factor* $V(\mathrm{d}^2(bV)/\mathrm{d}V^2)$. This factor uniquely depends on the shape of the refractive index profile and may substantially modify the spectral dependence of the waveguide dispersion, making possible a great variety of dispersion characteristics.

Due to this degree of freedom brought by waveguide dispersion it is possible to shift the zero GVD wavelength to the region of minimal attenuation at 1550 nm in silica optical fibers. **Fig. 7(a)** shows the total group velocity dispersion of a step-index core fiber, together with the separate contributions of material and waveguide GVD. In this case, the material GVD is clearly the dominating contribution, while the small waveguide GVD results in a shift of the zero GVD wavelength from 1273 nm to 1310 nm. A larger shift could be obtained by increasing the core-cladding index difference, but this also gives rise to an increased attenuation from the doping and no real benefit can be expected from such a modification. In **Fig. 7(b)**, the core shows a triangular index profile to enhance the shape factor in **Eq. (41)**, so that the contribution of waveguide GVD is significantly increased with no impairing attenuation due to excessive doping. This makes it possible to realize the ideal situation of an optical fiber showing a zero GVD at the wavelength of minimum attenuation. These dispersion-shifted fibers (DSF) have now become successful in modern telecommunication networks.

Nevertheless, the absence of dispersion favors the efficiency of nonlinear effects and several classes of fibers are now proposed, showing a small but nonzero GVD at 1550 nm, with positive or negative sign, nonzero DSF (NZDSF). By interleaving fibers with positive and negative GVDs it is possible to propagate along the optical link in a dispersive medium and thus minimize the impact of nonlinearities, while maintaining the overall GVD of the link close to zero and canceling any pulse spreading accordingly.

See also: Dispersion Management. Guided Wave Optics

Further Reading

Ainslie, B.J., Day, C.R., 1986. A review of single-mode fibers with modified dispersion characteristics. IEEE Journal of Lightwave Technology LT-4 (8), 967–979.
Bass M (ed.) (1994) *Handbook of Optics*, sponsored by the Optical Society of America, 2nd edn, vol. II, chap. 10 & 33. New York: McGraw-Hill.
Buck JA (1995) In: Goodman JW (ed.) *Fundamentals of Optical Fibers*, chap. 1, 5, 6. New York: Wiley Series in Pure and Applied Optics.
Gloge, D., 1971. Dispersion in weakly guiding fibers. Applied Optics 10, 2442–2445.
Jeunhomme, L.B., 1989. Single Mode Fiber Optics: Principles and Applications., Optical Engineering Series, No. 23., New York: Marcel Dekker.
Marcuse, D., 1974. Theory of Dielectric Optical Waveguides., Series on Quantum Electronics – Principles and Applications., New York: Academic Press, chap. 2.
Marcuse, D., 1980. Pulse distortion in single-mode fibers. Applied Optics 19, 1653–1660.
Marcuse, D., 1981. Pulse distortion in single-mode fibers. Applied Optics 20, 2962–2974.
Marcuse, D., 1981. Pulse distortion in single-mode fibers. Applied Optics 20, 3573–3579.
Mazurin, O.V., Streltsina, M.V., Shvaiko-Shvaikovskaya, T.P., 1983. Silica Glass and Binary Silicate Glasses (Handbook of Glass Data; Part A), Series on Physical Sciences Data, vol. 15. Amsterdam: Elsevier.
Murata, H., 1988. Handbook of Optical Fibers and Cables., Optical Engineering Series, No. 15., New York: Marcel Dekker, chap. 2.
Saleh BEA and Teich MC (1991) In: Goodman JW (ed.) *Fundamentals of Photonics*, chap. 5, 8 & 22. New York: Wiley Series in Pure and Applied Optics.

Nonlinear Optics

K Thyagarajan and AK Ghatak, Indian Institute of Technology, New Delhi, India

Glossary

Attenuation The decrease in optical power of a propagating mode usually expressed in dB/km

Birefringent medium A medium characterized by two different modes of propagation characterized by two different polarization states of a plane light wave along any direction

Cross-phase modulation (XPM) The phase modulation of a propagating light wave due to the nonlinear change in refractive index of the medium brought about by another propagating light wave

Effective length (L_{eff}) The length over which most of the nonlinear effect is accumulated in a propagating light beam

Four-wave mixing (FWM) The mixing of four propagating light waves due to intensity dependent refractive index of the medium. This mixing leads to the generation of new frequencies from a set of two or three input light beams

Mode effective area (A_{eff}) The cross sectional area of the mode which determines the nonlinear effects on the propagating mode. This is different from the core area

Nonlinear polarization (P_{NL}) When the response of the medium to an applied electric field is not proportional to the applied electric field, then this results in a nonlinear polarization

Optical waveguides Devices which are capable of guiding optical beams by overcoming diffraction effects using the phenomenon of total internal reflection

Phase coherence length The minimum crystal length upto which the second harmonic power generated by the nonlinear interaction increases

Phase matching Condition under with the nonlinear polarization generating an electromagnetic wave propagates at the same phase velocity as the generated electromagnetic wave

Pulse dispersion The broadening of a light pulse as it propagates in an optical fiber

Quasi-phase matching (QPM) Periodic modulation of the nonlinear coefficient to achieve efficient nonlinear interaction

Second-harmonic generation (SHG) The generation of a wave of frequency 2ω from an incident wave of frequency ω through nonlinear interaction in a medium

Self-phase modulation (SPM) The modulation of the phase of a propagating light wave due to the nonlinear change in refractive index of the medium brought about by itself

Supercontinuum generation (SC) The phenomenon in which a nearly contiunuous spectrally broadband output is produced through nonlinear effects on high peak power picosecond and sub-picosecond pulses

Walk off length (L_{wo}) Length of the fiber required for two interacting pulses at different wavelengths to walk off relative to each other

Introduction

The invention of the laser provided us with a light source capable of generating extremely large optical power densities (several MW/m^2). At such large power densities, matter behaves in a nonlinear fashion and we come across new optical phenomena such as second-harmonic generation (SHG), sum and difference frequency generation, intensity dependent refractive index, mixing of various frequencies, etc. In SHG, an incident light beam at frequency ω interacts with the medium and generates a new light wave at frequency 2ω. In sum and difference frequency generation, two incident beams at frequencies ω_1 and ω_2 mix with each other producing sum $(\omega_1+\omega_2)$ and difference $(\omega_1-\omega_2)$ frequencies at the output. Higher-order nonlinear effects, such as self-phase modulation, four-wave mixing, etc. can also be routinely observed today. The field of nonlinear optics dealing with such nonlinear interactions is gaining importance due to numerous demonstrated applications in many diverse areas such as optical fiber communications, all-optical signal processing, realization of novel sources of optical radiation, etc.

Nonlinear optical interactions become prominent when the optical power densities are high and interaction takes place over long lengths. The usual method to increase optical intensity is to focus the light beam using a lens system. For a given optical power, the tighter the focusing, the larger will be the intensity for a given optical power; however, greater will be the divergence of the beam. Thus tighter focusing produces larger intensities, but over shorter interaction lengths (see Fig. 1(a)). Optical waveguides, in which the light beam is confined to a small cross-sectional area, are currently being explored for realizing efficient nonlinear devices. In contrast to bulk media, in waveguides, diffraction effects are balanced by waveguiding and the beam can have small cross-sectional areas over much longer interaction lengths (see Fig. 1(b)). A simple optical waveguide consists of a high index dielectric medium surrounded by a lower index dielectric medium so that light waves can be trapped in the high index region by the phenomenon of total internal reflection. Fig. 2 shows a planar waveguide, a channel waveguide and an optical fiber. In the planar waveguide a film of refractive index n_f is deposited/diffused on a substrate of refractive index n_s and has a cover of refractive index n_c (with n_s, $n_c < n_f$). The waveguide has typical cross-sectional dimensions of a few micrometers. Unlike planar waveguides,

 doi:10.1016/B978-0-12-809283-5.00700-X

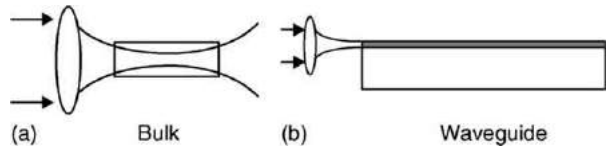


Fig. 1 (a) In bulk media, tighter focusing produces larger intensities, but over shorter interaction lengths. (b) In optical waveguides diffraction effects are balanced by waveguiding and the interaction lengths are much larger.

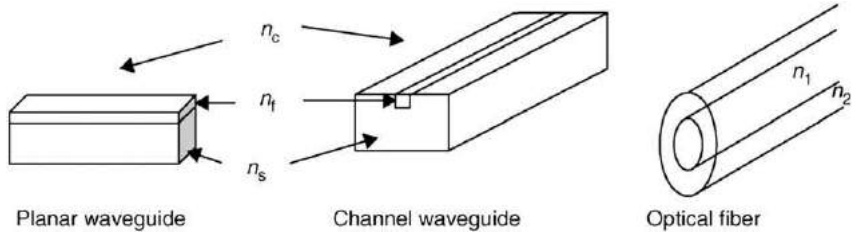


Fig. 2 A planar waveguide, a channel waveguide and an optical fiber.

in which guidance takes place only in one dimension, in channel waveguides the diffused region in a substrate is surrounded on all sides by lower index media. Optical fibers are structures with cylindrical symmetry and have a central cylindrical core of doped silica surrounded by a concentric cylindrical cladding of pure silica which has a slightly lower refractive index. Light guidance in all these waveguides takes place through the phenomenon of total internal reflection.

In contrast to bulk interactions requiring beam focusing, in the case of optical waveguides, the beam can be made to have extremely small cross-sectional areas ($\sim 25\ \mu m^2$ over very long interaction lengths (~ 20 mm in integrated optical waveguides and tens of thousands of kilometers in the case of optical fibers). This leads to very much increased nonlinear interaction efficiencies even at moderate powers (approximately a few tens of mW).

In the following sections, we will discuss some of the important nonlinear interactions that are being studied with potential applications to various branches of science and engineering. Obviously it is impossible to cover all aspects of nonlinear optics – several books have been written in this area (see Further Reading section at the end of this article) – what we will do is to discuss the physics of some of the important nonlinear effects.

Apart from intensity and length of interaction, one of the most important requirements for many efficient nonlinear interactions is the requirement of phase matching. The concept of phase matching can be easily understood from the point of view of SHG. In SHG, the incident wave at frequency ω generates a nonlinear polarization at frequency 2ω and this nonlinear polarization is responsible for the generation of the wave at 2ω. Now, the phase velocity of the nonlinear polarization wave at 2ω is the same as the phase velocity of the electromagnetic wave at frequency ω, which is usually different from the phase velocity of the electromagnetic wave at frequency 2ω; this happens due to wavelength dispersion in the medium. When the two phase velocities are unequal, the polarization wave at 2ω (which is the source) and the electromagnetic wave at 2ω pass in and out of phase with each other as they propagate through the medium. Due to this, the energy flowing in from ω to 2ω cannot add constructively and the efficiency of second-harmonic generation is limited. If the phase velocities of the waves at ω and 2ω are matched then the polarization wave and the wave at 2ω remain in phase leading to drastically increased efficiencies. This condition is referred to as phase matching and plays a very important role in most nonlinear interactions.

Nonlinear Polarization

In a linear medium, the electric polarization P is assumed to be a linear function of the electric field E:

$$P = \varepsilon_0 \chi E \tag{1}$$

where, for simplicity, a scalar relation has been written. The quantity χ is termed as linear dielectric susceptibility. At high optical intensities (which corresponds to high electric fields), all media behave in a nonlinear fashion. Thus **Eq. (1)** is modified to

$$P = \varepsilon_0 (\chi E + \chi^{(2)} E^2 + \chi^{(3)} E^3 + \ldots) \tag{2}$$

where $\chi^{(2)}, \chi^{(3)}, \ldots$ are higher order susceptibilities giving rise to the nonlinear terms. The second term on the right-hand side is responsible for SHG, sum and difference frequency generation, parametric interactions, etc. while the third term is responsible for third-harmonic generation, intensity dependent refractive index, self-phase modulation, four-wave mixing, etc. For media possessing an inversion symmetry, $\chi^{(2)}$ is zero and there is no second-order nonlinear effect. Thus silica optical fibers, which form the heart of today's communication networks, do not possess the second-order nonlinearity.

We will first discuss second-harmonic generation and parametric amplification which arises due to the second-order nonlinear term and then go on to effects due to third-order nonlinear interaction.

Second-Harmonic Generation in Crystals

The first demonstration of SHG was made in 1961 by focusing a 3 kW ruby laser pulse ($\lambda_0 = 6943$ Å) on a quartz crystal. An incident beam from a ruby laser (red color) after passing through a crystal of KDP, gets partly converted into blue light which is the second harmonic. Ever since, SHG has been one of the most widely studied nonlinear interactions.

We consider a plane wave of frequency ω propagating along the z-direction through a medium and consider the generation of the second-harmonic frequency 2ω as the beam propagates through the medium. Now, the field at ω generates a polarization at 2ω, which acts as a source for the generation of an electromagnetic wave at 2ω. Corresponding to the frequencies ω and 2ω, the electric fields are assumed to be given by

$$E^{(\omega)} = \frac{1}{2}(E_1(z)e^{i(\omega t - k_1 z)} + \text{c.c.}) \tag{3}$$

and

$$E^{(2\omega)} = \frac{1}{2}(E_2(z)e^{i(2\omega t - k_2 z)} + \text{c.c.}) \tag{4}$$

respectively; c.c. stands for the complex conjugate of the preceding quantities. The quantities:

$$k_1 = \omega\sqrt{(\varepsilon_1 \mu_0)} = (\omega/c)n^{\omega} \tag{5}$$

and

$$k_2 = 2\omega\sqrt{(\varepsilon_2 \mu_0)} = (2\omega/c)n^{2\omega} \tag{6}$$

represent the propagation vectors at ω and 2ω, respectively; ϵ_1 and ϵ_2 represent the dielectric permittivities at ω and 2ω, and n^{ω} and $n^{2\omega}$ represent the corresponding wave refractive indices. It should be noted that the amplitudes E_1 and E_2 are assumed to be z dependent – this is because at $z=0$ (where the beam is incident on the medium) the amplitude E_2 is zero and this would develop as the beam propagates through the medium. We will now develop an approximate theory for the generation of the second harmonic.

We start with the wave equation:

$$\nabla^2 E - \varepsilon\mu_0 \frac{\partial^2 E}{\partial t^2} = \mu_0 \frac{\partial^2 P_{\text{NL}}}{\partial t^2} \tag{7}$$

where P_{NL} is the nonlinear polarization. An incident wave at frequency ω generates a polarization at 2ω which acts as a source for the generation of an electromagnetic wave at 2ω.

In order to consider SHG, we write the wave equation corresponding to 2ω with the nonlinear polarization at 2ω given by

$$P_{\text{NL}}^{(2\omega)} = \frac{1}{2}(\tilde{P}_{\text{NL}}^{(2\omega)} e^{2i(\omega t - k_1 z)} + \text{c.c.}) \tag{8}$$

where

$$\tilde{P}_{\text{NL}}^{(2\omega)} = dE_1 E_1 \tag{9}$$

and

$$d = \frac{\varepsilon_0 \chi^{(2)}}{2} \tag{10}$$

represents the effective nonlinear coefficient and depends on the nonlinear material, the polarization states of the fundamental and the second harmonic and also on the propagation direction. Simple manipulations give us, for the rate of change of amplitude of the second harmonic wave:

$$\frac{dE_2}{dz} = -\frac{i\mu_0 dc\omega}{n^{2\omega}} E_1^2(z) e^{i(\Delta k)z} \tag{11}$$

where

$$\Delta k = k_2 - 2k_1 = (2\omega/c)(n^{2\omega} - n^{\omega}) \tag{12}$$

and we have assumed:

$$d^2E_2/dz^2 \ll k_2(dE_2/dz) \tag{13}$$

In order to solve **Eq. (11)** we neglect depletion of the fundamental field, i.e., $E_1(z)$ is almost a constant and the quantity E_1^2 on the right-hand side can be assumed to be independent of z. If we now integrate **Eq. (11)**, we obtain:

$$E_2(z) = -\frac{i\mu_0 dc\omega}{n^{2\omega}} z E_1^2 e^{i\sigma} \frac{\sin\sigma}{\sigma} \tag{14}$$

where

$$\sigma = \frac{1}{2}(\Delta k)z = (\omega/c)(n^{2\omega} - n^{\omega})z \tag{15}$$

Now, the powers associated with the beams corresponding to ω and 2ω are given by:

$$P_1 = \frac{n^{\omega}}{2c\mu_0} S|E_1|^2, \quad P_2 = \frac{n^{2\omega}}{2c\mu_0} S|E_2|^2 \tag{16}$$

where S represents the cross-sectional area of the beams. Substituting for $|E_2|^2$ from **Eq. (16)**, we get after some elementary simplifications:

$$\eta = \frac{P_2}{P_1} = \frac{2c^2\mu_0^3 d^2\omega^2}{(n^{\omega})^2 n^{2\omega}} z^2 \frac{P_1}{S} \left(\frac{\sin\sigma}{\sigma}\right)^2 \tag{17}$$

when η represents the SHG efficiency. Note that the efficiency of SHG increases if P_1/S the intensity of the fundamental wave increases. Also η increases if the nonlinear coefficient d increases (obviously) and if the frequency increases. However, for a given power P_1, the conversion efficiency increases if the area of the beam decreases – thus a focused beam will have a greater SHG efficiency. The most important factor is $(\sin\sigma/\sigma)^2$ which is a sharply peaked function around $\sigma = 0$, attaining a maximum value of unity at $\sigma = 0$. Thus for maximum SHG efficiency:

$$\sigma = 0 \Rightarrow n^{2\omega} = n^{\omega} \tag{18}$$

i.e., the refractive index at 2ω must be equal to the refractive index at ω – this is known as the phase matching condition.

The phase matching condition can be pictorially represented by a vector diagram (see **Fig. 3**). In **Fig. 3**, $k_1\omega$ and $k_2\omega$ represent the wave vectors of the fundamental and the second harmonic respectively. To achieve reasonably effective SHG, it is very important to satisfy the phase-matching condition. Efficiencies under nonphase matched operation can be orders of magnitude lower than under phase-matched operation.

We see from **Eq. (17)** that the smallest z for which η is maximum is:

$$z = L_c = \frac{\pi}{\Delta k} = \frac{\pi c}{2\omega(n^{2\omega} - n^{\omega})} \tag{19}$$

where we have used **Eq. (15)** for Δk. The length L_c is called the phase coherence length and represents the maximum crystal length up to which the second-harmonic power increases. Thus, if the length of the crystal is less than L_c, the second-harmonic power increases almost quadratically with z. For $z > L_c$, the second-harmonic power begins to reduce again.

In general, because of dispersion, it is very difficult to satisfy the phase-matching condition. However, in a birefringent medium, for example with $n_o > n_e$, it may be possible to find a direction along which the refractive index of the *o*-wave for ω equals the refractive index of the *e*-wave for 2ω (see **Fig. 4**). For media with $n_e > n_o$, the direction would correspond to that along which the refractive index of the *e*-wave for ω equals the refractive index for the *o*-wave for 2ω. This can readily be understood by

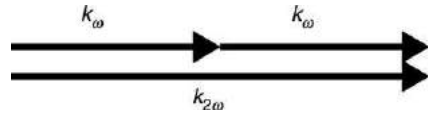


Fig. 3 Vector diagram representing the phase matching condition for SHG.

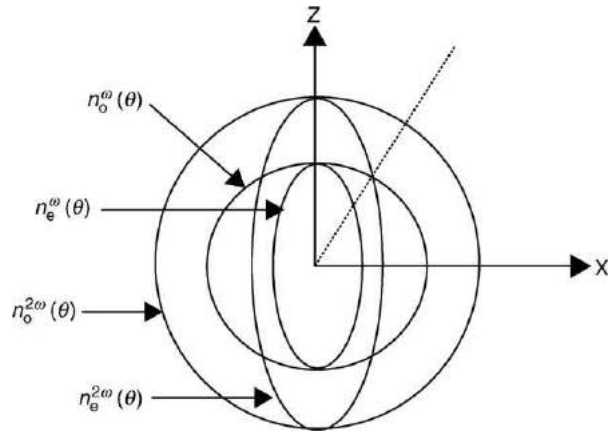


Fig. 4 In a birefringent medium, for example with $n_o > n_e$, it may be possible to find a direction along which the refractive index of the *o*-wave for ω equals the refractive index of the *e*-wave for 2ω.

considering a specific example. We consider the SHG in KDP corresponding to the incident ruby laser wavelength ($\lambda_0 = 0.6943$ μm, $\omega = 2.7150 \times 10^{15}$ Hz). For this frequency:

$$\left\{ \begin{array}{ll} n_o^{\omega} = 1.50502, & n_e^{\omega} = 1.46532 \\ n_o^{2\omega} = 1.53269, & n_e^{2\omega} = 1.48711 \end{array} \right\} \tag{20}$$

The refractive index variation for the *e*-wave is given by:

$$n_e(\theta) = \left(\frac{\sin^2\theta}{n_e^2} + \frac{\cos^2\theta}{n_o^2} \right)^{-\frac{1}{2}} \tag{21}$$

where θ is the angle that the wave propagation direction makes with the optic axis. Now, as can be seen from the above equations:

$$n_e < n_e(\theta) < n_o \tag{22}$$

Since n_o at 0.6943 μm lies between n_e and n_o at 0.34715 μm, there will always exist an angle θ_m along which $n_e^{2\omega}(\theta_m) = n_o^{\omega}$. We can solve for θ_m and obtain:

$$\cos\theta_m = \left[\frac{(n_o^{\omega})^2 - (n_e^{2\omega})^2}{(n_o^{2\omega})^2 - (n_e^{2\omega})^2} \right]^{1/2} \frac{n_o^{2\omega}}{n_o^{\omega}} \tag{23}$$

For the values given by **Eq. (20)**, we find $\theta_m = 50.5^{\circ}$.

Quasi Phase Matching (QPM)

As mentioned earlier, phase matching is extremely important for any efficient nonlinear interaction. Recently the technique of quasi phase matching (QPM) has become a convenient way to achieve phase matching at any desired wavelength in any material. QPM relies on the fact that the phase mismatch in the two interacting beams can be compensated by periodically readjusting the phase of interaction through periodically modulating the nonlinear characteristics of the medium at a spatial frequency equal to the wavevector mismatch of the two interacting waves. Thus in SHG, when the nonlinear polarization at 2ω and the electromagnetic wave at 2ω have an accumulated phase difference of π, then the sign of the nonlinear coefficient is reversed so that the energy flowing from the polarization to the wave can add constructively with the existing energy (see **Fig. 5**). Thus, by properly choosing the period of the spatial modulation of the nonlinear coefficient, one could achieve phase matching. This scheme, QPM, is being very widely studied for application to nonlinear interactions in bulk and in waveguides.

In a ferroelectric material such as lithium niobate, the signs of the nonlinear coefficients are linked to the direction of the spontaneous polarization. Thus a periodic reversal of the domains of the crystal can be used for achieving QPM (see **Fig. 6**). This is the currently used technique to obtain high-efficiency SHG and other nonlinear interactions in $LiNbO_3$, $LiTaO_3$, and KTP. The most popular technique today, to achieve periodic domain reversal in $LiNbO_3$, is the technique of electric field poling. In this method, a high electric field pulse is applied to properly oriented lithium niobate crystal using lithographically defined electrode patterns to produce a permanent periodic domain reversed pattern. Such a periodically domain reversed $LiNbO_3$ crystal with the periodically reversed domains going through the entire depth of the crystal is also referred to as PPLN (periodically poled lithium niobate, pronounced 'piplin') and is now commercially available.

In order to analyze SHG in a periodically poled material, let us assume that the nonlinear coefficient d varies sinusoidally with a period Λ. In such a case we have:

$$d = d_0 \sin(Kz) \tag{24}$$

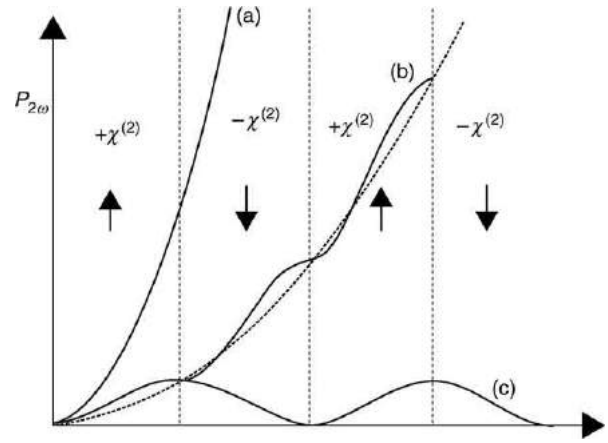


Fig. 5 By reversing the sign of the nonlinear coefficient after every coherence length, the energy in the second harmonic can be made to grow. (a) Perfect phase matching; (b) Quasi phase matching; (c) Nonphase matched.

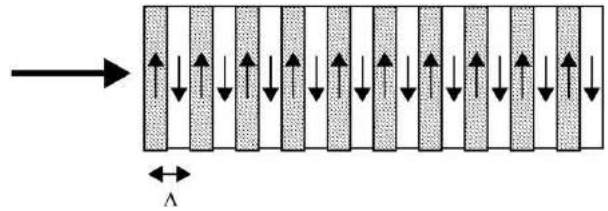


Fig. 6 QPM can be achieved by a periodic reversal of the domains of a ferroelectric material. Arrows represent the direction of the spontaneous polarization of the crystal.

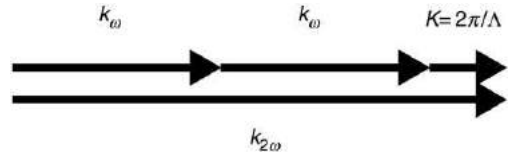


Fig. 7 Vector diagram corresponding to QPM-SHG.

where d_0 is the amplitude of modulation of the nonlinear coefficient and $K(=2\pi/\Lambda)$ represents the spatial frequency of the periodic modulation. For easier understanding we are assuming the modulation to be sinusoidal; in general, the modulation will be periodic but not sinusoidal. Any periodic modulation can be written as a superposition of sinusoidal and cosinusoidal variations. Thus our discussion is valid for one of the Fourier components of the variation.

By using **Eq. (24)**, **Eq. (11)** for the variation of the amplitude of the second harmonic becomes:

$$\frac{dE_2}{dz} = -\frac{i\mu_0 d_0 c\omega}{n^{2\omega}} E_1^2(z) e^{i(\Delta k)z} \sin(Kz) \tag{25}$$

which can be written as:

$$\frac{dE_2}{dz} = -\frac{\mu_0 d_0 c\omega}{2n^{2\omega}} E_1^2(z) e^{i(\Delta k)z} [e^{iKz} - e^{-iKz}] \tag{26}$$

Using similar arguments as earlier, it can be shown that if $\Delta k - K = 0$, then only the second term within the brackets in **Eq. (26)** contributes to the growth of the second harmonic and similarly if $\Delta k + K = 0$, then only the first term within the brackets contributes to the growth of the second harmonic. The first condition implies that:

$$2\frac{2\pi}{\lambda_0}(n^{2\omega} - n^{\omega}) - \frac{2\pi}{\Lambda} = 0 \tag{27}$$

where $\Lambda = 2\pi/K$ represents the spatial period of the modulation of the nonlinear coefficient, and λ_0 is the wavelength of the fundamental. Thus the modulation period Λ required for QPM SHG is:

$$\Lambda = \frac{\lambda_0}{2(n^{2\omega} - n^{\omega})} = 2L_c \tag{28}$$

Fig. 7 shows the vector diagram corresponding to QPM SHG. The phase mismatch between the fundamental and second harmonic is compensated by the spatial frequency vector of the periodic variation of d.

In the case of waveguides, n^{ω} and $n^{2\omega}$ would represent the effective indices of the modes at the fundamental and the second-harmonic frequency. It may be noted that the index difference between the fundamental and the second harmonic is typically 0.1 and thus the required spatial periods for a fundamental wavelength of 800 nm is ~ 4 μm.

The main advantage of QPM is that it can be used at any wavelength within the transparency window of the material; only the period needs to be correctly chosen for a specific fundamental wavelength. One can also choose appropriate polarization to make use of the largest nonlinear optical coefficients. Another advantage is the possibility of varying the domain reversal period (chirp) to achieve specific interaction characteristics such as increased bandwidth, etc.

In general, the spatial variation of the nonlinear grating is not sinusoidal. In this case the efficiency of interaction would be determined by the Fourier component of the spatial variation at the spatial frequency corresponding to the period given by **Eq. (28)**. Also, since the required periods are very small, it is possible to use a higher spatial period of modulation and use one of the Fourier components for the nonlinear interaction process. Thus, in the case of periodic reversal of the nonlinear coefficient with a spatial period given by:

$$\Lambda_g = m\frac{\lambda_0}{2(n^{2\omega} - n^{\omega})}; \quad m = 1, 3, 5, \ldots \tag{29}$$

which happens to be the m^{th} harmonic of the fundamental spatial frequency required for QPM, the corresponding nonlinear coefficient that would be responsible for SHG would be the Fourier amplitude at that spatial frequency. This can be taken into

account by defining an effective nonlinear coefficient:

$$d_{\mathrm{QPM}} = \frac{2d_0}{m\pi} \tag{30}$$

Of course the largest effective nonlinear coefficient is achieved by using the fundamental frequency with $m=1$. Higher spatial periods are easier to fabricate but would lead to reduced nonlinear efficiencies.

As compared to bulk devices, in the case of waveguides, the interacting waves are propagating in the form of modes having specific intensity distributions in the transverse plane. Because of this, the nonlinear interaction also depends on the overlap between the periodically inverted nonlinear medium, the fields of the fundamental and second-harmonic waves. Thus if $E_\omega(x, y)$ and $E_{2\omega}(x, y)$ represent the electric field distributions of the waveguide modes at the fundamental and second-harmonic frequency, then the efficiency of SHG depends on the following overlap integral:

$$I_{\mathrm{ovl}} = \int\int d_0(x,y)E_\omega^2(x,y)E_{2\omega}(x,y)\mathrm{d}x\,\mathrm{d}y \tag{31}$$

where $d_0(x, y)$ represents the transverse variation of the nonlinear coefficient of the waveguide. Thus optimization of SHG efficiency in the case of waveguide interactions has to take account of the overlap integral.

Since QPM relies on periodic phase matching, it is highly wavelength dependent. Thus the required period Λ is different for different fundamental frequencies. Any deviation in the frequency of the fundamental would lead to a reduction in the efficiency due to deviation from QPM condition. Thus the pump laser needs to be highly stabilized at a frequency corresponding to the fabricated period.

Apart from SHG, QPM has been used for other three-wave interaction processes such as difference frequency generation, parametric amplification, etc. Among the many materials that have been studied for SHG using QPM, the important ones are lithium niobate, lithium tantalite, and potassium titanyl phosphate (KTP). Many techniques have been developed for periodic domain reversal to achieve a periodic variation in the nonlinear coefficient. This includes electric field poling, electron bombardment, thermal diffusion treatment, etc.

Third-Order Nonlinear Effects

In the earlier sections, we discussed effects arising out of second-order nonlinearity, i.e., the term proportional to E^2 in the nonlinear polarization. This nonlinear term is found only in media not possessing an inversion symmetry. Thus amorphous materials or crystals possessing inversion symmetry do not exhibit second-order effects. The lowest-order nonlinear effects in such a medium is of an order three wherein the nonlinear polarization is proportional to E^3.

Self-phase modulation (SPM), cross-phase modulation (XPM) and four-wave mixing (FWM) represent some of the very important consequences of third-order nonlinearity. These effects have become all the more important as they play a significant and important role in wavelength division multiplexed optical fiber communication systems.

Self-Phase Modulation (SPM)

Consider the propagation of a plane light wave at frequency ω through a medium having $\chi^{(3)}$ nonlinearity. The polarization generated in the medium is given by

$$P = \varepsilon_0\chi E + \varepsilon_0\chi^{(3)}E^3 \tag{32}$$

If we consider a single frequency wave with an electric field given by

$$E = E_0\cos(\omega t - kz) \tag{33}$$

then

$$P = \varepsilon_0\chi E_0\cos(\omega t - kz) + \varepsilon_0\chi^{(3)}E_0^3\cos^3(\omega t - kz) \tag{34}$$

Expanding $\cos^3\theta$ in terms of $\cos\theta$ and $\cos 3\theta$, we obtain the following expression for the polarization at frequency ω:

$$P = \varepsilon_0\left(\chi + \frac{3}{4}\chi^{(3)}E_0^2\right)E_0\cos(\omega t - kz) \tag{35}$$

For a plane wave given by Eq. (33), the intensity is

$$I = \frac{1}{2}c\varepsilon_0 n_0 E_0^2 \tag{36}$$

where n_0 is the refractive index of the medium at low intensity. Then

$$P = \varepsilon_0\left(\chi + \frac{3}{2}\frac{\chi^{(3)}}{c\varepsilon_0 n_0}I\right)E \tag{37}$$

The polarization P and electric field are related through the following equation:

$$P = \varepsilon_0(n^2 - 1)E \tag{38}$$

where n is the refractive index of the medium. Comparing Eqs. (37) and (38), we get

$$n^2 = n_0^2 + \frac{3}{2}\frac{\chi^{(3)}}{c\varepsilon_0 n_0} I \tag{39}$$

where

$$n_0^2 = 1 + \chi \tag{40}$$

Since the last term in the Eq. (39) is usually very small, we get

$$n = n_0 + n_2 I \tag{41}$$

where

$$n_2 = \frac{3}{4}\frac{\chi^{(3)}}{c\varepsilon_0 n_0^2} \tag{42}$$

is the nonlinear coefficient.

For fused silica $n_0 \approx 1.47$, $n_2 \approx 3.2 \times 10^{-20}$ m^2/W and if we consider power of 100 mW having a cross-sectional area of 100 μm^2, the resultant intensity is 10^9 W/m^2 and the corresponding change in refractive index is

$$\Delta n \approx n_2 I \approx 3.2 \times 10^{-11} \tag{43}$$

Although this is very small, when the beam propagates over an optical fiber over long distances (a few hundred to a few thousand kilometers), the accumulated nonlinear effects can be significant.

In the case of an optical fiber, the light beam propagates as a mode having a specific transverse electric field distribution and thus the intensity is not constant across the cross-section. In such a case, it is convenient to express the nonlinear effect in terms of the power carried by the mode (rather than in terms of intensity). If the linear propagation constant of the mode is represented by β, then in the presence of nonlinearity, the effective propagation constant is given by

$$\beta_{\mathrm{NL}} = \beta + \frac{k_0 n_2}{A_{\mathrm{eff}}} P \tag{44}$$

where $k_0 = 2\pi/\lambda_0$, P is the power carried by the mode. The quantity A_{eff} represents the effective transverse cross-sectional area of the mode and is defined by

$$A_{\mathrm{eff}} = \frac{\left[\int_0^\infty \int_0^{2\pi} \psi^2(r) r \, \mathrm{d}r \, \mathrm{d}\phi\right]^2}{\int_0^\infty \int_0^{2\pi} \psi^4(r) r \, \mathrm{d}r \, \mathrm{d}\phi} \tag{45}$$

where $\psi(r)$ represents the transverse mode field distribution of the mode. For example, under the Gaussian approximation:

$$\psi(r) = \psi_0 e^{-r^2/w_0^2} \tag{46}$$

where ψ_0 is a constant and $2w_0$ represents the mode field diameter (MFD), we get:

$$A_{\mathrm{eff}} = \pi w_0^2 \tag{47}$$

It is usual to express the nonlinear characteristic of an optical fiber by the coefficient given by

$$\gamma = \frac{k_0 n_2}{A_{\mathrm{eff}}} \tag{48}$$

Thus, for the same input power and same wavelength, smaller values of A_{eff} lead to greater nonlinear effects in the fiber. Typically:

$$\begin{aligned} &A_{\mathrm{eff}} \sim 50 - 80\mu\mathrm{m}^2 \quad \text{and} \\ &\gamma \approx 2.4\mathrm{W}^{-1}\mathrm{km}^{-1} \quad \text{to} \quad 1.5\mathrm{W}^{-1}\mathrm{km}^{-1} \end{aligned} \tag{49}$$

When a light beam propagates through an optical fiber, the power decreases because of attenuation. Thus, the corresponding nonlinear effects also reduce. Indeed, the phase change suffered by a beam in propagating from 0 to L is given by

$$\phi = \int_0^L \beta_{\mathrm{NL}} \quad \mathrm{d}z = \beta L + \gamma \int_0^L P \mathrm{d}z \tag{50}$$

If α represents the attenuation coefficient, then

$$P(z) = P_0 \mathrm{e}^{-\alpha z} \tag{51}$$

and we get

$$\phi = \beta L + \gamma P_0 L_{\mathrm{eff}} \tag{52}$$

where

$$L_{\mathrm{eff}} = \frac{1 - \mathrm{e}^{-\alpha L}}{\alpha} \tag{53}$$

is referred to as the effective length. For $\alpha L \gg 1$, $L_{\mathrm{eff}} \approx 1/\alpha$ and for $\alpha L \gg 1$, $L_{\mathrm{eff}} \approx L$.

The effective length represents the length of the fiber over which most of the nonlinear effects has accumulated. For a loss coefficient of 0.20 dB/km, $L_{\text{eff}} \approx 21$ km.

If we consider a fiber length much longer than L_{eff}, then to have reduced impact of SPM, we must have

$$\gamma P_0 L_{\text{eff}} \ll 1 \tag{54}$$

or

$$P_0 \ll \frac{1}{\gamma L_{\text{eff}}} \approx \frac{\alpha}{\gamma} \tag{55}$$

For $\alpha = 4.6 \times 10^{-2}$ km^{-1} (which corresponds to an attenuation of 0.2 dB/km) and $\gamma = 2.4$ W^{-1} km^{-1}, we get

$$P_0 \ll 19\text{mW} \tag{56}$$

Propagation of a Pulse

When an optical pulse propagates through a medium, it suffers from the following effects:

(i) attenuation;
(ii) dispersion; and
(iii) nonlinearity.

Attenuation refers to the reduction in the pulse energy due to various mechanisms, such as scattering, absorption, etc. Dispersion is caused by the fact that a light pulse consists of various frequency components and each frequency component travels at a different group velocity. Dispersion causes the temporal width of the pulse to change; in most cases it results in an increase in pulse width, however, in some cases the temporal width could also decrease. Dispersion is accompanied by chirping, the variation of the instantaneous frequency of the pulse within the pulse duration. Since both attenuation and dispersion cause a change in the temporal variation of the optical power, they closely interact with nonlinearity in deciding the pulse evolution as it propagates through the medium.

Let $E(x, y, z, t)$ represent the electric field variation of an optical pulse. It is usual to express E in the following way:

$$E(x, y, z, t) = \frac{1}{2}[A(z, t)\psi(x, y)e^{i(\omega_0 t - \beta_0 z)} + \text{c.c.}] \tag{57}$$

where $A(z, t)$ represents the slowly varying complex envelope of the pulse, $\psi(x, y)$ represents the transverse electric field distribution, ω_0 represents center frequency, and β_0 represents the propagation constant at β_0.

In the presence of attenuation, second-order dispersion and third-order nonlinearity, the complex envelope $A(z, t)$ can be shown to satisfy the following equation:

$$\frac{\partial A}{\partial z} = -\frac{\alpha}{2}A - \beta_1 \frac{\partial A}{\partial t} + i\frac{\beta_2}{2}\frac{\partial^2 A}{\partial t^2} - i\gamma|A|^2 A \tag{58}$$

Here

$$\beta_1 = \left|\frac{d\beta}{d\omega}\right|_{\omega = \omega_0} = \frac{1}{v_g} \tag{59}$$

represents the inverse of the group velocity of the pulse, and

$$\beta_2 = \left|\frac{d^2\beta}{d\omega^2}\right|_{\omega = \omega_0} = -\frac{\lambda_0^2}{2\pi c}D \tag{60}$$

where D represents the group velocity dispersion (measured in ps/km nm).

The various terms on the RHS of the **Eq. (58)** represent the following:

I term: attenuation
II term: group velocity term
III term: second-order dispersion
IV term: nonlinear term

If we change to a moving frame defined by coordinates $T = t - \beta_1 z$, **Eq. (58)** becomes

$$\frac{\partial A}{\partial z} = -\frac{\alpha}{2}A + i\frac{\beta_2}{2}\frac{\partial^2 A}{\partial T^2} - i\gamma|A|^2 A \tag{61}$$

If we neglect the attenuation term, we obtain the following equation which is also referred to as the nonlinear Schrödinger equation:

$$\frac{\partial A}{\partial z} = i\frac{\beta_2}{2}\frac{\partial^2 A}{\partial T^2} - i\gamma|A|^2 A \tag{62}$$

The above equation has a solution given by

$$A(z,t) = A_0 \operatorname{sech}\sigma T e^{-igz} \tag{63}$$

with

$$A_0^2 = -\frac{\beta_2}{\gamma}\sigma^2, \quad g = -\frac{\sigma^2}{2}\beta_2 \tag{64}$$

Eq. (63) represents an envelope soliton and has the property that it propagates undispersed through the medium. The full width at half maximum (FWHM) of the pulse envelope will be given by $\tau_f = 2\tau_0$, where

$$\operatorname{sech}^2 \sigma\tau_0 = \frac{1}{2} \tag{65}$$

which gives the FWHM τ_f:

$$\tau_f = 2\tau_0 = \frac{2}{\sigma}\ln(1+\sqrt{2}) \approx \frac{1.7627}{\sigma} \tag{66}$$

The peak power of the pulse is:

$$P_0 = |A_0|^2 = \frac{|\beta_2|}{\gamma}\sigma^2 \tag{67}$$

Replacing σ by τ_f, we obtain

$$P_0\tau_f^2 \approx \frac{\lambda_0^2}{2c\gamma}D \tag{68}$$

where we have used Eq. (60). The above equation gives the required peak for a given σ by τ_f for the formation of a soliton pulse.

As an example, we have σ by $\tau_f = 10$ ps, $\gamma = 2.4\ \mathrm{W}^{-1}\ \mathrm{km}^{-1}$, $\lambda_0 = 1.55\ \mu$m, $D = 2$ ps/km nm and the required peak power will be $P_0 = 33$ mW.

Soliton pulses are being extensively studied for application to long distance optical fiber communication. In actual systems, the pulses have to be optically amplified at regular intervals to compensate for the loss suffered by the pulses. The amplification could be carried out using erbium doped fiber amplifiers (EDFAs) or fiber Raman amplifiers.

Spectral Broadening due to SPM

In the presence of only nonlinearity, Eq. (61) becomes

$$\frac{dA}{dz} = -i\gamma|A|^2A \tag{69}$$

whose solution is given by

$$A(z,t) = A(z=0,t)e^{-i\gamma Pz} \tag{70}$$

where $P = |A|^2$ is the power in the pulse. If P is a function of time, then time dependent phase term at $z = L$ becomes

$$e^{i\varphi(t)} = e^{i[\omega_0 t - \gamma P(t)L]} \tag{71}$$

We can define an instantaneous frequency as

$$\omega(t) = \frac{d\varphi}{dt} = \omega_0 - \gamma L\frac{dP}{dt} \tag{72}$$

for a Gaussian pulse:

$$P = P_0 e^{-2T^2/\tau_0^2} \tag{73}$$

giving

$$\omega(t) = \omega_0 + \frac{4\gamma LTP_0 e^{-2T^2/\tau_0^2}}{\tau_0^2} \tag{74}$$

Thus the instantaneous frequency within the pulse changes with time, leading to chirping of the pulse (see Fig. 8). Note that since the pulsewidth has not changed, but the pulse is chirped, the frequency spectrum of the pulse has increased. Thus SPM leads to the generation of new frequencies. By Fourier transform theory, an increased spectral width implies that the pulse can now be compressed in the temporal domain by passing it through a medium with the proper sign of dispersion. This is indeed one of the standard techniques to realize ultrashort femtosecond optical pulses.

Cross Phase Modulation (XPM)

Like SPM, cross-phase modulation also arises due to the intensity dependence of refractive index, leading to spectral broadening. Unlike SPM, in the case of XPM, intensity variations of a light beam at a particular frequency modulate the phase of light beam at another frequency. If the signals at both frequencies are pulses, then due to difference in group velocities of the pulses, there is a

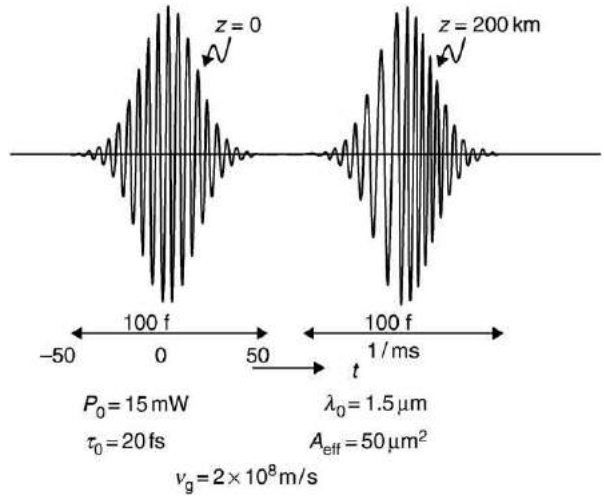


Fig. 8 Due to self phase modulation, the instantaneous frequency within the pulse changes with time leading to chirping of the pulse.

walk off between the two pulses, i.e., if they start together, they will separate as they propagate through the medium. Nonlinear interaction takes place as long as they physically overlap in the medium. Smaller the dispersion, smaller will be the difference in group velocities (assuming closely spaced wavelengths) and the longer they will overlap. This would lead to stronger XPM effects. At the same time, if two pulses pass through each other, then since one pulse will interact with both the leading and the trailing edge of the other pulse, XPM effects will be nil provided there is no attenuation. In the presence of attenuation in the medium, the pulse will still get modified due to XPM. A similar situation can occur when the two interacting pulses are passing through an optical amplifier.

To study XPM, we assume simultaneous propagation of two waves at two different frequencies through the medium. If ω_1 and ω_2 represent the two frequencies, then one obtains for the variation of the amplitude A_1 of the frequency ω_1:

$$\frac{dA_1}{dz} = -i\gamma(\tilde{P}_1 + 2\tilde{P}_2)A_1 \tag{75}$$

where $\tilde{P}_1$ and $\tilde{P}_2$ represent the powers at frequencies ω_1 and ω_2, respectively. The first term in **Eq. (75)** represents SPM while the second term corresponds to XPM. If the powers are assumed to attenuate at the same rate, i.e.:

$$\tilde{P}_1 = P_1 e^{-\alpha z}, \quad \tilde{P}_2 = P_2 e^{-\alpha z} \tag{76}$$

then the solution of **Eq. (75)** is

$$A_1(L) = A_1(0)e^{-i\gamma(P_1+2P_2)L_{\text{eff}}} \tag{77}$$

where, as before, L_{eff} represents the effective length of the medium. When we are studying the effect of power at ω_2 on the light beam at frequency ω^1 we will refer to the wave at frequency ω_2 as pump, and the wave at frequency ω_1 as probe or signal. From **Eq. (77)** it is apparent that the phase of signal at frequency ω_1 is modified by the power at another frequency. This is referred to as XPM. Note also that XPM is twice as effective as SPM.

Similar to the case of SPM, we can now write for the instantaneous frequency in the presence of XPM as **Eq. (72)**:

$$\omega(t) = \omega_0 - 2\gamma L_{\text{eff}} \frac{dP_2}{dt} \tag{78}$$

Hence the part of the signal that is influenced by the leading edge of the pump will be down-shifted in frequency (since in the leading edge $dP_2/dt > 0$) and the part overlapping with the trailing edge will be up-shifted in frequency (since $dP_2/dt < 0$). This leads to a frequency chirping of the signal pulse just as in the case of SPM.

If the probe and pump beams are pulses, then XPM can lead to induced frequency shifts depending on whether the probe pulse interacts only with the leading edge or trailing edge or both, as they both propagate through the medium. Let us consider a case when the group velocity of pump pulse is greater than that of the probe pulse. Thus, if both pulses enter the medium together, then since the pump pulse travels faster, the probe pulse will interact only with the trailing edge of the pump. Since in this case dP_2/dt is negative, the probe pulse suffers a blue-induced frequency shift. Similarly if the pulses enter at different instants but completely overlap at the end of the medium, then $dP_2/dt > 0$ and the probe pulse would suffer a red-induced frequency shift. Indeed, if the two pulses start separately and walk through each other, then there is no induced shift due to cancellation of shifts induced by leading and trailing edges of the pump. **Fig. 9** shows measured induced frequency shift of 532 nm probe pulse as a function of the delay between this pulse and a pump pulse at 1064 nm, when they propagate through a 1 m long optical fiber.

When pulses of light at two different wavelengths propagate through an optical fiber, due to different group velocities of the pulses, they pass through each other, resulting in what could be termed as a collision. In the linear case, the pulses will pass through without affecting each other, but when intensity levels are high, XPM induces phase shifts in both pulses. We can define a

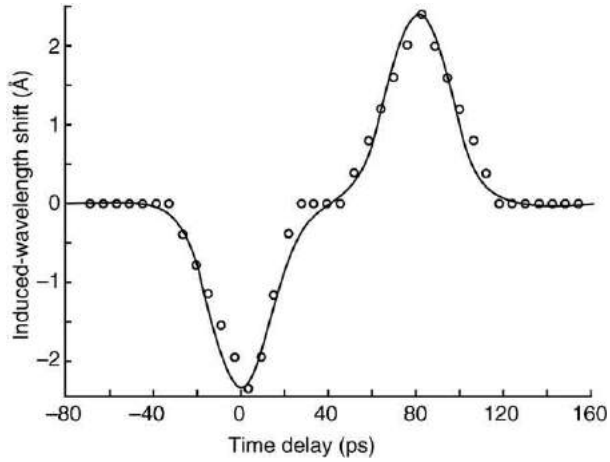


Fig. 9 Measured induced frequency shift of 532 nm probe pulse as a function of the delay between this pulse and a pump pulse at 1064 nm, when they propagate through a 1 m long optical fiber. Reproduced from Baldeck PL, Alfano RR and Agrawal GP (1998) Induced frequency shift of copropagating ultrafast optical pulses. *Applied Physics Letters* 52: 1939–1941 with permission from the American Institute of Physics.

parameter termed walk off length L_{wo} which is the length of the fiber required for the interacting pulses to walk off relative to each other. The walk off length is given by

$$L_{wo} = \frac{\Delta\tau}{D\Delta\lambda} \tag{79}$$

where D represents the dispersion coefficient, and $\Delta\lambda$ represents the wavelength separation between the interacting pulses. For return to zero (RZ) pulses, $\Delta\tau$ represents the pulse duration while for nonreturn to zero (NRZ) pulses, $\Delta\tau$ represents the rise term or fall time of the pulse. Closely spaced channels will thus interact over longer fiber lengths, thus leading to greater XPM effects. Larger dispersion coefficients will reduce L_{wo} and thus the effects of XPM. Since the medium is attenuating, the power carried by the pulses decreases as they propagate and thus leading to reduced XPM effect. The characteristic length for attenuation is the effective length L_{eff}, defined by **Eq. (53)**. If $L_{wo} \ll L_{eff}$, then over the length of interaction of the pulses, the intensity levels do not change appreciably and the magnitude of the XPM-induced effects will be proportional to the wavelength spacing $\Delta\lambda$. For small $\Delta\lambda$'s, $L_{wo} \gg L_{eff}$ and the interaction length is now determined by the fiber losses (rather than by walk off) and the XPM-induced effects become almost independent of $\Delta\lambda$ Indeed, if we consider XPM effects between a continuous wave (cw) probe beam and a sinusoidally industry modulated pump beam, then the amplitude of the XPM-induced phase shift ($\Delta\Phi_{pr}$) in the probe beam is given by:

$$\begin{aligned} \Delta\Phi_{pr} &\approx 2\gamma P_{2m} L_{eff} \quad \text{for} \quad L_{wo} \gg L_{eff} \\ \Delta\Phi_{pr} &\approx 2\gamma P_{2m} L_{wo} \quad \text{for} \quad L_{wo} \ll L_{eff} \end{aligned} \tag{80}$$

Here P_{2m} is the amplitude of the sinusoidal power modulation of the pump beam.

XPM-induced intensity interference can be studied by simultaneously propagating an intensity modulated pump signal and a cw probe signal at a different wavelength. The intensity modulated signal will induce phase modulation on the cw probe signal and the dispersion of the medium will convert the phase modulation to intensity modulation of the probe. Thus, the magnitude of the intensity fluctuation of the probe signal serves as an estimate of the XPM induced interference. **Fig. 10** shows the intensity fluctuations on a probe signal at 1550 nm, induced by a modulated pump for a channel separation of 0.6 nm. **Fig. 11** shows the variation of the RMS value of probe intensity modulation with the wavelength separation between the intensity modulated signal and the probe. The experiment has been performed over four amplified spans of 80 km of standard single mode fiber (SMF) and nonzero dispersion shifted fiber (NZDSF). The large dispersion in SMF has been compensated using dispersion compensating chirped gratings. The probe modulation, in the case of SMF, decreases approximately linearly with $1\Delta\lambda$ for all $\Delta\lambda$s; the modulation is independent of $\Delta\lambda$. In contrast the NZDSF shows a constant modulation for $\Delta\lambda \leq 1$ nm. This is consistent with the earlier discussion in terms of L_{wo} and L_{eff}.

Four-Wave Mixing (FWM)

Four-wave mixing (FWM) is a nonlinear interaction that occurs in the presence of multiple wavelengths in a medium, leading to the generation of new frequencies. Thus, if light waves at three different frequencies ω_2, ω_3, ω_4 are launched simultaneously into a medium, the same nonlinear polarization that led to intensity dependent refractive index, leads to nonlinear polarization component at a frequency:

$$\omega_1 = \omega_3 + \omega_4 - \omega_2 \tag{81}$$

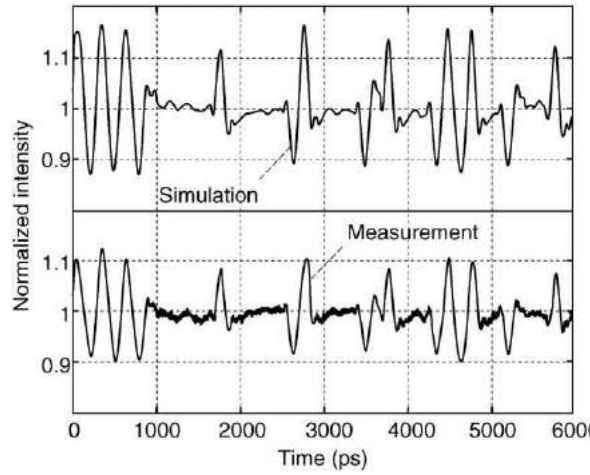


Fig. 10 Intensity fluctuations induced by cross phase modulation on a probe signal at 1550 nm by a modulated pump for a channel separation of 0.6 nm. Reproduced with permission from Rapp L (1997) Experimental investigation of signal distortions induced by cross phase modulation combined with dispersion. *IEEE Photon Technical Letters* 9: 1592–1595, © 2004 IEEE.

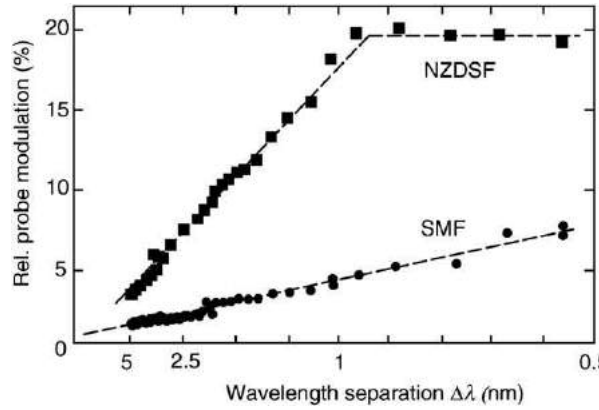


Fig. 11 Variation of the RMS value of probe intensity modulation with the wavelength separation between the intensity modulated signal and the probe. Reproduced with permission from Shtaif M, Eiselt M and Garret LD (2000) Cross-phase modulation distortion measurements in multispan WDM systems. *IEEE Photon Technical Letters* 12: 88–90, © 2004 IEEE.

This nonlinear polarization, under certain conditions, leads to the generation of electromagnetic waves at ω_1. This process is referred to as four-wave mixing due to the interaction between four different frequencies. Degenerate four-wave mixing corresponds to the case when two of the input waves have the same frequency. Thus, if $\omega_3 = \omega_4$, then inputting waves at ω_2 and ω_3 leads to the generation of waves at the frequency

$$\omega_1 = 2\omega_3 - \omega_2 \tag{82}$$

During FWM process, there are four different frequencies present at any point in the medium. If we write the electric field of the waves as

$$E_i = \frac{1}{2}[A_i(z)\psi_i(x,y)e^{i(\omega_i t - \beta_i z)} + \text{c.c.}], \quad i = 1,2,3,4 \tag{83}$$

where, as before, $A_i(z)$ represents the amplitude of the wave, $\psi_i(x, y)$ the transverse field distribution, and β_i the propagation constant of the wave. The total electric field is given by

$$E = E_1 + E_2 + E_3 + E_4 \tag{84}$$

Substituting for the total electric field in the equation for nonlinear polarization, the term with frequency ω_1 comes out as

$$P_{\text{NL}}^{(\omega_1)} = \frac{1}{2}[P_{\text{NL}}^{(\omega_1)} e^{i(\omega_1 t - \beta_1 z)} + \text{c.c}] \tag{85}$$

where

$$P_{\mathrm{NL}}^{(\omega_1)} = \frac{3\varepsilon_0}{2}\chi^{(3)}A_2^* A_3 A_4 \psi_2\psi_3\psi_4\, \mathrm{e}^{-\mathrm{i}\Delta\beta z} \tag{86}$$

and

$$\Delta\beta = \beta_3 + \beta_4 - \beta_2 - \beta_1 \tag{87}$$

In writing Eq. (86), we have only considered the FWM term, neglecting the SPM and XPM terms.

Substituting the expression for $P_{\mathrm{NL}}^{(\omega_1)}$ in the wave equation for ω_1 and making the slowly varying approximation (in a manner similar to that employed in the case of SPM and XPM), we obtain the following equation for $A_1(z)$:

$$\frac{\mathrm{d}A_1}{\mathrm{d}z} = -2\mathrm{i}\gamma A_2^* A_3 A_4 \mathrm{e}^{-\mathrm{i}\Delta\beta z} \tag{88}$$

where γ is defined by Eq. (48) with $k_0 = \omega/c$, ω representing the average frequency of the four interacting waves, and A_{eff} is the average effective area of the modes.

Assuming all waves to have the same attenuation coefficient α and neglecting depletion of waves at frequencies ω_2, ω_3 and ω_4, due to nonlinear conversion we obtain for the power in the frequency ω_1 as

$$P_1(L) = 4\gamma^2 P_2 P_3 P_4 L_{\mathrm{eff}}^2 \eta\, \mathrm{e}^{-\alpha L} \tag{89}$$

where

$$\eta = \frac{\alpha^2}{\alpha^2 + \Delta\beta^2}\left[1 + \frac{4\mathrm{e}^{-\alpha L}\sin^2\frac{\Delta\beta L}{2}}{(1-\mathrm{e}^{-\alpha L})^2}\right] \tag{90}$$

and L_{eff} is the effective length (see Eq. (53)). Maximum four-wave mixing takes place when $\Delta\beta = 0$, since in such a case $\eta = 1$. Now:

$$\Delta\beta = \beta(\omega_3) + \beta(\omega_4) - \beta(\omega_2) - \beta(\omega_1) \tag{91}$$

Since the frequencies are usually close to each other, we can make a Taylor series expansion about any frequency, say ω_2. In such a case, we obtain

$$\Delta\beta = (\omega_3 - \omega_2)(\omega_3 - \omega_1)\left|\frac{\mathrm{d}^2\beta}{\mathrm{d}\omega^2}\right|_{\omega=\omega_2} \tag{92}$$

In optical fiber communication systems, the channels are usually equally spaced. Thus, we assume the frequencies to be given by

$$\begin{aligned} &\omega_4 = \omega_2 + \Delta\omega, \quad \omega_3 = \omega_2 - 2\Delta\omega \quad \text{and} \\ &\omega_1 = \omega_2 - \Delta\omega \end{aligned} \tag{93}$$

Using these frequencies and Eq. (60), Eq. (92) gives us:

$$\Delta\beta = -\frac{4\pi D\lambda^2}{c}(\Delta\nu)^2 \tag{94}$$

where $\Delta\omega = 2\pi\Delta\nu$. Thus maximum FWM takes place when $D = 0$. This is the main problem in using wavelength division multiplexing (WDM) in dispersion shifted fibers which are characterized by zero dispersion at the operating wavelength of 1550 nm, as FWM will then lead to crosstalk among various channels. FWM efficiency can be reduced by using fiber with nonzero dispersion. This has led to the development of nonzero dispersion shifted fiber (NZDSF) which have a finite nonzero dispersion of about ± 2 ps/km nm at the operating wavelength.

From Eq. (94), we notice that for a given dispersion coefficient D, FWM efficiency will reduce as $\Delta\nu$ increases.

In order to get a numerical appreciation, we consider a case with $D = 0$, i.e., $\Delta\beta = 0$. For such a case $\eta = 1$. If all channels were launched with equal power P_{in} then:

$$P_1(L) = 4\lambda^2 P_{\mathrm{in}}^3 L_{\mathrm{eff}}^2\, \mathrm{e}^{-\alpha L} \tag{95}$$

Thus the ratio of power generated at ω_1, due to FWM and that existing at the same frequency, is

$$\frac{P_{\mathrm{g}}}{P_{\mathrm{out}}} = \frac{P_1(L)}{P_{\mathrm{in}}\ \mathrm{e}^{-\alpha L}} = 4\gamma^2 P_{\mathrm{in}}^2 L_{\mathrm{eff}}^2 \tag{96}$$

Typical values are $L_{\mathrm{eff}} = 20$ km, $\gamma = 2.4$ W^{-1} km^{-1}. Thus:

$$\frac{P_{\mathrm{g}}}{P_{\mathrm{out}}} \approx 0.01 P_{\mathrm{in}}^2(\mathrm{mW}^2) \tag{97}$$

Fig. 12 shows the output spectrum measured at the output of a 25 km-long dispersion shifted fiber ($D = -0.2$ ps/km nm) when three 3 mW wavelengths are launched simultaneously. Notice the generation of many new frequencies by FWM. Fig. 13 shows the ratio of generated power to the output as a function of channel spacing $\Delta\lambda$ for different dispersion coefficients. It can be seen that by choosing a nonzero value of dispersion, the four-wave mixing efficiency can be reduced. Larger the dispersion coefficient, smaller can be the channel spacing for the same crosstalk.

Since dispersion leads to increased bit error rates in fiber optic communication systems, it is important to have low dispersion. On the other hand, lower dispersion leads to crosstalk due to FWM. This problem can be resolved by noting that FWM depends on

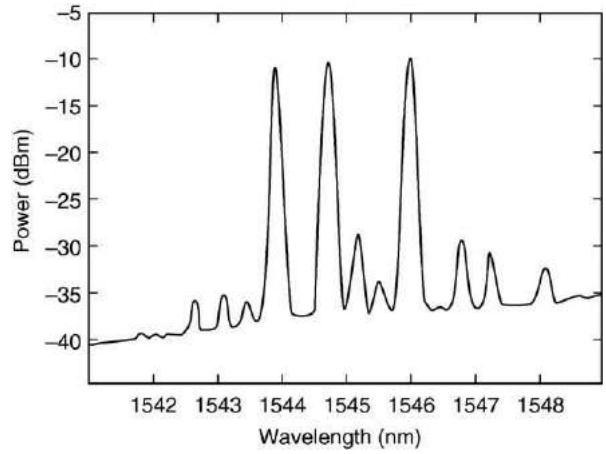


Fig. 12 Generation of new frequencies because of FWM when waves at three frequencies are incident in the fiber. Reproduced with permission from Tkach RW, Chraplyvy AR, Forghiari F, Gnanck AH and Derosier RM (1995) Four-photon mixing and high speed WDM systems. *Journal of lightwave Technology* 13: 841–849, © 2004 IEEE.

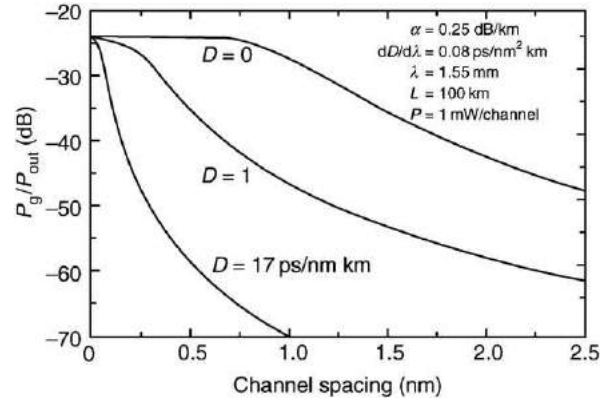


Fig. 13 Ratio of generated power to the output as a function of channel spacing $\Delta\lambda$ for different dispersion coefficients. Reproduced with permission from Tkach RW, Chraplyvy AR, Forghiari F, Gnanck AH and Derosier RM (1995) Four-photon mixing and high speed WDM systems. *Journal of lightwave Technology* 13: 841–849, © 2004 IEEE.

the local dispersion value in the fiber, while the pulse spreading at the end of a link depends on the overall dispersion in the fiber link. If one chooses a link made up of positive and negative dispersion coefficients, then by an appropriate choice of the lengths of the positive and negative dispersion fibers, it would be possible to achieve a zero total link dispersion while at the same time maintaining a large local dispersion. This is referred to as dispersion management in fiber optic systems.

Although FWM leads to crosstalk among different wavelength channels in an optical fiber communication system, it can be used for various optical processing functions such as wavelength conversion, high-speed time division multiplexing, pulse compression, etc. For such applications, there is a concerted worldwide effort to develop highly nonlinear fibers with much smaller mode areas and higher nonlinear coefficients. Some of the very novel fibers that have been developed recently include holey fibers, photonic bandgap fibers, or photonic crystal fibers which are very interesting since they possess extremely small mode effective areas ($\sim$2.5 μm^2 at 1550 nm) and can be designed to have zero dispersion even in the visible region of the spectrum. This is expected to revolutionize nonlinear fiber optics by providing new geometries to achieve highly efficient nonlinear optical processing at lower powers.

Supercontinuum Generation

Supercontinuum (SC) generation is the phenomenon in which a nearly continuous spectrally broadened output is produced through nonlinear effects on high peak power picosecond and subpicosecond pulses. Such broadened spectra find applications in spectroscopy, wavelength characterization of optical components such as a broadband source from which many wavelength channels can be obtained by slicing the spectrum.

Supercontinuum generation in an optical fiber is a very convenient technique since it provides a very broad bandwidth ($>$200 nm), the intensity levels can be maintained high over long interaction lengths by choosing small mode areas, and the dispersion profile of the fiber can be appropriately designed by varying the transverse refractive index profile of the fiber.

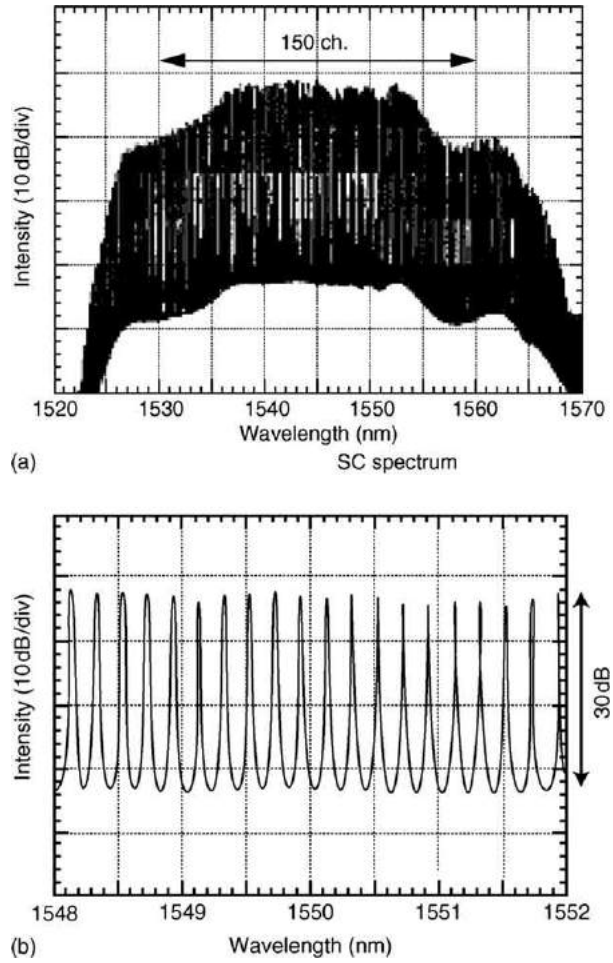


Fig. 14 Super Continuum spectrum generated by passing 25 GHz optical pulse train at 1544 nm generated by a mode locked laser diode and amplified by an erbium doped fiber. Reproduced from Yamada E, Takara H, Ohara T, Satc K, Morioka T, Jinguji K, Itoh M and Ishi M (2001) A high SNR, 150 Ch. Supercontinuum CW optical source with high SNR and precise 25 GHz spacing for 10 Gbit/s DWDM systems. *Electronics Letters* 37: 304–306 with permission from IEEE.

The spectral broadening that takes place in the fiber is attributed to a combination of various third-order effects such as SPM, XPM, FWM, and Raman scattering. Since dispersion plays a significant role in the temporal evolution of the pulse, different dispersion profiles have been used in the literature to achieve broadband SC.

Some studies have used dispersion decreasing fibers, dispersion flattened fibers, while others have used a constant anamolous dispersion fiber followed by a normal dispersion fiber.

Fig. 14 shows the SC spectrum generated by passing 25 GHz optical pulse train at 1544 nm generated by a mode locked laser diode and amplified by an erbium doped fiber. The fiber used for SC generation is a polarization maintaining dispersion flattened fiber. The output is a broad spectrum containing more than 150 spectral components at a spacing of 25 GHz with a flat top spectrum over 18 nm. The output optical powers range from -9 dBm to $+3$ dBm. Such sources are being investigated as attractive solutions for dense WDM (DWDM) optical fiber communication systems.

Conclusions

The advent of lasers provided us with an intense source of light and led to the birth of nonlinear optics. Nonlinear optical phenomena exhibit a rich variety of effects and coupled with optical waveguides and fibers, such effects can be observed even at moderate optical powers. With the realization of compact femtosecond lasers, highly nonlinear holey optical fibers, quasi phase matching techniques, etc, the field of nonlinear optics is expected to grow further and lead to commercial exploitation such as in compact blue lasers, soliton laser sources and multiwavelength sources and for fiber optic communication.

See also: Guided Wave Optics. Nonlinear Effects (Basics)

Further Reading

Agrawal, G.P., 1989. Nonlinear Fiber Optics. Boston, MA: Academic Press.
Armstrong, J.A., Bloembergen, N., Ducuing, J., Pershan, P.S., 1962. Interactions between light waves in a nonlinear dielectric. Physics Review 127, 1918.
Baldeck, P.L., Alfano, R.R., Agrawal, G.P., 1998. Induced frequency shift of copropagating ultrafast optical pulses. Applied Physics Letters 52, 1939–1941.
Baldi P (1992) *Generation de fluorescence parametrique guidee sur niobate et tantalite de lithium polarizes periodiquement. Etudes preliminaire d'un oscillateur parmetrique optique integree*. Doctoral Thesis of the University of Nice, France.
Chiang, T.K., Kagi, N., Marhic, M.E., Kazovsky, L.G., 1996. Cross phase modulation in fiber links with multiple optical amplifiers and dispersion compensators. Journal of Lightwave Technology 14, 249–259.
Franken, P.A., Hill, A.E., Peter, C.W., Weinreich, G., 1961. Generation of optical harmonics. Physics Review Letters 7, 118.
Ghatak, A.K., Thyagarajan, K., 1987. Optical Electronics. Cambridge, UK: Cambridge University Press.
Ghatak, A.K., Thyagarajan, K., 1998. Introduction to Fiber Optics. Cambridge: Cambridge University Press.
Lim, E.J., Fejer, M.M., Byer, R.L., 1989. Second-harmonic generation green light in periodically poled planar lithium niobate waveguide. Electronic Letters 25, 174.
Lim, E.J., Fejer, M.M., Byer, R.L., Kozovsky, W.J., 1989. Electronic Letters 25, 731–732.
Myers, L.E., Bosenberg, W.R., 1997. Periodically poled lithium niobate and quasi phase matcher parametric oscillators. IEEE Journal of Quantum Electronics 33, 1663–1672.
Newell, A.C., Moloney, J.V., 1992. Nonlinear Optics. Redwood City, CA: Addison Wesley Publishing Co.
Rapp, L., 1997. Experimental investigation of signal distortions induced by cross phase modulation combined with dispersion. IEEE Photon Technical Letters 9, 1592–1595.
Shen, Y.R., 1989. Principles of Nonlinear Optics. New York: Wiley.
Shtaif, M., Eiselt, M., Garret, L.D., 2000. Cross-phase modulation distortion measurements in multispan WDM systems. IEEE Photon Technical Letters 12, 88–90.
Tkach, R.W., Chraplyvy, A.R., Forghiari, F., Gnanck, A.H., Derosier, R.M., 1995. Four-photon mixing and high speed WDM systems. Journal of Lightwave Technology 13, 841–849.
Wang, Q.Z., Lim, Q.D., Lim, D.H., Alfano, R., 1994. High resolution spectra of self phase modulation in optical fibers. Journal of Optical Society of America B-11, 1084.
Webjorn, J., Siala, S., Nan, D.W., Waarts, R.G., Lang, R.J., 1997. Visible laser sources based frequency doubling in nonlinear waveguides. IEEE Journal of Quantum Electronics 33, 1673–1686.
Yamada, M., Nada, N., Saitoh, M., Watanabe, K., 1993. Applied Physics Letters 62, 435–436.
Yamada, E., Takara, H., Ohara, T., Sato, K., Morioka, T., Jinguji, K., Itoh, M., Ishi, M., 2001. A high SNR, 150 Ch. Supercontinuum CW optical source with high SNR and precise 25 GHz spacing for 10 Gbit/s DWDM systems. Electronics Letters 37, 304–306.
Yariv, A., 1997. Optical Electronics in Modern Communications, 5th ed. New York: Oxford University Press.

Fabrication of Optical Fiber

D Hewak, University of Southampton, Southampton, UK

Glossary

Dopants Elements or compounds added, usually in small amounts, to a glass composition to modify its properties.
Fiber drawing The process of heating and thus softening an optical fiber preform and then drawing out a thin thread of glass.
Fiber loss The transmission loss of light as it propagates through a fiber, usually measured in dB of loss per unit length of fiber. Loss can occur through the absorption of light in the core or scattering of light out of the core.
Glass An amorphous solid formed by cooling from the liquid state to a rigid solid with no long range structure.
Modified chemical vapor deposition (MCVD) A process for the fabrication of an optical preform where gases flow into the inside of a rotating tube, are heated and react to form particles of glass which are deposited onto the wall of the glass tube. After deposition, the glass particles are consolidated into a solid preform.
Outside vapor deposition (OVD) A process for the fabrication of an optical preform where glass soot particles are formed in an oxy-hydrogen flame and deposited on a rotating rod. After deposition, the glass particles are consolidated into a solid preform.
Preform A fiber blank, a bulk glass rod consisting of a core and cladding glass composite which is drawn into fiber.
Refractive index A characteristic property of glass, which is defined by the speed of light in the material relative to the speed of light in a vacuum.
Silica A transparent glass formed from silicon dioxide.
Vapor-axial deposition A process similar to OVD, where the core and cladding layers are deposited simultaneously.

Introduction

The drawing of optical fibers from silica preforms has, over a short period of time, progressed from the laboratory to become a manufacturing process capable of producing millions of kilometers of telecommunications fiber a year. Modern optical fiber fabrication processes produce low-cost fiber of excellent quality, with transmission losses close to their intrinsic loss limit. Today, fiber with transmission losses of 0.2 dB per kilometer of fiber are routinely drawn through a two-stage process that has been refined since the 1970s.

Although fibers of glass have been fabricated and used for hundreds of years, it was not until 1966 that serious interest in the use of optical fibers for communication emerged. At this time, it was estimated that the optical transmission loss in bulk glass could be as low as 20 dB km^{-1} if impurities were sufficiently reduced, a level at which practical applications were possible. At this time, no adequate fabrication techniques were available to synthesize glass of high purity, and fiber-drawing methods were crude.

Over the next five years, efforts worldwide addressed the fabrication of low-loss fiber. In 1970, a fiber with a loss of 20 dB km^{-1} was achieved. The fiber consisted of a titania doped core and pure silica cladding. This result generated much excitement and a number of laboratories worldwide actively began researching optical fiber. New fabrication techniques were introduced, and by 1986, fiber loss had been reduced close to the theoretical limit.

All telecommunications fiber that is fabricated today is made of silica glass, the most suitable material for low-loss fibers. Early fiber research studied multicomponent glasses, which are perhaps more familiar optical materials; however, low-loss fiber, could not be realized, partly due to the lack of a suitable fabrication method. Today other glasses, in particular the fluorides and sulfides, continue to be developed for speciality fiber applications, but silica fiber dominates in most applications.

Silica is a glass of simple chemical structure containing only two elements, silicon and oxygen. It has a softening temperature of about 2,000°C at which it can be stretched, i.e. drawn into fiber. An optical fiber consists of a high purity silica glass core, doped with suitable oxide materials to raise its refractive index (Fig. 1). This core, typically on the order of 2–10 microns in diameter, is surrounded by silica glass of lower refractive index. This cladding layer extends the diameter to typically 125 microns. Finally a protective coating covers the entire structure. It is the phenomenon of total internal reflection at the core cladding interface that confines light to the core and allows it to be guided. The basic requirements of an optical fiber are as follows:

1. The material used to form the core of the fiber must have a higher refractive index than the cladding material, to ensure the fiber is a guiding structure.
2. The materials used must be low loss, providing transmission with no absorption or scattering of light.
3. The materials used must have suitable thermal and mechanical properties to allow them to be drawn down in diameter into a fiber.

Silica (SiO_2) can be made into a glass relatively easily. It does not easily crystallize, which means that scattering from unwanted crystalline centers within the glass is negligible. This has been a key factor in the achievement of a low-loss fiber. Silica glass has

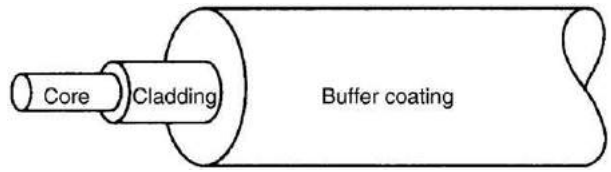


Fig. 1 Structure of an optical fiber.

high transparency in the visible and near-infrared wavelength regions and its refractive index can be easily modified. It is stable and inert, providing excellent chemical and mechanical durability. Moreover, the purification of the raw materials used to synthesize silica glass is quite straightforward.

In the first stage of achieving an optical fiber, silica glass is synthesized by one of three main chemical vapor processes. All use silicon tetrachloride ($SiCl_4$) as the main precursor, with various dopants to modify the properties of the glass. The precursors are reacted with oxygen to form the desired oxides. The end result is a high purity solid glass rod with the internal core and cladding structure of the desired fiber.

In the second stage, the rod, or preform as it is known, is heated to its softening temperature and stretched to diameters of the order of 125 microns. Tens to hundreds of kilometers of fiber are produced from a single preform, which is drawn continuously, with minimal diameter fluctuations. During the drawing process one or more protective coatings are applied, yielding long lengths of strong, low-loss fiber, ready for immediate application.

Preform Fabrication

The key step in preparing a low-loss optical fiber is to develop a technique for completely eliminating transition metal and OH ion contamination during the synthesis of the silica glass. The three methods most commonly used to fabricate a glass optical fiber preform are: the modified chemical vapor deposition process (MCVD); the outside vapor deposition process (OVD); and the vapor-axial deposition process (VAD).

In a typical vapor phase reaction, halide precursors undergo a high temperature oxidation or hydrolysis to form the desired oxides. The completed chemical reaction for the formation of silica glass is, for oxidation:

$$SiCl_4 + O_2 \rightarrow SiO_2 + 2Cl_2 \quad (1)$$

For hydrolysis, which occurs when the deposition occurs in a hydrogen-containing flame, the reaction is:

$$SiCl_4 + 2H_2O \rightarrow SiO_2 + 4HCl \quad (2)$$

These processes produce fine glass particles, spherical in shape with a size of the order of 1 nm. These glass particles, known as soot, are then deposited and subsequently sintered into a bulk transparent glass. The key to low-loss fiber is the difference in vapor pressures of the desired halides and the transition metal halides that cause significant absorption loss at the wavelengths of interest. The carrier gas picks up a pure vapor of, for example, $SiCl_4$, and any impurities are left behind.

Part of the process requires the formation of the desired core/cladding structure in the glass. In all cases, silica-based glass is produced with variations in refractive index produced by the incorporation of dopants. Typical dopants used are germania (GeO_2), titania (TiO_2), alumina (Al_2O_3), and phosphorous pentoxide (P_2O_5) for increasing the refractive index, and boron oxide (B_2O_2) and fluorine (F) for decreasing it. These dopants also allow other properties to be controlled, such as the thermal expansion of the glass and its softening temperatures. In addition, other materials, such as the rare earth elements, have also been used to fabricate active fibers that are used to produce optical fiber amplifiers and lasers.

MCVD Process

Optical fibers were first produced by the MCVD method in 1974, a breakthrough that completely solved the technical problems of low-loss fiber fabrication (**Fig. 2**).

As shown schematically in **Fig. 3**, the halide precursors are carried in the vapor phase by oxygen carrier gas into a pure silica substrate tube. An oxy-hydrogen burner traverses the length of the tube, which it heats externally. The tube is heated to temperatures of about 1,400°C which then oxidizes the halide vapor materials. The deposition temperature is sufficiently high to form a soot made of glassy particles which are deposited on the inside wall of the substrate tube but low enough to prevent the softened silica substrate tube from collapsing. The process usually takes place on a horizontal glass-working lathe.

During the deposition process, the repeated traversing of the burner forms multiple layers of soot. Changes to the precursors entering the tube and thus the resulting glass composition are introduced for layers which will form the cladding and then the core. The MCVD method allows germania to be doped into the silica glass and the precise control of the refractive index profile of the preform. When deposition is complete, the burner temperature is increased and the hollow, multilayered structure is collapsed to a solid rod. A characteristic of fiber formed by this process is a refractive index dip in the center of the core of the fiber. In addition, the deposition takes place in a closed system, which dramatically reduces contamination by OH^- ions and maintains

Fig. 2 Fabrication of an optical fiber preform by the MCVD method.

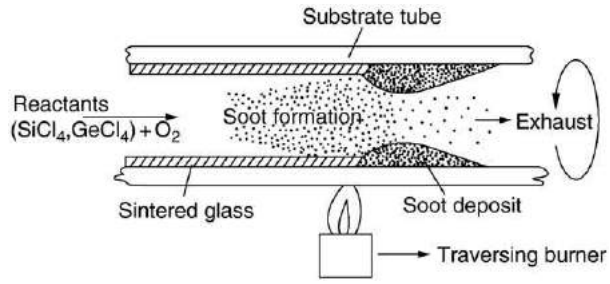


Fig. 3 Schematic of preform fabrication by the MCVD method.

low levels of other impurities. The high temperature allows high deposition rates compared to traditional chemical vapor deposition (CVD) and large performs, built up from hundreds of layers can be produced.

The MCVD method is still widely used today though it has some limitations, particularly on the preform size that can be achieved and thus the manufacturing cost. Diameter of the final preform is determined by the size of the initial silica substrate tube, which, due to the high purity required, accounts for a significant portion of the cost of the preform. A typical preform fabricated by MCVD yields about 5 km of fiber. Its success has spurred improvements to the process, in particular to address the fiber yield from a single preform.

OVD Process

The outside vapor deposition (OVD) methods, also known as outside vapor phase oxidation (OVPO), synthesize the fine glass particles within a burner flame. The precursors, oxygen, and fuel for the burner, are introduced directly into the flame. The soot is deposited onto a target rod that rotates and traverses in front of the burner. As in MCVD, the preform is built up layer by layer, though now initially by depositing the core glass and then building up the cladding layers over this. After the deposition process is complete, the preform is removed from the target rod and is collapsed and sintered into a transparent glass preform. The center hole remains, but disappears during the fiber drawing process. This technique has advantages in both size of preform which can be obtained and the fact that a high-quality silica substrate tube is no longer required. These two advantages combine to make a more economical process. From a single preform, several hundred kilometers of fiber can be produced.

VAD Process

The most recent refinement to the fabrication process was developed again to aid the mass production of high-quality fibers. In the VAD process, both core and cladding glasses are deposited simultaneously. Like OVD, the soot is synthesized and deposited by flame hydrolysis, as shown in Fig. 4, the precursors are blown from a burner, oxidized, and deposited onto a silica target rod.

Burners for the VAD process consist of a series of concentrate nozzles. The first delivers an inert carrier and the main precursors $SiCl_4$, the second delivers an inert carrier glass and the dopants, the third delivers hydrogen fuel, and the fourth delivers oxygen. Gas flows are up to a liter per minute and deposition rates can be very high. With the VAD process, both core and cladding glasses can be deposited simultaneously. The main advantage is that this is a continuous process, as the soot which forms the core and cladding glasses are deposited axially onto the end of the rotating silica rod, it is slowly drawn upwards into a furnace which sinters

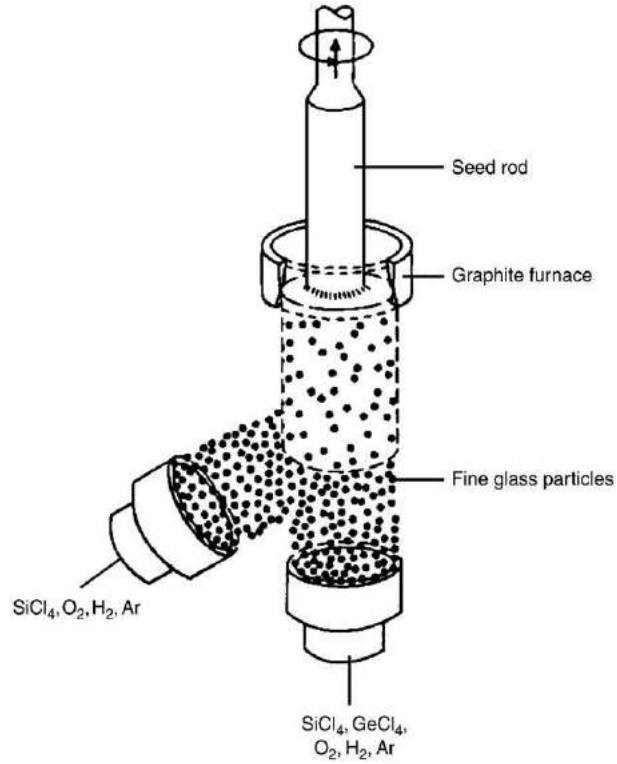


Fig. 4 Schematic of preform fabrication by the VAD method.

and consolidates the soot into a transparent glass. The upward motion is such that the end at which deposition is occurring remains in a fixed position and essentially the preform is grown from this base.

The advantages of the VAD process are a preform without a central dip and, most importantly, the mass production associated with a continuous process. The glass quality produced is uniform and the resulting fibers have excellent reproducibility and low loss.

Other Methods of Preform Fabrication

There are other methods of preform fabrication, though these are not generally used for today's silica glass based fiber. Some methods, such as fabrication of silica through sol–gel chemistry could not provide the large low-loss preforms obtained by chemical vapor deposition techniques. Other methods, such as direct casting of molten glass into rods, or formation of rods and tubes by extrusion, are better suited to and find application in other glass chemistries and speciality fibers.

Fiber Drawing

Once a preform has been made, the second step is to draw the preform down in diameter on a fiber drawing tower. The basic components and configuration of the drawing tower have remained unchanged for many years, although furnace design has become more sophisticated and processes like coating have become automated. In addition, fiber drawing towers have increased in height to allow faster pulling speeds (Fig. 5).

Essentially, the fiber drawing process takes place as follows. The preform is held in a chuck which is mounted on a precision feed assembly that lowers the preform into the drawing furnace at a speed which matches the volume of preform entering the furnace to the volume of fiber leaving the bottom of the furnace. Within the furnace, the fiber is drawn from the molten zone of glass down the tower to a capstan, which controls the fiber diameter, and then onto a drum. During the drawing process, immediately below the furnace, is a noncontact device that measures the diameter of the fiber. This information is fed back to the capstan which speeds up to reduce the diameter or slows down to increase the fiber diameter. In this way, a constant diameter fiber is produced. One or more coatings are also applied in-line to maintain the pristine surface quality as the fiber leaves the furnace and thus to maximize the fiber strength.

The schematic drawing in Fig. 6 shows a fiber drawing tower and its key components. Draw towers are commercially available, ranging in height from 3 m to greater than 20 m, with the tower height increasing as the draw speed increases.

Typical drawing speeds are on the order of 1 m s^{-1}. The increased height is needed to allow the fiber to cool sufficiently before entering the coating applicator, although forced air-cooling of the fiber is commonplace in a manufacturing environment.

Fig. 5 A commercial scale optical fiber drawing tower.

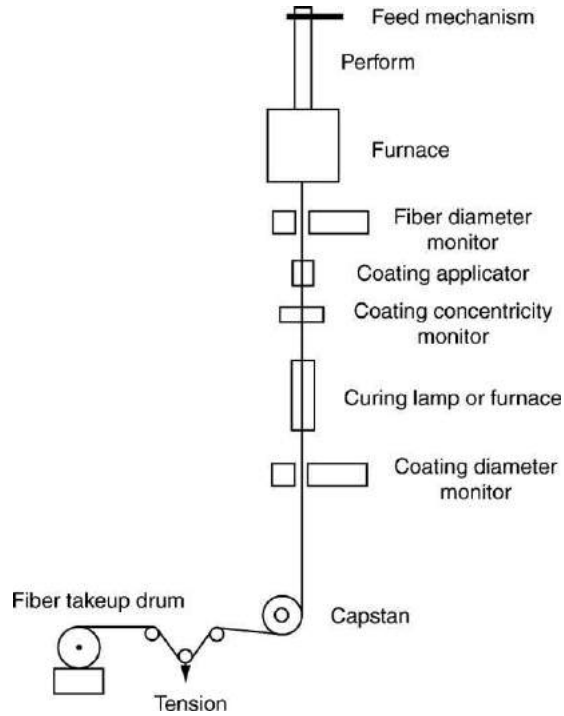


Fig. 6 Schematic diagram of an optical fiber drawing tower and its components.

Furnaces

A resistance furnace is perhaps the most common and economical heating source used in a fiber drawing tower. A cylindrical furnace, with graphite or tungsten elements, provides heating to the preform by blackbody radiation. These elements must be

surrounded by an inert gas, such as nitrogen or argon, to prevent oxidation. Gas flow and control is critical to prevent variations in the diameter of the fiber. In addition, the high temperature of the elements may lead to contamination of the preform surface that can then weaken the fiber.

An induction furnace provides precise clean heating through radio frequency (RF) energy that is inductively coupled to a zirconia susceptor ring. This type of furnace is cleaner than the graphite resistance type and is capable of continuous running for several months. It also has the advantage that it does not require a protective inert atmosphere and consequently the preform is drawn in a turbulence-free environment. These advantages result in high-strength fibers with good diameter control.

Diameter Measurement

A number of noncontact methods of diameter measurement can be applied to measure fiber diameter. These include laser scanning, interferometry, and light scattering techniques. To obtain precise control of the fiber diameter, the deviation between the desired diameter and the measured diameter is fed into the diameter control system. In order to cope with high drawing speeds, sampling rates as high as 1,000 times per second are used. The diameter control is strongly affected by the gas flow in the drawing furnace and is less affected by the furnace temperature variation. The furnace gas flow can be used to achieve suppression of fast diameter fluctuations. This is used in combination with drawing speed control to achieve suppression of both fast and slow diameter fluctuations. Current manufacturing processes are capable of producing several hundreds of kilometers of fiber with diameter variations of ± 1 micron.

There are two major sources of fiber diameter fluctuations: short-term fluctuations caused by temperature fluctuations in the furnace; and long-term fluctuations caused by variations in the outer diameter of the preform. Careful control of the furnace temperature, the length of the hot zone, and the flow of gas minimize the short-term fluctuations. Through optimization of these parameters, diameter errors of less than ± 0.5 microns can be realized. The long-term fluctuations in diameter are controlled by the feedback mechanism between the diameter measurement and the capstan.

Fiber Coating

A fiber coating is primarily used to preserve the strength of a newly drawn fiber and therefore must be applied immediately after the fiber leaves the furnace. The fiber coating apparatus is typically located below the diameter measurement gauge at a distance determined by the speed of fiber drawing, the tower height, and whether there is external cooling of the fiber. Coating is usually

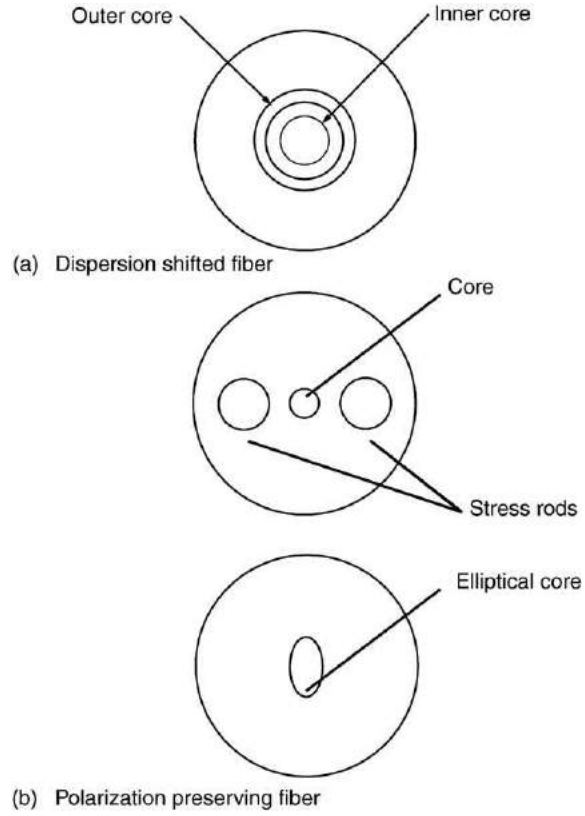


Fig. 7 Structure of (a) dispersion shifted fiber and (b) two methods of achieving polarization preserving fiber.

one of the limiting factors in the speed of fiber drawing. To minimize any damage or contamination of the pristine fiber leaving the furnace, this portion of the tower, from furnace to the application of the first coating, is enclosed in a clean, filtered-air chamber.

A wide variety of coating materials has been applied; however, conventional, commercial fiber generally relies on a UV curable polymer as the primary coating. Coating thickness is typically 50–100 microns. Subsequent coatings can then be applied for specific purposes. A dual coating is often used with an inner primary coating that is soft and an outer, secondary coating which is hard. This ratio of low to high elastic modulus can minimize stress on the fiber and reduce bending loss.

Alternative coatings include hermetic coatings of a low melting temperature metal, ceramics, or amorphous carbon. These can be applied in-line before the polymeric coating. Metallic coatings are applied by passing the fiber through a molten metal while ceramic or amorphous coatings utilize an in-line chemical vapor deposition reactor.

Types of Fiber

The majority of silica fiber drawn today is single mode. This structure consists of a core whose diameter is chosen such that, with a given refractive index difference between the core and cladding, only a single guide mode propagates at the wavelength of interest. With a discrete index difference between core and cladding, it is often referred to as a step index fiber. In typical telecommunications fiber, single mode operation is obtained with core diameters of 2–10 microns with a standard outer diameter of 125 microns.

Multimode fiber has core diameters considerably larger, typically 50, 62.5, 85, and 110 microns, again with a cladded diameter of 125 microns. Multimode fibers are often graded index, that is the refractive index is a maximum in the center of the fiber and smoothly decreases radially until the lower cladding index is reached. Multimode fibers find use in non-telecommunication applications, for example optical fiber sensing and medicine.

Single mode fibers, which are capable of maintaining a linear polarization input to the fiber, are known as polarization preserving fibers. The structure of these fibers provides a birefringence that removes the degeneracy of the two possible polarization modes. This birefringence is a small difference in the effective refractive index of the two polarization modes that can be guided and it is achieved in one of two ways. Common methods for the realization of this birefringence are an elliptical core in the fiber, or through stress rods, which modify the refractive index in one orientation. These fiber structures are shown in **Fig. 7**.

While the loss minimum of silica-based fiber is near 1.55 microns, step index single-mode fiber offers zero dispersion close to 1.3 micron wavelengths and dispersion at the loss minimum is considerable. A modification of the structure of the fiber, and in particular a segmented refractive index profile in the core, can be used to shift this dispersion minimum to 1.55 microns. This fiber, illustrated in **Fig. 7** is known as dispersion shifted fiber. Similarly fibers, with a relatively low dispersion over a wide wavelength range, known as dispersion flattened fibers, can be obtained by the use of multiple cladding layers.

See also: Guided Wave Optics. Optical Fiber Cables

Further Reading

Agrawal, G.P., 1992. Fiber-Optic Communication Systems. New York: John Wiley & Sons.
Bass, M., Van Stryland, E.V. (Eds.), 2001. Fiber Optics Handbook: Fiber, Devices, and Systems for Optical Communications. New York: McGraw-Hill Professional Publishing.
Fujiura, K., Kanamori, T., Sudo, S., 1997. Fiber materials and fabrications. In: Sudo, S. (Ed.), Optical Fiber Amplifiers. Boston, MA: Artech House, pp. 193–404. ch. 4.
Goff, D.R., Hansen, K.S., 2002. Fiber Optic Reference Guide: A Practical Guide to Communications Technology, 3rd edn Oxford, UK: Butterworth-Heinemann UK.
Hewak, D. (Ed.), 1998. Glass and Rare Earth Doped Glasses for Optical Fibres.EMIS Datareview Series, INSPEC. London: The Institution of the Electrical Engineers.
Keck, D.B., 1981. Optical fibre waveguides. In: Barnoski, M.K. (Ed.), Fundamentals of Optical Fiber Communications. New York: Academic Press.
Li, T. (Ed.), 1985. Optical Fiber Communications: Fiber Fabrication. New York: Academic Press.
Personick, S.D., 1985. Fiber Optics: Technology and Applications. New York: Plenum Press.
Schultz, P.C., 1979. In: Bendow, B., Mitra, S.S. (Eds.), Fiber Optics. New York: Plenum Press.

Overview: Coherence

A Sharma and AK Ghatak, Indian Institute of Technology, New Delhi, India
HC Kandpal, National Physical Laboratory, New Delhi, India

Introduction

Coherence refers to the characteristic of a wave that indicates whether different parts of the wave oscillate in tandem, or in a definite phase relationship. In other words, it refers to the degree of confidence by which one can predict the amplitude and phase of a wave at a point, from the knowledge of these at another point at the same or a different instant of time. Emission from a thermal source, such as a light bulb, is a highly disordered process and the emitted light is incoherent. A well-stabilized laser source, on the other hand, generates light in a highly ordered manner and the emitted light is highly coherent. Incoherent and coherent light represent two extreme cases. While describing the phenomena of physical optics and diffraction theory, light is assumed to be perfectly coherent in both spatial as well as temporal senses, whereas in radiometry it is generally assumed to be incoherent. However, practical light sources and the fields generated by them are in between the two extremes and are termed as partially coherent sources and fields. The degree of order that exists in an optical field produced by a source of any kind may be described in terms of various correlation functions. These correlation functions are the basic theoretical tools for the analyses of statistical properties of partially coherent light fields.

Light fields generated from real physical sources fluctuate randomly to some extent. On a microscopic level quantum mechanical fluctuations produce randomness and on macroscopic level the randomness occurs as a consequence of these microscopic fluctuations, even in free space. In real physical sources, spontaneous emission causes random fluctuations and even in the case of lasers, spontaneous emission cannot be suppressed completely. In addition to spontaneous emission, there are many other processes that give rise to random fluctuations of light fields. Optical coherence theory was developed to describe the random nature of light and it deals with the statistical similarity between light fluctuations at two (or more) space–time points.

As mentioned earlier, in developing the theory of interference or diffraction, light is assumed to be perfectly coherent, or, in other words, it is taken to be monochromatic and sinusoidal for all times. This is, however, an idealization since the wave is obviously generated at some point of time by an atomic or molecular transition. Furthermore, a wavetrain generated by such a transition is of a finite duration, which is related to the finite lifetime of the atomic or molecular levels involved in the transition. Thus, any wave emanating from a source is an ensemble of a large number of such wavetrains of finite duration, say τ_c. A simplified visualization of such an ensemble is shown in **Fig. 1** where a wave is shown as a series of wavetrains of duration τ_c. It is evident from the figure that the fields at time t and $t+\Delta t$ will have a definite phase relationship if $\Delta t \ll \tau_c$ and will have no or negligible phase relationship when $\Delta t \gg \tau_c$. The time τ_c is known as the coherence time of the radiation and the field is said to remain coherent for time $\sim \tau_c$. This property of waves is referred to as the time coherence or the temporal coherence and is related to the spectral purity of the radiation. If one obtains the spectrum of the wave shown in **Fig. 1** by taking the Fourier transform of the time variation, it would have a width of $\Delta\nu$ around ν_0 which is the frequency of the sinusoidal variation of individual wavetrains. The spectral width $\Delta\nu$ is related to the coherence time as

$$\Delta\nu \sim 1/\tau_c \tag{1}$$

For thermal sources such as a sodium lamp, $\tau_c \sim 100$ ps, whereas for a laser beam it could be as large as a few milliseconds. A related quantity is the coherence length l_c, which is the distance covered by the wave in time τ_c,

$$l_c = c\tau_c \sim \frac{c}{\Delta\nu} = \frac{\lambda_0^2}{\Delta\lambda} \tag{2}$$

where λ_0 is the central wavelength ($\lambda_0 = c/\nu_0$) and $\Delta\lambda$ is the wavelength spread corresponding to the spectral width $\Delta\nu$. In a two-beam interference experiment (e.g., Michelson interferometer, **Fig. 2**), the interfering beam derived from the same source will produce good interference fringes if the path difference between the two interfering beams is less than the coherence length of the radiation given out by the source.

It must be added here that for the real fields, generated by innumerable atoms or molecules, the individual wavetrains have different lengths around an average value τ_c. Furthermore, several wavetrains in general are propagating simultaneously,

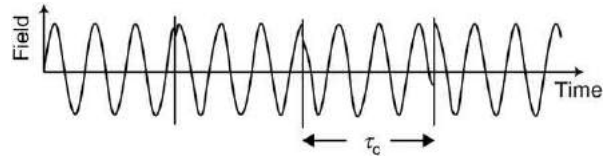


Fig. 1 Typical variation of the radiation field with time. The coherence time $\sim \tau_c$.

 doi:10.1016/B978-0-12-809283-5.00698-4

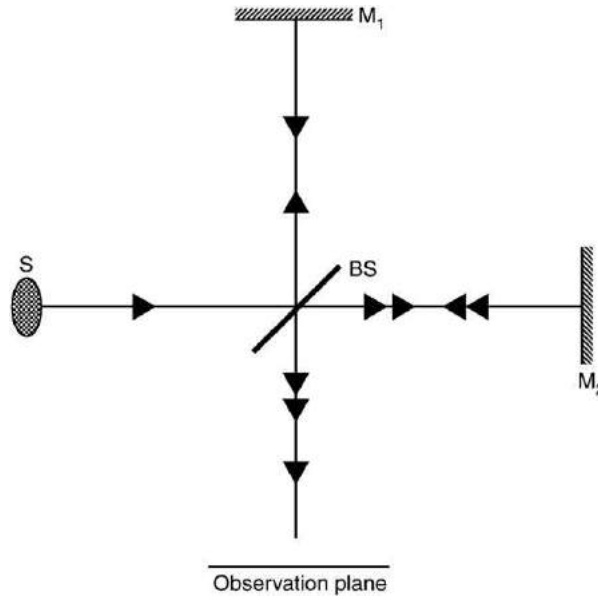


Fig. 2 Michelson interferometer to study the temporal coherence of radiation from source S; M_1 and M_2 are mirrors and BS is a 50%–50% beamsplitter.

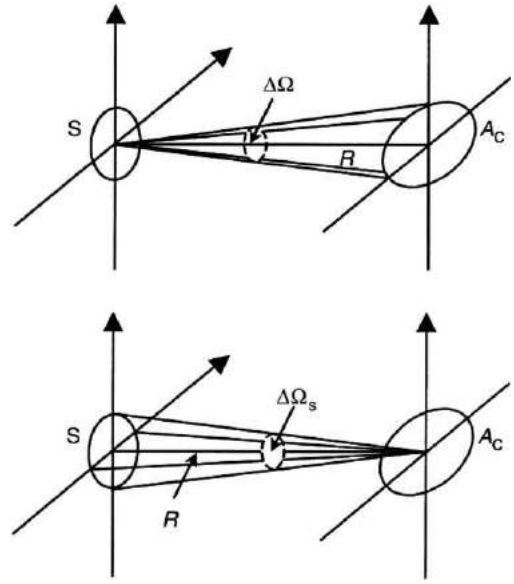


Fig. 3 Spatial coherence and coherence area.

overlapping in space and time, to produce an ensemble whose properties are best understood in a statistical sense as we shall see in later sections.

Another type of coherence associated with the fields is the space coherence or the spatial coherence, which is related to the size of the source of radiation. It is evident that when the source is an ideal point source, the field at any two points (within the coherence length) would have definite phase relationship. However, the field from a thermal source of finite area S can be thought as resultant of the fields from each point on the source. Since each point source would usually be independent of the other, the phase relationship between the fields at any two points would depend on their position and size of the source. It can be shown that two points will have a strong phase relationship if they lie within the solid angle $\Delta\Omega$ from the source (Fig. 3) such that

$$\Delta\Omega \sim \frac{\lambda_0^2}{S} \tag{3}$$

Thus, on a plane R distance away from the source, one can define an area $A_c = R^2\ \Delta\Omega$ over which the field remains spatially coherent. This area A_c is called the coherence area of the radiation and its square root is sometimes referred to as the transverse

coherence length. It is trivial to show that

$$A_c \sim \frac{\lambda_0^2}{\Delta\Omega_S} \tag{4}$$

where $\Delta\Omega_S$ is the solid angle subtended by the source on the plane at which the coherence area is defined. In Young's double-hole experiment, if the two pinholes lie within the coherence area of the radiation from the primary source, good contrast in fringes would be observed.

Combining the concepts of coherence length and the coherence area, one can define a coherence volume as $V_c = A_c l_c$. For the wavefield from a thermal source, this volume represents that portion of the space in which the field is coherent and any interference produced using the radiation from points within this volume will produce fringes of good contrast.

Mathematical Description of Coherence

Analytical Field Representation

Coherence properties associated with fields are best analyzed in terms of complex representation for optical fields. Let the real function $V^{(r)}(\mathbf{r}, t)$ represent the scalar field, which could be one of the transverse Cartesian components of the electric field associated with the electromagnetic wave. One can then define a complex analytical signal $V(\mathbf{r}, t)$ such that

$$V^{(r)}(\mathbf{r}, t) = \mathrm{Re}[V(\mathbf{r}, t)], \tag{5}$$

$$V(\mathbf{r}, t) = \int_0^\infty \tilde{V}^{(\mathbf{r})}(\mathbf{r}, \nu) e^{-2\pi i \nu t} \mathrm{d}\nu$$

where the spectrum $\tilde{V}^{(\mathbf{r})}(\mathbf{r}, \nu))$ is the Fourier transform of the scalar field $V^{(r)}(\mathbf{r}, t)$ and the spectrum for negative frequencies has been suppressed as it does not give any new information since $\tilde{V}^{(r)*}(\mathbf{r}, \nu) = \tilde{V}^{(r)}(\mathbf{r}, -\nu)$.

In general, the radiation from a quasi-monochromatic thermal source fluctuates randomly as it is made of a large number of mutually independent contributions from individual atoms or molecules in the source. The field from such a source can be regarded as an ensemble of a large number of randomly different analytical signals such as $V(\mathbf{r}, t)$. In other words, $V(\mathbf{r}, t)$ is a typical member of an ensemble, which is the result of a random process representing the radiation from a quasi-monochromatic source. This process is assumed to be stationary and ergodic so that the ensemble average is equal to the time average of a typical member of the ensemble and that the origin of time is unimportant. Thus, the quantities of interest in the theory of coherence are defined as time averages:

$$\langle f(t) \rangle = \underset{T \to \infty}{\mathrm{Lim}} \frac{1}{2T} \int_{-T}^{T} f(t) \mathrm{d}t \tag{6}$$

Mutual Coherence

In order to define the mutual coherence, the key concept in the theory of coherence, we consider Young's interference experiment, as shown in Fig. 4, where the radiation from a broad source of size S is illuminating a screen with two pinholes P_1 and P_2. The light emerging from the two pinholes produces interference, which is observed on another screen at a distance R from the first screen. The field at point P, due to the pinholes, would be $K_1 V(\mathbf{r}_1, t - t_1)$ and $K_2\ V(\mathbf{r}_2, t - t_2)$ respectively, where $\mathbf{r}_1$ and $\mathbf{r}_2$ define the positions of P_1 and P_2, t_1 and t_2 are the times taken by the light to travel to P from P_1 and P_2, and K_1 and K_2 are imaginary constants that depend on the geometry and size of the respective pinhole and its distance from the point P. Thus, the resultant field at point P would be given by

$$V(\mathbf{r}, t) = K_1 V(\mathbf{r}_1, t - t_1) + K_2 V(\mathbf{r}_2, t - t_2) \tag{7}$$

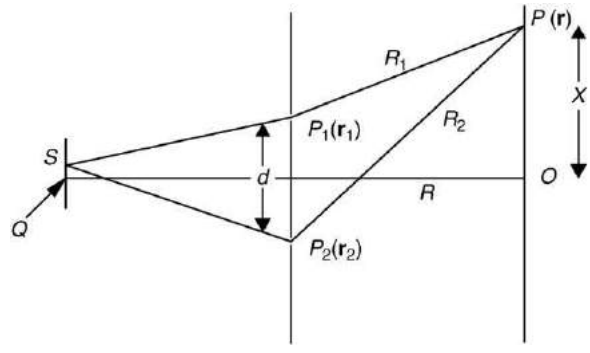


Fig. 4 Schematic of Young's double-hole experiment.

Since the optical periods are extremely small as compared to the response time of a detector, the detector placed at point P will record only the time-averaged intensity:

$$I(P) = \langle V^*(\mathbf{r}, t)V(\mathbf{r}, t)\rangle \tag{8}$$

where some constants have been ignored. With **Eq. (7)**, the intensity at point P would be given by

$$I(P) = I_1 + I_2 + 2\,\mathrm{Re}\{K_1K_2^*\Gamma(\mathbf{r}_1, \mathbf{r}_2, t - t_1, t - t_2)\} \tag{9}$$

where I_1 and I_2 are the intensities at point P, respectively, due to radiations from pinholes P_1 and P_2 independently (defined as $I_i = \langle |K_i\, V(\mathbf{r}_i,\, t)|^2\rangle$) and

$$\begin{aligned}\Gamma(\mathbf{r}_1, \mathbf{r}_2, t - t_{1,} t - t_2) &\equiv \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)\\ &= \langle V^*(\mathbf{r}_1, t)V(\mathbf{r}_2, t + \tau)\rangle\end{aligned} \tag{10}$$

is the mutual coherence function of the fields at P_1 and P_2, and it depends on the time difference $\tau = t_2 - t_1$, since the random process is assumed to be stationary. This function is also sometimes denoted by $\Gamma_{12}(\tau)$. The function $\Gamma_{ii}(\tau) \equiv \Gamma(\mathbf{r}_i, \mathbf{r}_i, \tau)$ defines the self-coherence of the light from the pinhole at P_i, and $|K_i|^2\,\Gamma_{ii}(0)$ defines the intensity I_i at point P, due to the light from pinhole P_i.

Degree of Coherence and the Visibility of Interference Fringes

One can define a normalized form of the mutual coherence function, namely:

$$\begin{aligned}\gamma_{12}(\tau) \equiv \gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) &= \frac{\Gamma_{12}(\tau)}{\sqrt{\Gamma_{12}(0)}\sqrt{\Gamma_{22}(0)}}\\ &= \frac{\langle V^*(\mathbf{r}_1, t)V(\mathbf{r}_2, t + \tau)\rangle}{\left[\langle |V(\mathbf{r}_1, t)|^2\rangle\langle |V(\mathbf{r}_2, t)|^2\rangle\right]^{1/2}}\end{aligned} \tag{11}$$

which is called the complex degree of coherence. It can be shown that $0 \le |\gamma_{12}(\tau)| \le 1$. The intensity at point P, given by **Eq. (9)**, can now be written as

$$I(P) = I_1 + I_2 + 2\sqrt{I_1 I_2}\,\mathrm{Re}[\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)] \tag{12}$$

Expressing $\gamma(\mathbf{r}_1,\, \mathbf{r}_2,\, \tau)$ as

$$\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = |\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)|\exp[\mathrm{i}\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau) - 2\pi\mathrm{i}v_0\tau] \tag{13}$$

where $\alpha(\vec{r}_1, \vec{r}_2, \tau) = \arg[\gamma(\vec{r}_1, \vec{r}_2, \tau)] + 2\pi v_0\tau$ and v_0 is the mean frequency of the light, the intensity in **Eq. (12)** can be written as

$$\begin{aligned}I(P) \quad &= I_1 + I_2 + 2\sqrt{I_1 I_2}|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)|\\ &\times\cos[\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau) - 2\pi v_0\tau]\end{aligned} \tag{14}$$

Now, if we assume that the source is quasi-monochromatic, i.e., its spectral width $\Delta v \ll v_0$, the quantities $\gamma(\mathbf{r}_1,\, \mathbf{r}_2,\, \tau)$ and $\alpha(\mathbf{r}_1,\, \mathbf{r}_2,\, \tau)$ vary slowly on the observation screen, and the interference fringes are mainly obtained due to the cosine term. Thus, defining the visibility of fringes as $\upsilon = (I_{\max} - I_{\min})/(I_{\max} + I_{\min})$, we obtain

$$\upsilon = \frac{2(I_1 I_2)^{1/2}}{I_1 + I_2}|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| \tag{15}$$

which shows that for maximum visibility, the two interfering fields must be completely coherent ($|\gamma(\mathbf{r}_1,\, \mathbf{r}_2,\, \tau)| = 1$). On the other hand, if the fields are completely incoherent ($|\gamma(\mathbf{r}_1,\, \mathbf{r}_2,\, \tau)| = 0$), no fringes are observed ($I_{\max} = I_{\min}$). The fields are said to be partially coherent when $0 < |\,\gamma\,(\mathbf{r}_1,\, \mathbf{r}_2,\, \tau)| < 1$. When $I_1 = I_2$, the visibility is the same as $|\gamma(\mathbf{r}_1,\, \mathbf{r}_2,\, \tau)|$. The relation in **Eq. (15)** shows that in an interference experiment, one can obtain the modulus of the complex degree of coherence by measuring I_1, I_2, and the visibility. Similarly, **Eq. (14)** shows that from the measurement the positions of maxima, one can obtain the phase of the complex degree of coherence, $\alpha(\mathbf{r}_1,\, \mathbf{r}_2,\, \tau)$.

Temporal and Spatial Coherence

If the source illuminating the pinholes is a point source of finite spectral width situated at point Q, the fields at point P_1 and P_2 (**Fig. 4**) at any given instant are the same and the mutual coherence function becomes

$$\begin{aligned}\Gamma_{11}(\tau) &= \Gamma(\mathbf{r}_1, \mathbf{r}_1, \tau) = \langle V^*(\mathbf{r}_1, t)V(\mathbf{r}_1, t + \tau)\rangle\\ &= \langle V^*(\mathbf{r}_2, t)V(\mathbf{r}_2, t + \tau)\rangle = \Gamma_{22}(\tau)\end{aligned} \tag{16}$$

The self-coherence, $\Gamma_{11}(\tau)$, of the light from pinhole P_1, is a direct measure of the temporal coherence of the source. On the other hand, if the source is of finite size and we observe the interference at point O which corresponds to $\tau = 0$, the mutual coherence function would be

$$\Gamma_{12}(0) = \Gamma(\mathbf{r}_1, \mathbf{r}_2, 0) = \langle V^*(\mathbf{r}_1, t)V(\mathbf{r}_2, t)\rangle \equiv J_{12} \tag{17}$$

which is called the mutual intensity and is a direct measure of the spatial coherence of the source. In general, however, the function $\Gamma_{12}(\tau)$ measures, for a source of finite size and spectral width, a combination of the temporal and spatial coherence, and in only some limiting cases, the two types of coherence can be separated.

Spectral Representation of Mutual Coherence

One can also analyze the correlation between two fields in the spectral domain. In particular, one can define the cross-spectral density function $W(\mathbf{r}_1, \mathbf{r}_2, \nu)$ which defines the correlation between the amplitudes of the spectral components of frequency ν of the light at the points P_1 and P_2. Thus:

$$W(\mathbf{r}_1, \mathbf{r}_2, \nu)\delta(\nu - \nu') = \langle V^*(\mathbf{r}_1, \nu)V(\mathbf{r}_2, \nu')\rangle \tag{18}$$

Using the generalized Wiener–Khintchine theorem, the cross-spectral density function can be shown to be the Fourier transform of the mutual coherence function:

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \int_0^\infty W(\mathbf{r}_1, \mathbf{r}_2, \nu)\mathrm{e}^{-2\pi i\nu t}\mathrm{d}\nu \tag{19}$$

$$W(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_{-\infty}^{\infty} \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)\mathrm{e}^{2\pi i\nu t}\mathrm{d}\tau \tag{20}$$

If the two points P_1 and P_2 coincide (i.e., there is only one pinhole), the cross-spectral density function reduces to the spectral density function of the light, which we denote by $S(\mathbf{r}, \nu)$ $[\equiv W(\mathbf{r}, \mathbf{r}, \nu)]$. Thus, it follows from **Eq. (20)** that the spectral density of light is the inverse Fourier transform of the self-coherence function. This leads to the Fourier transform spectroscopy, as we shall see later. One can also define spectral degree of coherence at frequency ν as

$$\begin{aligned}\mu(\mathbf{r}_1, \mathbf{r}_2, \nu) &= \frac{W(\mathbf{r}_1, \mathbf{r}_2, \nu)}{\sqrt{W(\mathbf{r}_1, \mathbf{r}_1, \nu)W(\mathbf{r}_2, \mathbf{r}_2, \nu)}} \\ &= \frac{W(\mathbf{r}_1, \mathbf{r}_2, \nu)}{\sqrt{S(\mathbf{r}_1, \nu)S(\mathbf{r}_2, \nu)}}\end{aligned} \tag{21}$$

It is easy to see that $0 \le |\mu(\mathbf{r}_1, \mathbf{r}_2, \nu)| \le 1$. It is also sometimes referred to as the complex degree of coherence at frequency ν. It may be noted here that in the literature the notation $G^{(1,1)}(\nu) \equiv G(\mathbf{r}_1, \mathbf{r}_2, \gamma)$ has also been used for $W(\mathbf{r}_1, \mathbf{r}_2, \nu)$.

Propagation of Coherence

The van Cittert–Zernike Theorem

Perfectly coherent waves propagate through diffraction formulae, which have been discussed in this volume elsewhere. However, incoherent and partially coherent waves would evidently propagate somewhat differently. The earliest treatment of propagation noncoherent light is due to van Cittert, which was later generalized by Zernike to obtain what is now the van Cittert–Zernike theorem. The theorem deals with the correlations developed between fields at two points, which have been generated by a quasi-monochromatic and spatially incoherent planar source. Thus, as shown in **Fig. 5**, we consider a quasi-monochromatic ($\Delta\nu \ll \nu_0$) planar source σ, which has an intensity distribution $I(\mathbf{r}')$ on its plane and is spatially incoherent, i.e., there is no correlation between the fields at any two points on the source. The field, due to this source, would develop finite correlations after propagation, and the theorem states that

$$\begin{aligned}\gamma(\mathbf{r}_1, \mathbf{r}_2, 0) &= \frac{\Gamma_{12}(0)}{\sqrt{\Gamma_{11}(0)}\sqrt{\Gamma_{12}(0)}} \\ &= \frac{1}{\sqrt{I(\mathbf{r}_1)I(\mathbf{r}_2)}}\int\int_\sigma I(\mathbf{r}')\frac{\mathrm{e}^{2\pi i\nu_0(R_2 - R_1)/c}}{R_1R_2}\mathrm{d}^2r'\end{aligned} \tag{22}$$

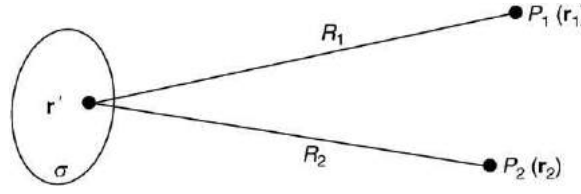


Fig. 5 Geometry for the van Cittert–Zernike theorem.

where $R_i = |\mathbf{r}_i - \mathbf{r}'|$ and $I(\mathbf{r}_i)$ is the intensity at point P_i. This relation is similar to the diffraction pattern produced at point P_1 due to a wave, with a spherical wavefront converging towards P_2 and with an amplitude distribution $I(\mathbf{r}')$ when it is diffracted by an aperture of the same shape, size, and the intensity distribution as those of the incoherent source σ. Thus, the theorem shows that a radiation, which was incoherent at the source, becomes partially coherent as it propagates.

Generalized Propagation

The expression $e^{-2\pi i \nu_0 R_1/c}/R_1$ can be interpreted as the field obtained at P_1 due to a point source located at the point $\mathbf{r}'$ on the planar source. Thus, this expression is simply the point spread function of the homogeneous space between the source and the observation plane containing the points P_1 and P_2. Hopkins generalized this to include any linear optical system characterized by a point spread function $h(\mathbf{r}_j - \mathbf{r}')$ and obtained the formula for the complex degree of coherence of a radiation emerging from an incoherent, quasi-monochromatic planar source after it has propagated through such a linear optical system:

$$\gamma(\mathbf{r}_1, \mathbf{r}_2, 0) = \frac{1}{\sqrt{I(\mathbf{r}_1) I(\mathbf{r}_2)}} \int \int_{\sigma} I(\mathbf{r}') h(\mathbf{r}_1 - \mathbf{r}') h^*(\mathbf{r}_2 - \mathbf{r}') \mathrm{d}^2 r' \tag{23}$$

It would thus seem that the correlations propagate in much the same way, as does the field. Indeed, Wolf showed that the mutual correlation function $\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ satisfies the wave equations:

$$\nabla_1^2 \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \frac{1}{c^2} \frac{\partial^2}{\partial \tau^2} \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) \tag{24}$$

$$\nabla_2^2 \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \frac{1}{c^2} \frac{\partial^2}{\partial \tau^2} \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$$

where ∇_j^2 is the Laplacian with respect to the point $\mathbf{r}_i$. Here the first of **Eqs. (24)**, for instance, gives the variation of the mutual coherence function with respect to $\mathbf{r}_1$ and τ for fixed $\mathbf{r}_2$. Further, the variable τ is the time difference defined through the path difference, since $\tau = (R_2 - R_1)/c$, and the actual time does not affect the mutual coherence function (as the fields are assumed to be stationary).

Using the relation in **Eq. (20)**, one can also obtain from **Eq. (24)**, the propagation equations for the cross-spectral density function $W(\mathbf{r}_1, \mathbf{r}_2, \nu)$:

$$\begin{aligned} \nabla_1^2 W(\mathbf{r}_1, \mathbf{r}_2, \nu) + k^2 W(\mathbf{r}_1, \mathbf{r}_2, \nu) = 0 \\ \nabla_2^2 W(\mathbf{r}_1, \mathbf{r}_2, \nu) + k^2 W(\mathbf{r}_1, \mathbf{r}_2, \nu) = 0 \end{aligned} \tag{25}$$

where $k = 2\pi\nu/c$.

Thompson and Wolf Experiment

One of the most elegant experiments for studying various aspects of coherence theory was carried out by Thompson and Wolf by making slight modifications in the Young's double-hole experiment. The experimental set-up shown in **Fig. 6** consists of a quasi-monochromatic broad incoherent source S of diameter $2a$. This was obtained by focusing filtered narrow band light from a mercury lamp (not shown in the figure) on to a hole of size $2a$ in an opaque screen A. A mask consisting of two pinholes, each of diameter $2b$ with their axes separated by a distance d, was placed symmetrically about the optical axis of the experimental setup at plane B between two lenses L_1 and L_2, each having focal length f. The source was at the front focal plane of the lens L_1 and the observations were made at the plane C at the back focus of the lens L_2. The separation d was varied to study the spatial coherence function on the mask plane by measuring the visibility and the phase of the fringes formed in plane C.

Using the van Cittert–Zernike theorem, the complex degree of coherence at the pinholes P_1 and P_2 on the mask after the lens L_1 is

$$\gamma_{12} = |\gamma_{12}| e^{i\beta_{12}} = \frac{2J_1(v)}{v} \text{ with } v = \left(\frac{2\pi v a d}{c f}\right) \tag{26}$$

for symmetrically placed pinholes about the optical axis. Here β_{12} is the phase of the complex degree of coherence and in this special case where the two holes are equidistant from the axis, β_{12} is either zero or π, respectively, for positive or negative values of $2J_1(v)/v$. Let us assume that the intensities at two pinholes P_1 and P_2 are equal. The interference pattern observed at the back focal

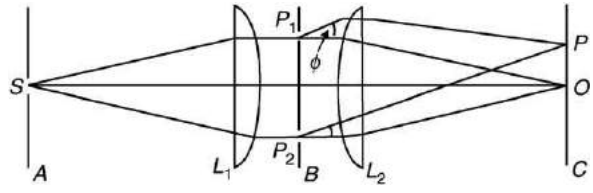


Fig. 6 Schematic of the Thompson and Wolf experiment.

plane of lens L_2 is due to the superposition of the light diffracted from the pinholes. The beams are partially coherent with degree of coherence given by Eq. (26). Since the pinholes are symmetrically placed, the intensity due to either of the pinholes at a point P at the back focal plane of the lens L_2 is the same and is given by the Fraunhofer formula for diffraction from circular apertures, i.e.:

$$I_1(P) = I_2(P) = \left|\frac{2J_1(u)}{u}\right|^2, \text{ with } u = \frac{2\pi v}{c} b \sin\varphi \tag{27}$$

and ϕ is the angle that the diffracted beam makes from normal to the plane of the pinholes. The intensity of the interference pattern produced at the back focal plane of the lens L_2 is:

$$I(P) = 2I_1(P)\left[1 + \left|\frac{2J_1(v)}{v}\right|^2 \cos(\delta + \beta_{12})\right] \tag{28}$$

where $\delta = \mathrm{d}\sin\phi$ is the phase difference between the two beams reaching P from P_1 and P_2. For the on-axis point O, the quantity δ is zero.

The maximum and minimum values of $I(P)$ are given by

$$I_{\max}(P) = 2I_1(P)\lfloor 1 + |2J_1(v)/v|^2 \rfloor \tag{29(a)}$$

$$I_{\max}(P) = 2I_1(P)\lfloor 1 - |2J_1(v)/v|^2 \rfloor \tag{29(b)}$$

Fig. 7 shows an example of the observed fringe patterns obtained by Thompson (1958) in a subsequent experiment.

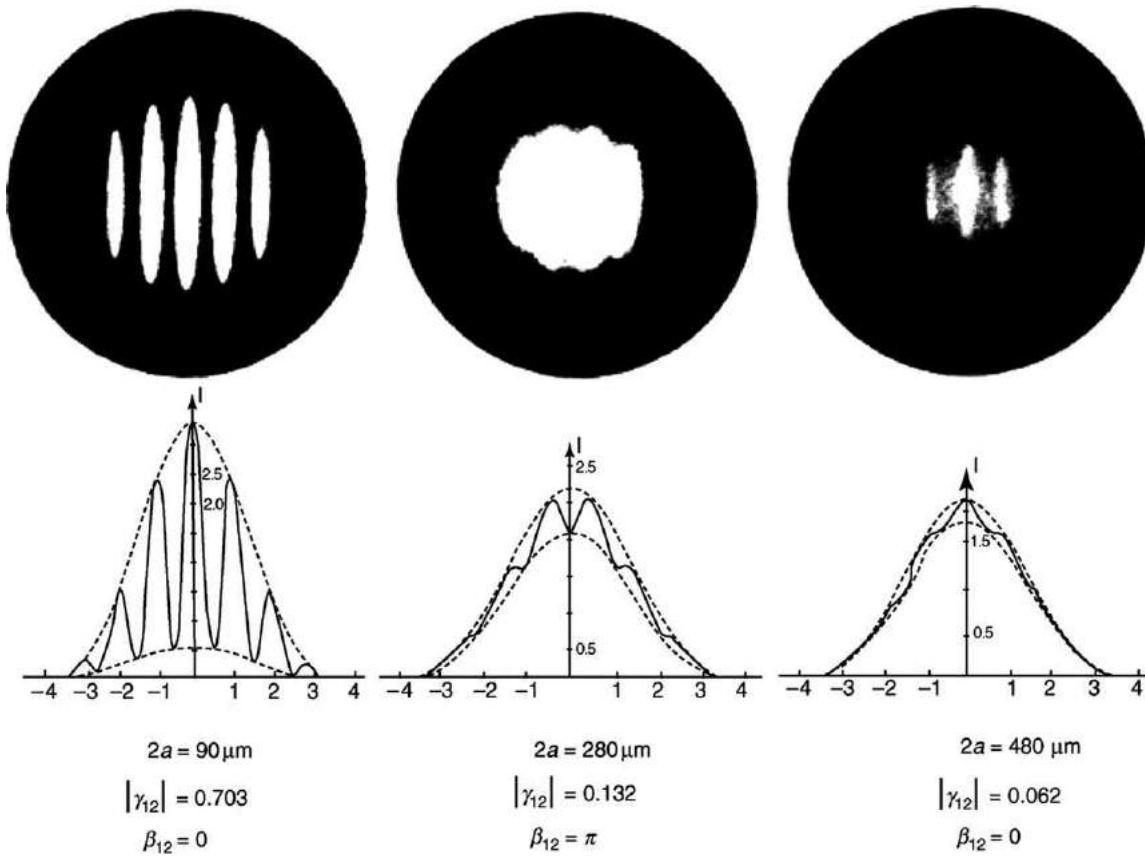


Fig. 7 Two beam interference patterns obtained by using partially coherent light. The wavelength of light used was 0.579 μm and the separation d of the pinholes in the screen at B was 0.5 cm. The figure shows the observed fringe pattern and the calculated intensity variation for three sizes of the secondary source. The dashed lines show the maximum and minimum intensity. The values of the diameter, $2a$, of the source and the corresponding values of the magnitude, $|\gamma_{12}|$, and the phase, β_{12}, are also shown in each case. Reproduced with permission from Thompson BJ (1958) Illustration of phase change in two-beam interference with partially coherent light. *Journal of the Optical Society of America* 48: 95–97.

Types of Fields

As mentioned above, $\gamma_{12}(\tau)$ is a measure of the correlation of the complex field at any two points P_1 and P_2 at specific time delay τ. Extending this definition, an optical field may be coherent or incoherent, if $|\gamma_{12}(\tau)|=1$ or $|\gamma_{12}(\tau)|=0$, respectively, for all pairs of points in the field and for all time delays. In the following, we consider some specific cases of fields and their properties.

Perfectly Coherent Fields

A field would be termed as perfectly coherent or self-coherent at a fixed point, if it has the property that $|\gamma(\mathbf{R}, \mathbf{R}, \tau)|=1$ at some specific point $\mathbf{R}$ in the domain of the field for all values of time delay τ. It can also be shown that $|\gamma(\mathbf{R}, \mathbf{R}, \tau)$ is periodic in time, i.e.:

$$\gamma(\mathbf{R}, \mathbf{R}, \tau) = e^{-2\pi i v_0 \tau} \quad \text{for} \quad v_0 > 0 \tag{30}$$

For any other point $\mathbf{r}$ in the domain of the field, $\gamma(\mathbf{R}, \mathbf{r}, \tau)$ and $\gamma(\mathbf{r}, \mathbf{R}, \tau)$ are also unimodular and also periodic in time.

A perfectly coherent optical field at two fixed points has the property that $|\gamma(\mathbf{R}_1, \mathbf{R}_2, \tau)|=1$ for any two fixed points $\mathbf{R}_1$ and $\mathbf{R}_2$ in the domain of the field for all values of time delay τ. For such a field, it can be shown that $\gamma(\mathbf{R}_1, \mathbf{R}_2, \tau)=\exp[i(\beta - 2\pi v_0 \tau)]$ with v_0 (> 0) and β are real constants. Further, as a corollary, $|\gamma(\mathbf{R}_1, \mathbf{R}_1, \tau)|=1$ and $|\gamma(\mathbf{R}_2, \mathbf{R}_2, \tau)|=1$ for all τ, i.e., the field is self-coherent at each of the field points $\mathbf{R}_1$ and $\mathbf{R}_2$, and

$$\gamma(\mathbf{R}_1, \mathbf{R}_1, \tau) = \gamma(\mathbf{R}_2, \mathbf{R}_2, \tau) = \exp(-2\pi i v_0 \tau) \text{ for } v_0 > 0$$

It can be shown that for any other point $\mathbf{r}'$ within the field

$$\gamma(\mathbf{r}, \mathbf{R}_2, \tau) = \gamma(\mathbf{R}_1, \mathbf{R}_2, 0)\gamma(\mathbf{r}, \mathbf{R}_1, \tau) = \gamma(\mathbf{r}, \mathbf{R}_1, \tau)e^{i\beta} \tag{31}$$

$$\gamma(\mathbf{R}_1, \mathbf{R}_2, \tau) = \gamma(\mathbf{R}_1, \mathbf{R}_2, 0)\exp[-2\pi i v_0 \tau]$$

A perfectly coherent optical field for all points in the domain of the field has the property that $|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)|=1$ for all pairs of points $\mathbf{r}_1$ and $\mathbf{r}_2$ in the domain of the field, for all values of time delay τ. It can then be shown that $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)=\exp\{i[\alpha(\mathbf{r}_1) - \alpha(\mathbf{r}_2)] - 2\pi i v_0 \tau\}$, where $\alpha(\mathbf{r}')$ is real function of a single position $\mathbf{r}'$ in the domain of the field. Further, the mutual coherence function $\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)$ for such a field has a factorized periodic form as

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = U(\mathbf{r}_1)U^*(\mathbf{r}_2)\exp(-2\pi i v_0 \tau) \tag{32}$$

which means the field is monochromatic with frequency v_0 and $U(\mathbf{r})$ is any solution of the Helmholtz equation:

$$\nabla^2 U(\mathbf{r}) = k_0^2 U(\mathbf{r}) = 0, \quad k_0 = 2\pi v_0 / c \tag{33}$$

The spectral and cross-spectral densities of such fields are represented by Dirac δ-functions with their singularities at some positive frequency v_0. Such fields, however, can never occur in nature.

Quasi-Monochromatic Fields

Optical fields are found in practice for which spectral spread Δv is much smaller than the mean frequency $\bar{v}$. These are termed as quasi-monochromatic fields. The cross-spectral density $W(\mathbf{r}_1, \mathbf{r}_2, v)$ of the quasi-monochromatic fields attains an appreciable value only in the small region Δv, and falls to zero outside this region.

$$W(\mathbf{r}_1, \mathbf{r}_2, v) = 0, \quad |v - \bar{v}| > \Delta v \quad \text{and} \quad \Delta v \ll \bar{v} \tag{34}$$

The mutual coherence function $\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ can be written as the Fourier transform of cross-spectral density function as

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = e^{-2\pi i \bar{v} \tau} \int_0^\infty W(\mathbf{r}_1, \mathbf{r}_2, v) e^{-2\pi i (v - \bar{v})\tau} dv \tag{35}$$

If we limit our attention to small τ such that $\Delta v|\tau| \ll 1$ holds, the exponential factor inside the integral is approximately equal to unity and Eq. (35) reduces to

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \exp(-2\pi i \bar{v} \tau) \int_0^\infty [W(\mathbf{r}_1, \mathbf{r}_2, v) dv \tag{36}$$

which gives

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \Gamma(\mathbf{r}_1, \mathbf{r}_2, 0)\exp(-2\pi i \bar{v} \tau) \tag{37}$$

Eq. (37) describes the behavior of $\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ for a limited range of τ values for quasi-monochromatic fields and in this range it behaves as a monochromatic field of frequency $\bar{v}$. However, due to the factor $\Gamma(\mathbf{r}_1, \mathbf{r}_2, 0)$ the quasi-monochromatic field may be coherent, partially coherent, or even incoherent.

Cross-Spectrally Pure Fields

The complex degree of coherence, if it can be factored into a product of a component dependent on spatial coordinates and a component dependent on time delay, is called reducible. In the case of perfectly coherent light, the complex degree of coherence is reducible, as we have seen above, e.g., in Eq. (32), and in the case of quasi-monochromatic fields, this is reducible approximately as

shown in Eq. (37). There also exists a very special case of a cross-spectrally pure field for which the complex degree of coherence is reducible. A field is called a cross-spectrally pure field if the normalized spectrum of the superimposed light is equal to the normalized spectrum of the component beams, a concept introduced by Mandel. In the space-frequency domain, the intensity interference law is expressed as the so-called spectral interference law:

$$S(\mathbf{r},\nu) = S^{(1)}(\mathbf{r},\nu) + S^{(2)}(\mathbf{r},\nu) + 2[\sqrt{S^{(1)}(\mathbf{r},\nu)}\sqrt{S^{(2)}(\mathbf{r},\nu)}]\mathrm{Re}[\mu(\mathbf{r}_1,\mathbf{r}_2,\nu)\mathrm{e}^{-2\pi i\nu\tau}] \tag{38}$$

where $\mu(\mathbf{r}_1, \mathbf{r}_2, \nu)$ is the spectral degree of coherence, defined in Eq. (21) and τ is the relative time delay that is needed by the light from the pinholes to reach any point on the screen; $S^{(1)}(\mathbf{r}, \nu)$ and $S^{(2)}(\mathbf{r}, \nu)$ are the spectral densities of the light reaching P from the pinholes P_1 and P_2, respectively (see Fig. 4) and it is assumed that the spectral densities of the field at pinholes P_1 and P_2 are the same $[S(\mathbf{r}_1, \nu) = CS(\mathbf{r}_2, \nu)]$. Now, if we consider a point for which the time delay is τ_0, then it can be seen that the last term in Eq. (38) would be independent of frequency, provided that

$$\mu(\mathbf{r}_1,\mathbf{r}_2,\nu)\exp(-2\pi i\nu\tau_0) = f(\mathbf{r}_1,\mathbf{r}_2,\tau_0) \tag{39}$$

where $f(\mathbf{r}_1, \mathbf{r}_2, \tau_0)$ is a function of $\mathbf{r}_1$, $\mathbf{r}_2$ and τ_0 only and the light at this point would have the same spectral density as that at the pinholes. If a region exists around the specified point on the observation plane, such that the spectral distribution of the light in this region is of the same form as the spectral distribution of the light at the pinholes, the light at the pinholes is cross-spectrally pure light.

In terms of the spectral distribution of the light $S(\mathbf{r}_1, \nu)$ at pinhole P_1 and $S(\mathbf{r}_2, \nu)[= CS(\mathbf{r}_1, \nu)]$ at pinhole P_2, the mutual coherence function at the pinholes can be written as

$$\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau) = \sqrt{C}\int S(\mathbf{r}_1,\nu)\mu(\mathbf{r}_1,\mathbf{r}_2,\nu)\exp(-2\pi i\nu\tau)\mathrm{d}\nu \tag{40}$$

and using Eq. (39), we get the very important condition for the field to be cross-spectrally pure, i.e.:

$$\gamma(\mathbf{r}_1,\mathbf{r}_2,\tau) = \gamma(\mathbf{r}_1,\mathbf{r}_2,\tau_0)\gamma(\mathbf{r}_1,\mathbf{r}_1,\tau-\tau_0) \tag{41}$$

The complex degree of coherence $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ is thus expressible as the product of two factors: one factor characterizes the spatial coherence at the two pinholes at time delay τ_0 and the other characterizes the temporal coherence at one of the pinholes. Eq. (41) is known as the reduction formula for cross-spectrally pure light. It can further be shown that

$$\mu(\mathbf{r}_1,\mathbf{r}_2,\nu) = \gamma(\mathbf{r}_1,\mathbf{r}_2,\tau_0)\exp(2\pi i\nu\tau_0) \tag{42}$$

Thus, the absolute value of the spectral degree of coherence is the same for all frequencies and is equal to the absolute value of the degree of coherence for the point specified by τ_0 It has been shown that cross-spectrally pure light can be generated, for example, by linear filtering of light that emerges from the two pinholes in Young's interference experiments.

Types of Sources

Primary Sources

A primary source is a set of radiating atoms or molecules. In a primary source the randomness comes from true source fluctuations, i.e., from the spatial distributions of fluctuating charges and currents. Such a source gives rise to a fluctuating field. Let $Q(\mathrm{r},t)$ represent the fluctuating source variable at any point r at time t, then the field generated by the source is represented by fluctuating field variable $V(\mathbf{r},t$. The source is assumed to be localized in some finite domain such that $Q(\mathbf{r},t) = 0$ at any time $t > 0$ outside the domain. Assuming that field variable $V(\mathbf{r},t)$ and the source variable $Q(\mathbf{r},t)$ are scalar quantities, they are related by an inhomogeneous equation as

$$\nabla^2 V(\mathbf{r},t) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}V(\mathbf{r},t) = -4\pi Q(\mathbf{r},\mathrm{t}) \tag{43}$$

The mutual coherence functions of the source $\Gamma_Q(\mathbf{r}_1,\mathbf{r}_2,\tau) = \langle Q^*(\mathbf{r}_1,t)Q(\mathbf{r}_2,t+\tau)\rangle$ and of the field $\Gamma_V(\mathbf{r}_1,\mathbf{r}_2,\tau) = \langle V^*(\mathbf{r}_1,t)V(\mathbf{r}_2,t+\tau)\rangle$ characterize the statistical similarity of the fluctuating quantities at the points $\mathbf{r}_1$ and $\mathbf{r}_2$. Following Eq. (20), one can define, respectively, the cross-spectral density function of the source and the field as

$$W_Q(\mathbf{r}_1,\mathbf{r}_2,\nu) = \int_{-\infty}^{\infty}\Gamma_Q(\mathbf{r}_1,\mathbf{r}_2,\tau)\mathrm{e}^{-2\pi i\nu t}\mathrm{d}\tau \tag{44}$$

$$W_V(\mathbf{r}_1,\mathbf{r}_2,\nu) = \int_{-\infty}^{\infty}\Gamma_V(\mathbf{r}_1,\mathbf{r}_2,\tau)\mathrm{e}^{-2\pi i\nu t}\mathrm{d}\tau$$

The cross-spectral density functions of the source and of the field are related as

$$\left(\nabla_2^2 + k^2\right)\left(\nabla_1^2 + k^2\right)W_V(\mathbf{r}_1,\mathbf{r}_2,\nu) = 4\pi^2 W_Q(\mathbf{r}_1,\mathbf{r}_2,\nu) \tag{45}$$

The solution of Eq. (45) is represented as

$$W_V(\mathbf{r}_1,\mathbf{r}_2,\nu)=\int_S\int_S W_Q(\mathbf{r}_1',\mathbf{r}_2'\nu)\frac{e^{ik(R_2-R_1)}}{R_1R_2}d^3r_1'd^3r_2' \tag{46}$$

where $R_1=|\mathbf{r}_1-\mathbf{r}'_1|$ and $R_2=|\mathbf{r}_2-\mathbf{r}'_2|$ (see Fig. 8). Using Eq. (46), one can then obtain an expression for the spectrum at a point ($\mathbf{r}_1=\mathbf{r}_2=\mathbf{r}=r\mathbf{u}$) in the far field ($r\gg r'_1,r'_2$) of a source as

$$S^\infty(r\mathbf{u},\nu)=\frac{1}{r^2}\int_S\int_S W_Q(\mathbf{r}_1',\mathbf{r}_2',\nu)e^{-ik\mathbf{u}\cdot(\mathbf{r}_1'-\mathbf{r}_2')}d^3r_1'd^3r_2' \tag{47}$$

where u is the unit vector along r. The integral in Eq. (47), i.e., the quantity defined by $r^2S^\infty(r\mathbf{u},\nu)$, is also defined as the radiant intensity, which represents the rate of energy radiated at the frequency ν from the source per unit solid angle in the direction of u.

Secondary Sources

Sources used in a laboratory are usually secondary planar sources. A secondary source is a field, which arises from the primary source in the region outside the domain of the primary sources. This kind of source is an aperture on an opaque screen illuminated either directly or via an optical system by primary sources. Let $V(\boldsymbol{\rho},t$ represent the fluctuating field in a secondary source plane σ at $z=0$ (Fig. 9) and $W_0(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2,\nu)$ represent its cross-spectral density (the subscript 0 refers to $z=0$). One can then solve Eq. (25) to obtain the propagation of the cross-spectral density from this planar source. For two points P_1 and P_2 located at distances which are large compared to wavelength, the cross-spectral density is then given by

$$W(\mathbf{r}_1,\mathbf{r}_2,\nu)=\left(\frac{k}{2\pi}\right)^2\int_\sigma\int_\sigma W_0(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2,\nu)\frac{e^{ik(R_2-R_1)}}{R_1R_2}\cos\theta_1'\cos\theta_2'd^2\rho_1d^2\rho_2 \tag{48}$$

where $R_j=|\mathbf{r}_j-\boldsymbol{\rho}_j|$, $(j=1,2)$, θ'_1 and θ'_2 are the angles that R_1 and R_2 directions make with the z-axis (Fig. 9). Using Eq. (48), one can then obtain an expression for the spectrum at a point ($\mathbf{r}_1=\mathbf{r}_2=\mathbf{r}=r\mathbf{u}$) in the far field ($r\gg\rho_1,\rho_2$ of a planar source as

$$S^\infty(r\mathbf{u},\nu)=\frac{k^2}{2\pi^2}\frac{\cos^2\theta}{r^2}\int_\sigma\int_\sigma W_0(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2,\nu)e^{-ik\mathbf{u}_\perp\cdot(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2)}d^2\rho_1d^2\rho_2 \tag{49}$$

where $\mathbf{u}_\perp$ is the projection of the unit vector $\mathbf{u}$ on the plane σ of the source.

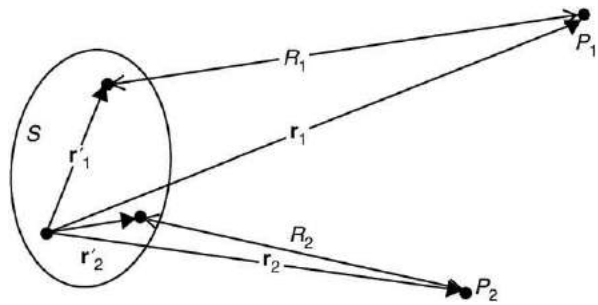


Fig. 8 Geometry of a 3-dimensional primary source S and the radiation from it.

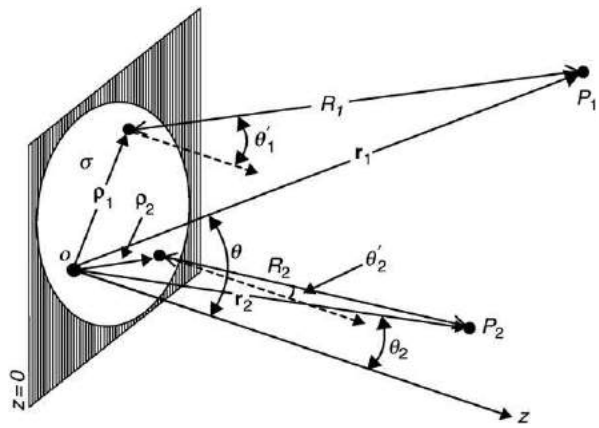


Fig. 9 Geometry of a planar source σ and the radiation from it.

Schell-Model Sources

In the framework of coherence theory in space–time domain, two-dimensional planar model sources of this kind were first discussed by Schell. Later, the model was adopted for formulation of coherence theory in space–frequency domain. Schell-model sources are the sources whose degree of spectral coherence $\mu_A(\mathbf{r}_1,\mathbf{r}_2,\nu)$ (for either primary or secondary source) is stationary in space. It means that $\mu_A(\mathbf{r}_1,\mathbf{r}_2,\nu)$ depends on $\mathbf{r}_1$ and $\mathbf{r}_2$ only through the difference $\mathbf{r}_2-\mathbf{r}_1$, i.e., of the form:

$$\mu_A(\mathbf{r}_1,\mathbf{r}_2,\nu)\equiv\mu_A(\mathbf{r}_2-\mathbf{r}_1,\nu) \tag{50}$$

for each frequency ν present in the source spectrum. Here A stands for field variables V, in the case of a Schell-model secondary source and for source variables Q, in the case of a Schell-model primary source. The cross-spectral density function of a Schell-model source is of the form:

$$W_A(\mathbf{r}_1,\mathbf{r}_2,\nu)=[S_A(\mathbf{r}_1,\nu)]^{1/2}[S_A(\mathbf{r}_2,\nu)]^{1/2}\mu_A(\mathbf{r}_2-\mathbf{r}_1,\nu) \tag{51}$$

where $S_A(\mathbf{r},\nu)$ is the spectral density of the light at a typical point in the primary source or on the plane of a secondary source. Schell model sources do not assume low coherence, and therefore, can be applied to spatially stationary light fields of any state of coherence. The Schell-model of the form given in **Eq. (51)** has been used to represent both three-dimensional primary sources and two-dimensional secondary sources.

Quasi-Homogeneous Sources

Useful models of partially coherent sources that are frequently encountered in nature or in the laboratory are the so-called quasi-homogeneous sources. These are an important sub-class of Schell-model sources. A Schell-model source is called quasi-homogeneous if the intensity of a Schell model source is essentially constant over any coherence area. Under these approximations, the cross-spectral density function for a quasi-homogeneous source is given by

$$\begin{aligned} W_A(\mathbf{r}_1,\mathbf{r}_2,\nu) &= S_A\left[\tfrac{1}{2}(\mathbf{r}_1+\mathbf{r}_2),\nu\right]\mu_A(\mathbf{r}_2-\mathbf{r}_1,\nu)\\ &= S_A[\mathbf{r},\nu]\mu_A(\mathbf{r}',\nu)\end{aligned} \tag{52}$$

where $\mathbf{r}=(\mathbf{r}_1+\mathbf{r}_2)/2$, and $\mathbf{r}'=\mathbf{r}_2-\mathbf{r}_1$. The subscript A stands for either V or Q for the field variable or a source variable, respectively. It is clear that for a quasi-homogeneous source the spectral density $S_A(\mathbf{r},\nu$ varies so slowly with position that it is approximately constant over distances across the sources that are of the order of the correlation length Λ, which is a measure of the effective width of $|\mu_A(\mathbf{r}',\nu)|$. Therefore, $S_A(\mathbf{r},\nu)$ is a slowly varying function of r (**Fig. 10(b)**) and $|\mu_A(\mathbf{r}',\nu)|$ is a fast varying function of $\mathbf{r}'$ (**Fig. 10(a)**). In addition, the linear dimensions of the source are large compared with the wavelength of light and with the correlation length Δ (**Fig. 10(c)**).

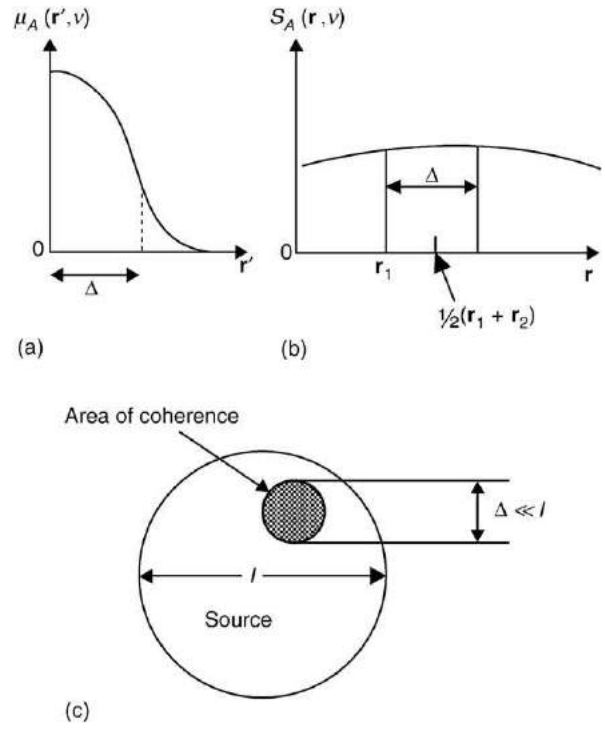


Fig. 10 Concept of quasi-homogeneous sources.

Quasi-homogeneous sources are always spatially incoherent in the 'global' sense, because their linear dimensions are large compared with the correlation length. This model is very good for representing two-dimensional secondary sources with sufficiently low coherence such that the intensity does not vary over the coherence area on the input plane. It has also been applied to three-dimensional primary sources, three-dimensional scattering potentials, and two-dimensional primary and secondary sources.

Equivalence Theorems

The study of partially coherent sources led to the formulations of a number of equivalence theorems, which show that sources of any state of coherence can produce the same distribution of the radiant intensity as a fully spatially coherent laser source. These theorems provide conditions under which sources of different spatial distribution of spectral density and of different state of coherence will generate fields, which have the same radiant intensity. It has been shown, by taking examples of Gaussian–Schell model sources, that sources of completely different coherence properties and different spectral distributions across the source generate identical distribution of radiant intensity. Experimental verifications of the results of these theorems have also been carried out. For further details on this subject, the reader is referred to Mandel and Wolf (1995).

Correlation-Induced Spectral Changes

It was assumed that spectrum is an intrinsic property of light that does not change as the radiation propagates in free-space, until studies on partially coherent sources and radiations from them in 1980s, revealed that this was true only for specific type of sources. It was discovered on general grounds that the spectrum of light, which originates in an extended source, either a primary source or a secondary source, depends not only on the source spectrum but also on the spatial coherence properties of the source. It was also predicted theoretically by Wolf that the spectrum of light would, in general, be different from the spectrum of the source, and be different at different points in space on propagation in free space.

For a quasi-homogeneous planar secondary source defined by **Eq. (52)**, whose normalized spectrum is the same at each source point, one can write the spectral density as

$$S_0(\boldsymbol{\rho}, \nu) = I_0(\boldsymbol{\rho}) g_0(\nu) \quad \text{with} \quad \int_0^\infty g_0(\nu)\mathrm{d}\nu = 1 \tag{53}$$

where $I_0(\boldsymbol{\rho})$ is the intensity of light at point $\boldsymbol{\rho}$ on the plane of the source, $g_0(\nu)$ is the normalized spectrum of the source and the subscript 0 refers to the quantities of the source plane. Using **Eq. (49)**, one can obtain an expression for the far-field spectrum due to this source as

$$S^\infty(r\mathbf{u}, \nu) = \frac{k^2}{2\pi^2}\frac{\cos^2\theta}{r^2}\int_\sigma\int_\sigma I_0(\boldsymbol{\rho}) g_0(\nu)\mu_0(\boldsymbol{\rho}', \nu)\mathrm{e}^{-\mathrm{i}k\mathbf{u}_\perp\cdot(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2)}\mathrm{d}^2\rho_1\mathrm{d}^2\rho_2 \tag{54}$$

Noting that $\boldsymbol{\rho} = (\boldsymbol{\rho}_1 + \boldsymbol{\rho}_2)/2$ and $\boldsymbol{\rho}' = \boldsymbol{\rho}_2 - \boldsymbol{\rho}_1$, one can transform the variables of the integration and obtain after some manipulation:

$$S^\infty(r\mathbf{u}, \nu) = \frac{k^2\cos^2\theta}{(2\pi r)^2}\tilde{I}_0\tilde{\mu}_0(k\mathbf{u}_\perp, \nu) g_0(\nu) \tag{55}$$

where

$$\tilde{\mu}_0(k\mathbf{u}_\perp, \nu) = \int_\sigma \mu_0(\boldsymbol{\rho}', \nu)\mathrm{e}^{-\mathrm{i}k\mathbf{u}_\perp\cdot\boldsymbol{\rho}'}\mathrm{d}^2\rho' \tag{56}$$

and

$$\tilde{I}_0 = \int_\sigma I_0(\boldsymbol{\rho})\mathrm{d}^2\rho \tag{57}$$

Eq. (55) shows that the spectrum of the field in the far-zone depends on the coherence properties of the source through its spectral degree of coherence $\mu_0(\boldsymbol{\rho}', \nu)$ and on the normalized source spectrum $g_0(\nu)$.

Scaling Law

The reason why coherence-induced spectral changes were not observed until recently is that the usual thermal sources employed in laboratories or commonly encountered in nature have special coherence properties and the spectral degree of coherence has the function form:

$$\mu^{(0)}(\boldsymbol{\rho}_2 - \boldsymbol{\rho}_1, \nu) = f[k(\boldsymbol{\rho}_2 - \boldsymbol{\rho}_1)] \quad \text{with} \quad k = \frac{2\pi\nu}{c} \tag{58}$$

which shows that the spectral degree of coherence depends only on the product of the frequency and space coordinates. This formula expresses the so-called scaling law, which was enunciated by Wolf. Commonly used sources satisfy this property.

For example, the spectral degree of coherence of Lambertian sources and black-body sources can be shown to be

$$\mu_0(\boldsymbol{\rho}_2 - \boldsymbol{\rho}_1, \nu) = \frac{\sin(k|\boldsymbol{\rho}_2 - \boldsymbol{\rho}_1|)}{k|\boldsymbol{\rho}_2 - \boldsymbol{\rho}_1|} \tag{59}$$

This expression evidently satisfies the scaling law. If the spectral degree of coherence does not satisfy the scaling law, the normalized far-field spectrum will, in general, vary in different directions in the far-zone and will differ from the source spectrum.

Spectral Changes in Young's Interference Experiment

Spectral changes in Young's interference experiment with broadband light are not as well understood as in experiments with quasi-monochromatic, probably because in such experiments no interference fringes are formed. However, if one were to analyze the spectrum of the light in the region of superposition, one would observe changes in the spectrum of light in the region of superposition in the form of a shift in the spectrum for narrowband spectrum and spectral modulation for broadband light. One can readily derive an expression for the spectrum of light in the region of superposition. Let $S^{(1)}(P, \nu)$ be the spectral density of the light at P which would be obtained if the small aperture at P_1 alone was open; $S^{(2)}(P, \nu)$ has a similar meaning if only the aperture at P_2 was open (**Fig. 4**).

Let us assume, as is usually the case, that $S^{(2)}(P, \nu) \approx S^{(1)}(P, \nu)$ and let d be the distance between the two pinholes. Consider the spectral density at the point P, at distance x from the axis of symmetry in an observation plane located at distance of R from the plane containing the pinholes. Assuming that $x/R \ll 1$, one can make the approximation $R_2 - R_1 \approx xd/R$. The spectral interference law (**Eq. (38)**) can then be written as

$$S(P, \nu) \approx 2S^{(1)}(P, \nu)\{1 + |_{\mu}(P_1, P_2, \nu)|\cos[\beta(P_1, P_2, \nu) + 2\pi\nu xd/cR]\} \tag{60}$$

where $\beta(P_1,P_2,\nu)$ denotes the phase of the spectral degree of coherence. **Eq. (60)** implies the two results:

(i) at any fixed frequency ν, the spectral density varies sinusoidally with the distance x of the point from the axis, with the amplitude and the phase of the variation depending on the (generally complex) spectral degree of coherence $\mu(P_1, P_2,\nu)$; and

(ii) at any fixed point P in the observation plane the spectrum $S(P,\nu)$ will, in general, differ from the spectrum $S^{(1)}(P,\nu)$, the change also depending on the spectral degree of coherence $\mu(P_1,P_2,\nu)$ of the light at the two pinholes.

Experimental Confirmations

Experimental tests of the theoretical prediction of spectral invariance and noninvariance due to correlation of fluctuations across the source were performed just after the theoretical predictions. **Fig. 11** shows results of one such experiment in which spectrum changes in the Young's experiment were studied. Several other experiments also reported confirmation of the source correlation-dependent spectral changes. One of the important applications of these observations has been to explain the discrepancies in the maintenance of the spectroradiometric scales by national laboratories in different countries. These studies also have potential application in determining experimentally the spectral degree of coherence of partially coherent fields. The knowledge of spectral degree of coherence is often important in remote sensing, e.g., for determining angular diameters of stars.

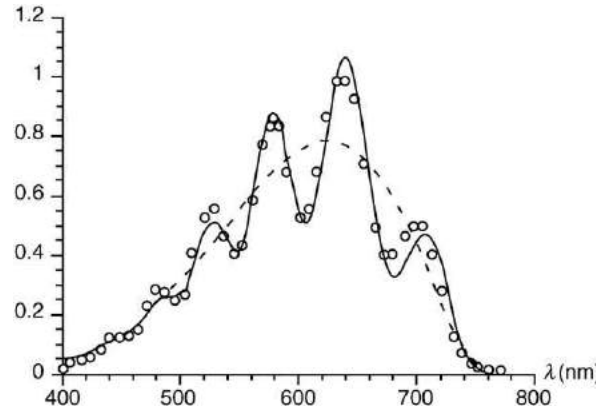


Fig. 11 Correlation-induced changes in spectrum in Young's interference. Dashed line represents the spectrum when only one of the slits is open, the continuous curve shows the spectrum when both the slits are open and the circles are the measured values in the latter case. Reproduced with permission from Santarsiero M and Gori F (1992) Spectral changes in Young interference pattern. *Physics Letters* 167: 123–128.

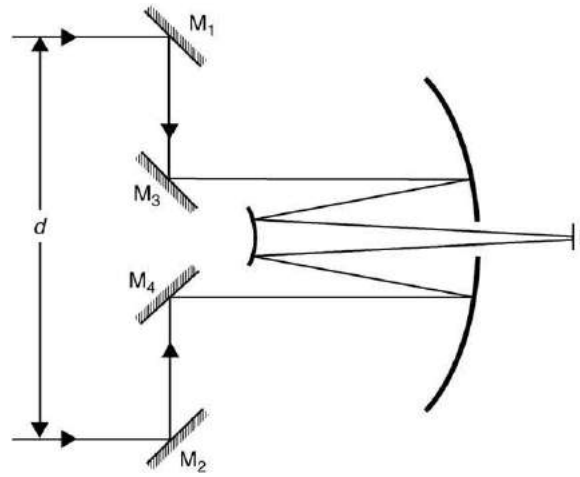


Fig. 12 Schematic of the Michelson stellar interferometer.

Applications of Optical Coherence

Stellar Interferometry

The Michelson stellar interferometer, named after Albert Michelson, was used to determine the angular diameters of stars and also the intensity distribution across the star. The method was devised by Michelson without using any concept of coherence, although subsequently the full theory of the method was developed on the basis of propagation of correlations. A schematic of the experiment is shown in **Fig. 12**. The interferometer is mounted in front of a telescope, a reflecting telescope in this case. The light from a star is reflected from mirrors M_1 and M_2 and is directed towards the primary mirror (or the objective lens) of the telescope. The two beams thus collected superpose in the focal plane F of telescope where an image crossed with fringes is formed. The outer mirrors M_1 and M_2 can be moved along the axis defined as $M_1M_3M_4M_2$ while the inner mirrors M_3 and M_4 remain fixed. The fringe spacing depends on the position of mirrors M_3 and M_4 and hence is fixed, while the visibility of the fringes depends on the separation of the mirrors M_1 and M_2 and hence, can be varied. Michelson showed that from the measurement of the variation of the visibility with the separation of the two mirrors, one could obtain information about the intensity distribution of the stars, which are rotationally symmetric. He also showed that if the stellar disk is circular and uniform, the visibility curve as a function of the separation d of the mirrors M_1 and M_2 will have zeros for certain values of d, and that the smallest of these d values for which zero occurs is given by $d_0 = 0.61\ \lambda_a/\alpha$, where α is the semi-angular diameter of the star and λ_a is the mean wavelength of the filtered quasi-monochromatic light from the star. Angular diameters of several stars down to 0.02 second of an arc were determined.

From the standpoint of second-order coherence theory the principles of the method can be readily understood. The star is considered an incoherent source and according to the van Cittert–Zernike theorem, the light reaching the outer mirrors M_1 and M_2 of the interferometer will be partially coherent. This coherence would depend on the size of and the intensity distribution across the star. Let (x_1, y_1) and (x_2, y_2) be the coordinates of the positions of the mirrors M_1 and M_2, respectively, and (ξ, η) the coordinates of a point on the surface plane of the star which is assumed to be at a very large (astronomical) distance R from the mirrors. The complex degree of coherence at the mirrors would then be given by **Eq. (22)** which can now be written as

$$\gamma(\Delta x, \Delta y, 0) = \frac{\int_\sigma I(u,v) e^{-ik_a(u\Delta x + v\Delta y)} du\,dv}{\int_\sigma I(u,v)\,du\,dv} \tag{61}$$

where $I(u, v)$ is the intensity distribution across the star disk σ as a function of the angular coordinates $u = \xi/R$, $v = \eta/R$, $\Delta x = x_1 - x_2$, $\Delta y = y_1 - y_2$, and $k_a = 2\pi/\lambda_a$, λ_a being the mean wavelength of the light from the star. **Eq. (61)** shows that the equal-time ($\tau = 0$) complex degree of coherence of the light incident at the outer mirrors of the interferometer is the normalized Fourier-transform of the intensity distribution across the stellar disk. Further, **Eq. (15)** shows that the visibility of the interference fringes is the absolute value of γ, if the intensity of the two interfering beams is equal, as in the present case. The phase of γ can be determined by the position of the intensity maxima of the fringe pattern (**Eq. (14)**). If one is interested in determining only the angular size of the star and the star is assumed to be a circularly symmetric disk of angular diameter 2α and of uniform intensity [$I(u, v)$ is constant across the disk], then **Eq. (61)** reduces to

$$\gamma(\Delta x, \Delta y) = \frac{2J_1(v)}{v}, \quad v = \frac{2\pi\alpha}{\lambda_a} d, \quad d = \sqrt{(\Delta x)^2 + (\Delta y)^2} \tag{62}$$

The smallest separation of the mirrors for which the visibility γ vanishes corresponds to $v = 3.832$, i.e., $d_0 = 0.61\ \lambda_a/\alpha$, which is in agreement with Michelson's result.

Interference Spectroscopy

Another contribution of Michelson, which was subsequently identified as an application of the coherence theory, was the use of his interferometer (Fig. 2) to determine the energy distribution in the spectral lines. The method he developed is capable to resolving spectral lines that are too narrow to be analyzed by the spectrometer. The visibility of the interference fringes depends on the energy distribution in the spectrum of the light and its measurement can give information about the spectral lines. In particular, if the energy distribution in a spectral line is symmetric about some frequency ν_0, its profile is simply the Fourier transform of the visibility variation as a function of the path difference between the two interfering beams. This method is the basis of interference spectroscopy or the Fourier transform spectroscopy.

Within the framework of second-order coherence theory, if the mean intensity of the two beams in a Michelson's interferometer is the same, then the visibility of the interference fringes in the observation plane is related to the complex degree of coherence of light at a point at the beamsplitter where the two beams superimpose (Eq. (15)). The two quantities are related as $\upsilon(\tau)=|\gamma(\tau)|$ where $\gamma(\tau) \equiv \gamma(\mathbf{r}_1, \mathbf{r}_1, \tau)$ is the complex degree of self-coherence at the point $\mathbf{r}_1$ on the beamsplitter. Following Eq. (19), $\gamma(\tau)$ can be represented by a Fourier integral as

$$\gamma(\tau) = \int_0^\infty s(\nu)\exp(-\mathrm{i}2\pi\nu\tau)\mathrm{d}\nu \tag{63}$$

where $s(\nu)$ is the normalized spectral density of the light defined as $s(\nu) = S(\nu)/\int_0^\infty S(\nu)\mathrm{d}\nu$ and $S(\nu) \equiv S(\mathbf{r}_1, \nu) = W(\mathbf{r}_1, \mathbf{r}_1, \nu)$ is the spectral density of the beam at the point $\mathbf{r}_1$. The method is usually applied to very narrow spectral lines for which one can assume that the peak that occurs at ν_0, and $\gamma(\tau)$ can be represented as

$$\begin{aligned} \gamma(\tau) &= \overline{\gamma}(\tau)\exp(-2\pi\mathrm{i}\nu_0\tau) \quad \text{with} \\ \overline{\gamma}(\tau) &= \int_{-\infty}^{\infty} \overline{s}(\mu)\exp(-\mathrm{i}2\pi\mu\tau)\mathrm{d}\mu \end{aligned} \tag{64}$$

where $\overline{s}(\mu)$ is the shifted spectrum such that

$$\begin{aligned} \overline{s}(\mu) &= s(\nu_0+\mu) \quad \mu \geq -\nu_0 \\ &= 0 \qquad\quad \mu < -\nu_0 \end{aligned}$$

From the above one readily gets

$$\upsilon(\tau) = |\gamma(\tau)| = \left| \int_{-\infty}^{\infty} \overline{s}(\mu)\exp(-\mathrm{i}2\pi\mu\tau)\mathrm{d}\mu \right| \tag{65}$$

If the spectrum is symmetric about ν_0, then $\upsilon(\tau)$ would be an even function of τ and the Fourier inversion would give:

$$\overline{s}(\mu) = s(\nu_0+\mu) = 2\int_0^\infty \upsilon(\tau)\cos(2\pi\mu\tau)\mathrm{d}\tau \tag{66}$$

which can be used to calculate the spectral energy distribution for a symmetric spectrum about ν_0 from the visibility curve. However, for an asymmetric spectral distribution, the visibility and the phase of the complex degree of coherence must be determined as the Fourier transform of the shifted spectrum is no longer real everywhere.

Higher-Order Coherence

So far we have considered correlations of the fluctuating field variables at two space–time $(\mathbf{r},t)$ points, as defined in Eq. (10). These are termed as second-order correlations. One can extend the concept of correlations to more than two space–time points, which will involve higher-order correlations. For example, one can define the space–time cross correlation function of order (M,N) of the random field $V(\mathbf{r},t)$, represented by $\Gamma^{(M,N)}$, as an ensemble average of the product of the field $V(\mathbf{r},t)$ values at N space–time points and $V^*(\mathbf{r},t)$ at other M points. In this notation, the mutual coherence function as defined in Eq. (10) would now be $\Gamma^{(1,1)}$. Among higher-order correlations, the one with $M=N=2$, is of practical significance and is called the fourth-order correlation function, $\Gamma^{(2,2)}(\mathbf{r}_1,t_1,\mathbf{r}_2,t_2;\mathbf{r}_3,t_3,\mathbf{r}_4,t_4)$. The theory of Gaussian random variables tells us that any higher correlation can be written in terms of second-order correlations over all permutations of pairs of points. In addition, if we assume that $(\mathbf{r}_3,t_3)=(\mathbf{r}_1,t_1)$ and $(\mathbf{r}_4,t_4)=(\mathbf{r}_2, t_2)$, and that the field is stationary, then $\Gamma^{(2,2)}$ is called the intensity–intensity correlation and is given as

$$\begin{aligned} &\Gamma^{(2,2)}(\mathbf{r}_1,\mathbf{r}_2,t_2-t_1) \\ &\quad = \langle V(\mathbf{r}_1,t_1)V(\mathbf{r}_2,t_2)V^*(\mathbf{r}_1,t_1)V^*(\mathbf{r}_2,t_2)\rangle \\ &\quad = \langle I(\mathbf{r}_1,t_1)I(\mathbf{r}_2,t_2)\rangle \\ &\quad = \langle I(\mathbf{r}_1,t_1)\rangle\langle I(\mathbf{r}_2,t_2)\rangle(1+|\gamma^{(1,1)}(\mathbf{r}_1,\mathbf{r}_2,t_2-t_1)|^2) \end{aligned} \tag{67}$$

where

$$\gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1) = \frac{\Gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1)}{[\langle I(\mathbf{r}_1, t_1)\rangle]^{1/2}[\langle I(\mathbf{r}_2, t_2)\rangle]^{1/2}} \tag{68}$$

We now define fluctuations in intensity at $(\mathbf{r}_j, t_j)$ as

$$\Delta I_j = I(\mathbf{r}_j, t_j) - \langle I(\mathbf{r}_j, t_j)\rangle$$

and then the correlation of intensity fluctuations becomes

$$\begin{aligned}\langle \Delta I_1 \Delta I_2\rangle &= \langle I(\mathbf{r}_1, t_1) I(\mathbf{r}_2, t_2)\rangle - \langle I(\mathbf{r}_1, t_1)\rangle\langle I(\mathbf{r}_2, t_2)\rangle \\ &= \langle I(\mathbf{r}_1, t_1)\rangle\langle I(\mathbf{r}_2, t_2)\rangle|\gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1)|^2\end{aligned} \tag{69}$$

where we have used **Eq. (67)**. **Eq. (69)** forms the basis for intensity interferometry.

Hanbury-Brown and Twiss Experiment

In this landmark experiment conducted both on the laboratory scale and astronomical scale, Hanbury-Brown and Twiss demonstrated the existence of intensity–intensity correlations in terms of the correlations in the photocurrents in the two detectors and thus measured the squared modulus of complex degree of coherence. In the laboratory experiment, the arc of a mercury lamp was focused onto a circular hole to produce a secondary source. The light from this source was then divided equally into two parts through a beamsplitter. Each part was, respectively, detected by two photomultiplier tubes, which had identical square apertures in front. One of the tubes could be translated normally to the direction of propagation of light and was so positioned that its image through the splitter could be made to coincide with the other tube. Thus, by suitable translation a measured separation d between the two square apertures could be introduced. The output currents from the photomultiplier tubes were taken by cables of equal length to a correlator. In the path of each cable a high-pass filter was inserted, so that only the current fluctuations could be transmitted to the correlator. Thus, the normalized correlations between the two current fluctuations:

$$\mathrm{C}(d) = \frac{\langle \Delta J_1(t)\Delta J_2(t)\rangle}{\langle[\Delta J_1(t)]^2\rangle^{1/2}\langle[\Delta J_2(t)]^2\rangle^{1/2}} \tag{70}$$

as a function of detector separation d could be measured. Now when the detector response time is much larger than the time-scale of the fluctuations in intensity, then it can be shown that the correlations in the fluctuations of the photocurrent are proportional to the correlations of the fluctuations in intensity of the light being detected. Thus, we would have

$$\mathrm{C}(d) \approx \delta|\gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, 0)|^2 \tag{71}$$

where δ is the average number of photocounts of the light of one polarization during the time-scale of the fluctuations (for general thermal sources, this is much less than one). **Eq. (71)** represents the Hanbury-Brown–Twiss effect.

Stellar Intensity Interferometry

Michelson stellar interferometry can resolve stars which have angular sizes of the order of 0.01″, since for smaller stars, the separation between the primary mirrors runs into several meters and maintaining stability of mirrors such that the optical paths do not change, even by fraction of a wavelength, is extremely difficult. The atmospheric turbulence further adds to this problem and obtaining stable fringe pattern becomes next to impossible for very small stars. Hanbury-Brown and Twiss applied the intensity interferometry based on their photoelectric correlation technique for determining the angular sizes of such stars. Two separate parabolic mirrors collected light from the star and the output of the photodetectors placed at the focus of each mirror was sent to a correlator. The cable lengths were made unequal so as to compensate for the time difference of the light arrival at the two mirrors. The normalized correlation of the fluctuations of the photocurrents was determined as described above. This would give the variation of the modulus-squared degree of coherence as a function of the mirror separation d from which the angular size of the stars can be estimated. The advantage of the stellar intensity interferometer over the stellar (amplitude) interferometer is that the light need not interfere as in the latter, since the photodetectors are mounted directly at the focus of the primary mirrors of the telescope. Thus, the constraint on the large path difference between the two beams is removed and large values of d can now be used. Moreover, the atmosphere turbulence and the mirror movements have very small effect. Stellar angular diameters as small as 0.0004″ of arc with resolution of 0.00003″ could be measured by such interferometers.

Further Reading

Beran, M.J., Parrent, G.B., 1964. Theory of Partial Coherence. Englewood Cliffs, NJ: Prentice-Hall.
Born, M., Wolf, E., 1999. Principles of Optics. New York: Pergamon Press.
Carter, W.H., 1996. Coherence theory. In: Bass, M. (Ed.), Hand Book of Optics. New York: McGraw-Hill.
Davenport, W.B., Root, W.L., 1960. An Introduction to the Theory of Random Signals and Noise. New York: McGraw-Hill.
Goodman, J.W., 1985. Statistical Optics. Chichester: Wiley.

Hanbury-Brown, R., Twiss, R.Q., 1957. Interferometry of intensity fluctuations in light: I basic theory: the correlation between photons in coherent beams of radiation. Proceedings of the Royal Society 242, 300–324.
Hanbury-Brown, R., Twiss, R.Q., 1957. Interferometry of intensity fluctuations in light: II an experimental test of the theory for partially coherent light. Proceedings of the Royal Society 243, 291–319.
Kandpal, H.C., Vaishya, J.S., Joshi, K.C., 1994. Correlation-induced spectral shifts in optical measurements. Optical Engineering 33, 1996–2012.
Mandel, L., Wolf, E., 1965. Coherence properties of optical fields. Review of Modern Physics 37, 231–287.
Mandel, L., Wolf, E., 1995. Optical Coherence and Quantum Optics. Cambridge: Cambridge University Press.
Marathay, A.S., 1982. Elements of Optical Coherence. New York: John Wiley.
Perina, J., 1985. Coherence of Light. Boston: M.A. Reidel.
Santarsiero, M., Gori, F., 1992. Spectral changes in Young interference pattern. Physics Letters 167, 123–128.
Schell, A.C., 1967. A technique for the determination of the radiation pattern of a partially coherent aperture. IEEE Transactions on Antennas and Propagation AP-15, 187.
Thompson, B.J., 1958. Illustration of phase change in two-beam interference with partially coherent light. Journal of the Optical Society of America 48, 95–97.
Thompson, B.J., Wolf, E., 1957. Two beam interference with partially coherent light. Journal of the Optical Society of America 47, 895–902.
Wolf, E., James, D.F.V., 1996. Correlation-induced spectral changes. Report on Progress in Physics 59, 771–818.

Diffraction Gratings

J Turunen and T Vallius, University of Joensuu, Joensuu, Finland

Nomenclature

U	Complex field amplitude	H	Grating thickness [m]
θ, ϕ	Diffraction angles [rad]	Q	Number of quantization levels
η	Diffraction efficiency	$\phi(r)$	Phase function [rad]
f, f_0	Focal length [m]	n	Refractive index
d	Grating period [m]	λ,λ_0	Wavelength [m]
		k	Wave vector [m^{-1}]

Introduction

The classic textbook example of a diffraction grating is a periodic arrangement of parallel opaque wires: if a plane wave is incident upon such a structure, it is divided into a multitude of plane waves propagating in well-defined directions. In particular, the dramatic wavelength-dependence of these directions has attracted great interest ever since the potential of diffraction gratings in spectroscopy was realized in the late 1700s. Today spectroscopy is only one, though important, application of diffraction gratings. Gratings of many different forms belong to the basic building blocks in modern wave-optics-based design of optical elements and systems.

Grating Equations

Some examples of the wide variety of possible grating structures are illustrated in **Fig. 1**, which shows the profiles of some commonly encountered diffraction gratings that are invariant in the y direction. These are known as linear gratings. Each grating may be of reflection type (the refractive index n_{III} is complex) or of transmission type (n_{III} is real). In all cases n_1 and n_2 can both be real or one of them can be complex. Here we assume that the incident field is a plane wave with wave vector k in the x–z plane; more general illumination waves can be represented as superpositions of plane waves.

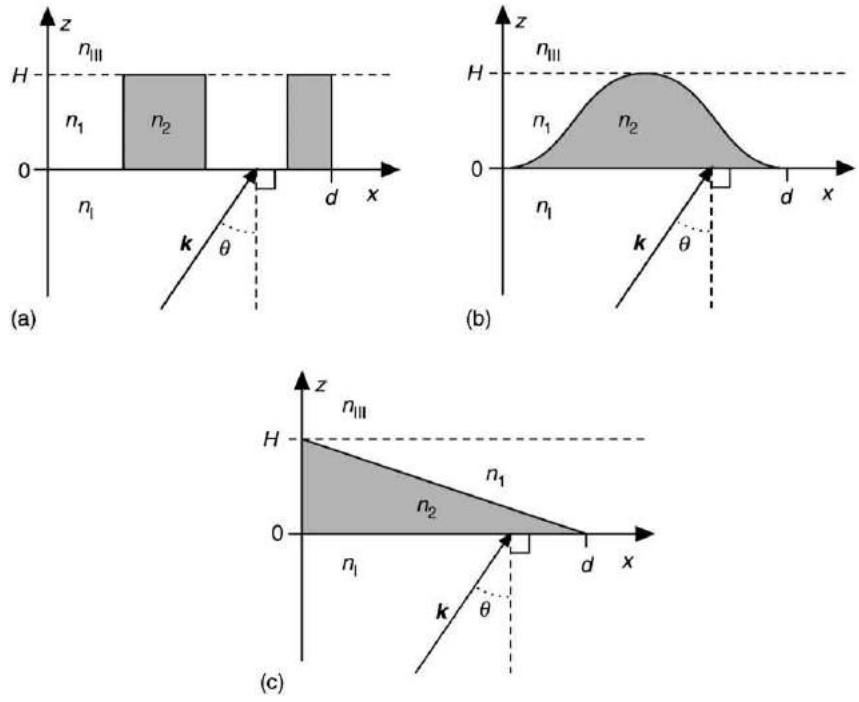


Fig. 1 Examples of linear gratings. (a) Binary grating; (b) sinusoidal grating; (c) triangular grating. Here d is the grating period, H is the thickness of the modulated region, θ is the angle of incidence, and n denotes refractive index.

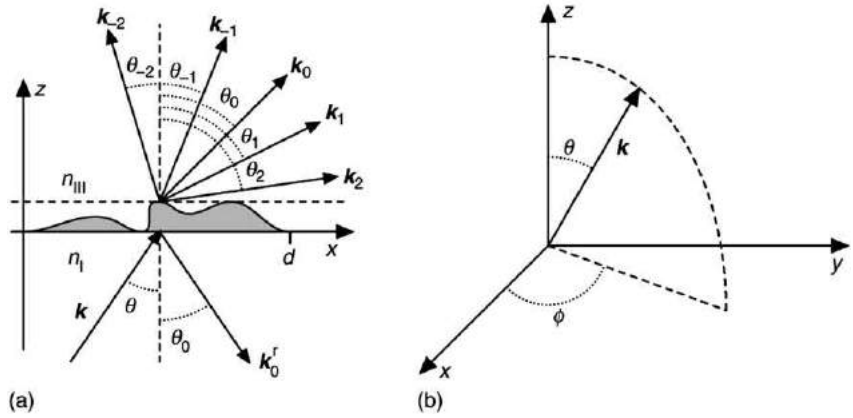


Fig. 2 Diffraction by gratings. (a) Linear grating; (b) definitions of direction angles in the case of a biperiodic grating.

Fig. 1(a) illustrates a binary grating: in the modulated region $0<z<H$ the refractive index alternates between n_1 and n_2, changing from one value to another at certain transition points within the period d. The structure in Fig. 1(b) represents a sinusoidal surface-relief grating as an example of smooth grating profiles. Finally, Fig. 1(c) shows a triangular grating profile.

The periodicity of a grating, illuminated by a plane wave, usually leads to the appearance of a set of reflected and transmitted plane waves (diffraction orders of the grating) that propagate in directions determined by the grating period d, the illumination wavelength λ, and the refractive indices n_{I} and n_{III} (see Fig. 2(a)). These directions are obtained from the transmission grating equation:

$$n_{\mathrm{III}}\sin\theta_m = n_{\mathrm{I}}\sin\theta + m\lambda/d \tag{1}$$

where m is the index of the diffraction order and θ_m is its propagation angle. For reflected orders (superscript r) one simply replaces n_{III} by n_{I} in Eq. (1).

Obviously only a limited number of orders satisfy the condition $|\theta_m|<\pi/2$ and thereby represent conventional propagating plane waves. Higher, so-called evanescent orders are inhomogeneous waves that propagate along the grating surface and decay exponentially in the z direction. However, they cannot be ignored in grating analysis, as we shall see.

In the case of a biperiodic grating with a period $d_x \times d_y$ the propagation direction of the incident beam is defined by two angles θ and ϕ illustrated in Fig. 2(b). The grating equations now read as:

$$n_{\mathrm{III}}\sin\theta_{\mathrm{mn}}\cos\phi_{\mathrm{mn}} = n_{\mathrm{I}}\sin\theta\cos\phi + m\lambda/d_x \tag{2}$$

$$n_{\mathrm{III}}\sin\theta_{\mathrm{mn}}\sin\phi_{\mathrm{mn}} = n_{\mathrm{I}}\sin\theta\sin\phi + n\lambda/d_y, \tag{3}$$

where θ_{mn} and ϕ_{mn} define the propagation directions of the diffraction order with indices m and n. For reflected orders n_{III} is again replaced by n_{I}. These expressions apply also to the special case of linear gratings (limit $d_y\to\infty$) illuminated by a plane wave not propagating in the $x-z$ plane (conical incidence).

Some two-dimensionally periodic gratings are illustrated in Fig. 3. The structure illustrated in Fig. 3(a) is an extension of a binary grating to the biperiodic geometry: here a single two-dimensional feature is placed within the grating period but, of course, more features with different outlines could be included as well. Moreover, each period could contain a continuous 'mountain-range' or consist of a 'New-York'-style pixel structure. Fig. 3(b) illustrates a structure consisting of an array of rectangular-shaped holes pierced in a flat screen, while Fig. 3(c) defines an array of isolated particles on a flat substrate. If the grid or the particles are metallic, surrounded by dielectrics from all sides, one speaks of inductive and capacitive structures, respectively: in the former case relatively free-flowing currents are induced in the metallic grid structure, but in the latter case the isolated metal particles become polarized (and therefore capacitive) when illuminated by an electromagnetic field.

Grating Theory

Gratings equations provide exact knowledge of the propagation directions of the diffraction orders, but they give no information about the distribution of light among different orders. To obtain such information, one must in general solve Maxwell's equations exactly (rigorous diffraction theory), but within certain limits simpler methods are available. In all cases, however, the transmitted and reflected fields are represented as superpositions of plane waves that propagate in the directions allowed by the grating equations (these are known as Rayleigh expansions).

The most popular method for rigorous grating analysis is the Fourier Modal Method (FMM), though many alternative analysis methods exist. To understand the principle of FMM, consider first binary gratings. The field inside the modulated region $0<z<H$ is represented as a superposition of waveguide modes, determined by imagining that the modulated region would extend from $x=-\infty$ to $x=\infty$. With given transition points the electric permittivity, its inverse, and the electric field inside the grating, can be represented in the form of Fourier series and one ends up with a matrix eigenvalue problem with finite dimensions when the

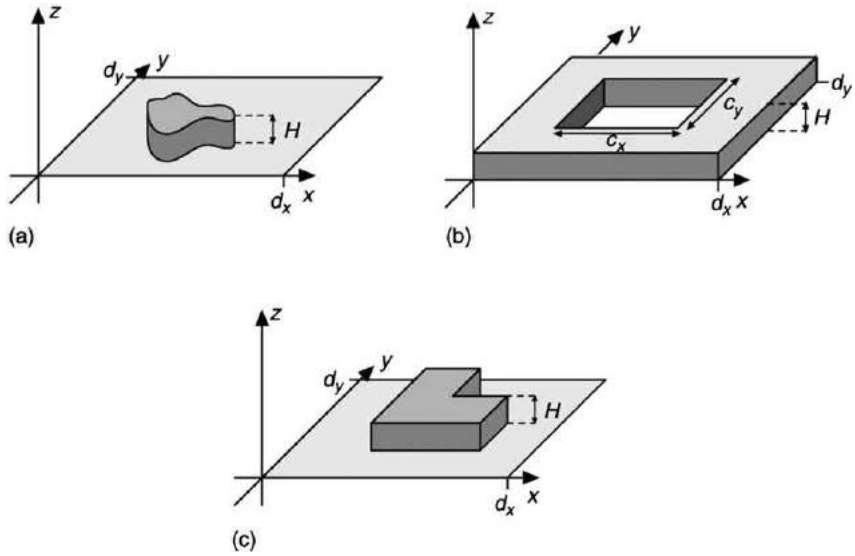


Fig. 3 Examples of biperiodic gratings. (a) General binary grating; (b) inductive structure; (c) capacitive structure.

matrices are truncated. A numerical solution of this problem provides a set of orthogonal modes that can propagate in the structure with certain propagation constants. The relative weights of the modes are then determined by applying the electromagnetic boundary conditions at the planes $z=0$ and $z=H$ to match the modal expansions inside the grating with the plane-wave (Rayleigh) expansions of the fields in the homogeneous materials on both sides of the grating. This matching procedure, which is numerically implemented by solving a set of simultaneous linear equations, yields the complex amplitudes of the electric and magnetic field components associated with the reflected and transmitted diffraction orders of the grating.

Of main interest are the diffraction efficiencies, η_m (or η_{mn} for biperiodic gratings) of the orders, which tell the distribution of energy between different propagating orders. In electromagnetic theory the diffraction efficiency is defined as the z-component of the time-averaged Poynting vector associated with the particular order, divided by that of the incident plane wave.

Gratings with z-dependent surface profiles can be handled similarly if they are 'sliced' into a sufficient number of layers, in each of which the grating is modeled as a binary structure. The electromagnetic boundary conditions are applied at each slice boundary and the approach models the real continuous structure with arbitrary degree of accuracy when the number of the slices is increased; in practice it is usually sufficient to use approximately 20–30 slices for sufficiently accurate modeling of any continuous grating profile by FMM.

The main problem with rigorous techniques such as FMM is bad scaling. If, in the case of 1D modulated (or y-invariant) gratings, the period is increased by a factor of two, the computation time required to solve the eigenvalue problem is increased by a factor of around eight. Thus we have a d^3 type scaling law. The situation is much worse for bi-periodic gratings, for which the exponent is around six. Thus no foreseeable developments in computer technology will solve the problem, and approximate analysis methods have to be sought. A multitude of such methods are under active development, but in this context we discuss the simplest example only.

If $d \gg \lambda$ and the grating profile does not contain wavelength-scale features, one can describe the effects of a grating on the incident field merely by considering it as a complex-amplitude-modulating object, which imposes a position-dependent phase shift and amplitude transmission (or reflection) coefficient on the incident field. This model is known as the thin-element approximation (TEA).

At normal incidence the plane-wave expansion of any scalar field component U immediately behind the grating (at $z=H$) takes the form:

$$U(x,y,H) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} T_{\mathrm{mn}} \times \exp\left[\mathrm{i}2\pi\left(\mathrm{m}x/d_x + \mathrm{n}y/d_y\right)\right] \tag{4}$$

where

$$T_{\mathrm{mn}} = \frac{1}{d_x d_y} \int_0^{d_x} \int_0^{d_y} U(x,y,H) \times \exp\left[-\mathrm{I}2\pi\left(\mathrm{m}x/d_x + \mathrm{n}y/d_y\right)\right] \mathrm{d}x\mathrm{d}y \tag{5}$$

are the complex amplitudes of the transmitted diffraction orders. In TEA the field at $z=H$ may be connected to the field at $z=0$ (a unit-amplitude plane wave is assumed here) by optical path integration along a straight line though the modulated region of the grating:

$$U(x,y,H) = \int_0^H \exp\left[\mathrm{i}\frac{2\pi}{\lambda}\hat{n}(x,y,z)\right] \mathrm{d}z \tag{6}$$

where $\hat{n}(x,y,z) = n(x,y,z) + i\kappa(x,y,z)$ is the complex refractive-index distribution in the region $0 < z < H$. With TEA the diffraction efficiencies can be calculated from a simple formula:

$$\eta_{mn} = |T_{mn}|^2 \tag{7}$$

which, however, ignores Fresnel reflections.

Eqs. (5)–(7) allow one to obtain closed-form expressions for the diffraction efficiencies η_{mn} in many grating geometries. For example, a y-invariant triangular profile with $H = \lambda/|n_I - n_{III}|$ gives $\eta_{-1} = 100\%$ if $n_1 = n_{III}$, $n_2 = n_I$, and $\theta = 0$. For a Q-level stair-step-quantized version of this grating we obtain:

$$\eta_m = \mathrm{sinc}^2(1/Q) \tag{8}$$

where $cx = \sin(\pi x)/(\pi x)$ It should be noted, however, that these results are valid only for gratings with $d \gg \lambda$, even if they are corrected by Fresnel transmission losses at the interface between the two media.

Fabrication of Gratings

Gratings can be manufactured in a number of different ways, perhaps the simplest being to plot an equidistant array of straight black and white lines and to reduce the scale of the pattern photographically. This method can be readily adapted to the generation of more complex grating structures and diffractive elements, but it is not suitable for the production of high-quality gratings required in, for example, spectroscopy.

Ruling engines represent the traditional way of making gratings with periods of the order of the wavelength of light, especially for spectroscopic applications. These are machines equipped with a triangularly shaped diamond ruling edge and an interferometrically controlled two-dimensional mechanical translation stage that moves the substrate with respect to the ruling edge. Triangular-profile gratings of high quality can be produced with such machines, but the fabrication process is slow and the cost of the manufactured gratings is therefore high. In addition, even with real-time interferometric control of the ruling-edge position, it is impossible to avoid small local errors in the grating geometry. Such errors are particularly harmful if they are periodic since, according to the grating equation, these errors will generate spurious spectral lines known as ghosts. Such lines generated by the grating 'super-period' are weak but they may be confused with real spectral lines in the source spectrum. If the ruling errors are random, they result in 'pedestal-type' background noise. While such errors do not introduce ghosts, they nevertheless reduce the visibility of weak original spectral lines.

The introduction of lasers in the early 1960s made it possible to form stable and high-contrast interference patterns over a sufficient time-scale to allow the exposure of photographic films and other photosensitive materials such as polymeric photoresists. This interferometric technique offers a direct method for the fabrication of large diffraction gratings without appreciable periodic errors. With this technique it is possible to make highly regular gratings for spectroscopy, but it does not permit easy generation of arbitrarily shaped grating structures. The two intersecting plane waves can also be generated with a linear grating using diffraction orders $m = +1$ and $m = -1$, suppressing the zero order by appropriate design of the grating profile and choosing the period in such a way that all higher transmitted orders are evanescent. This so-called phase mask technique is suitable, for example, for the exposure of Bragg gratings in fibers.

To generate arbitrary gratings one usually employs lithographic techniques. Some of the possible processes are illustrated in Fig. 4. The exposure can be performed with scanning photon, electron, or ion beams, etc. The process starts with substrate preparation, i.e., coating it with a beam-sensitive layer known as resist, and in the case of electron beam exposure with a thin conductive layer to prevent charging. Following the exposure the resist is developed chemically, and in this process either the exposed or nonexposed parts of the resist are dissolved (depending on whether the resist is of positive or negative type), leading to a surface-relief profile. Several alternative fabrication steps may follow the development, depending on the desired grating type. If a high-contrast resist is used, a slightly 'undercut' profile as illustrated in Fig. 4(c) is obtained. This can be coated with a metal layer and the resist can be 'lifted off' chemically. Thus a binary metallic grating is formed. To produce a binary dielectric profile one bombards the substrate with ions either purely mechanically (ion beam milling) or with chemical assistance (reactive ion etching), and finally the residual metal is dissolved.

Alternatively, the particle beam dose can be varied with position to obtain an analog profile in the resist after development. This resist profile can be transformed into the substrate by proportional etching, which leads to an analog dielectric surface-relief structure (see Fig. 4(f)). This profile can be made reflective by depositing a thin metal film on top of it. If large-scale fabrication of identical gratings is the goal, a thick (~ 100 μm) metal layer (for example nickel) can be grown on the surface by electroplating. After separation from the substrate, the electroplated metal film, which is a negative of the original profile, can be used as a stamp in replicating the original profile in plastics using methods such as hot embossing, ultraviolet radiation curing, and injection molding. In particular, the latter method provides the key to high-volume and low-cost production of various types of diffraction gratings and other grating-based diffractive structures in the industrial scale.

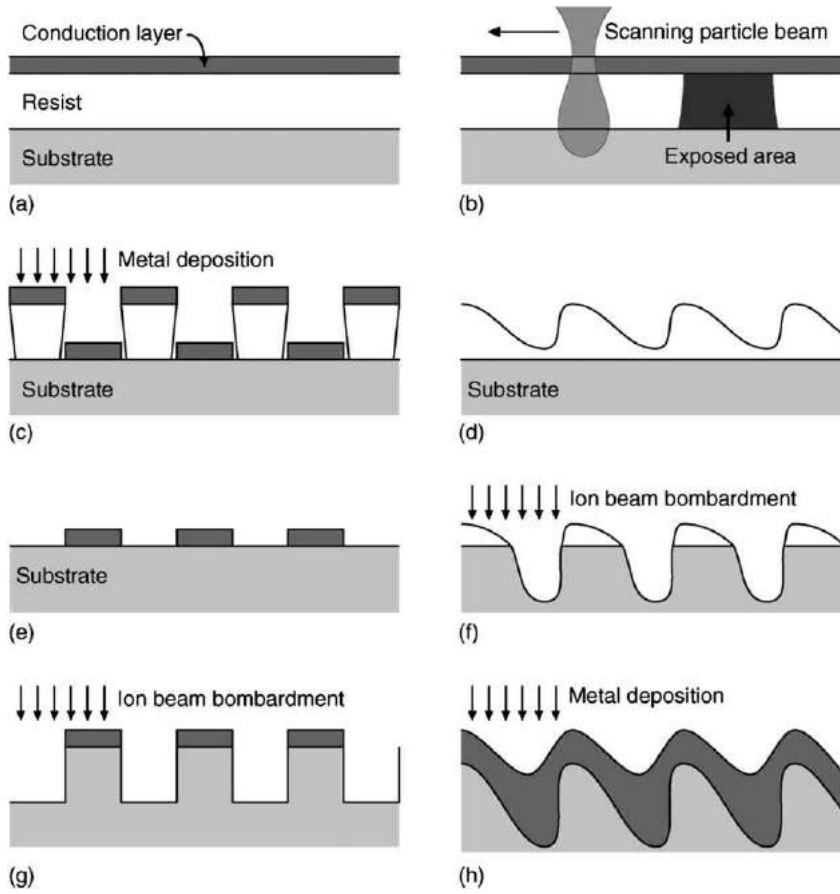


Fig. 4 Lithographic processes for grating fabrication. (a) Substrate after preparation for electron beam exposure; (b) exposure by a scanning electron beam; (c) slightly undercut profile in hight-contrast resist after development, under metal film evaporation; (d) analog resist profile after variable-dose exposure and resist development; (e) finished binary metallic grating after resist lift-off; (f) transfer of the analog surface-relief profile into the substrate by proportional etching; (g) etching of a binary profile into the substrate with a metal mask; (h) growth of a metal film on an analog surface profile.

Some Grating Applications

We now proceed to describe a collection of representative grating applications. More exhaustive coverage of grating applications can be found in the Further Reading section at the end of this article.

Spectroscopic Gratings

The purpose of a spectroscopic grating is to divide the incident light into a spectrum, i.e., to direct different frequency components of the incident beam into different directions given by the grating (**Eq. (1)**). When the grating period is reduced, the angular dispersion $\partial\theta_m/\partial\lambda$, and therefore the resolution of the spectrograph, increases.

In the numerical examples presented in **Fig. 5** we assume a reflection grating in a Littrow mount, i.e., $\theta_m^r = -\theta$. This mount is common in spectroscopy; it obviously requires that the grating is rotated when the spectrum is scanned. We see from **Eq. (1)** that in the Littrow mount:

$$\sin\theta = \frac{m\lambda}{2d} \tag{9}$$

By inserting this result back into **Eq. (1)** and differentiating we obtain a quantitative measure for dispersion in the Littrow mount:

$$\frac{\partial\theta_m}{\partial\lambda} = -\frac{m}{2d\cos\theta} \tag{10}$$

We consider TE polarization (electric field is linearly polarized in the y-direction) and order $m=-1$, optimize H to give maximum efficiency at $\lambda_0 = 550$ nm, and plot the diffraction efficiency η_{-1} over the visible wavelength range for two values of the grating period, $d=\lambda_0$ and $d=2\lambda_0$. Rigorous diffraction theory is used with $n_I=n_2=1$, and the metal is assumed to be aluminum (n_1 and n_{III} are complex-valued).

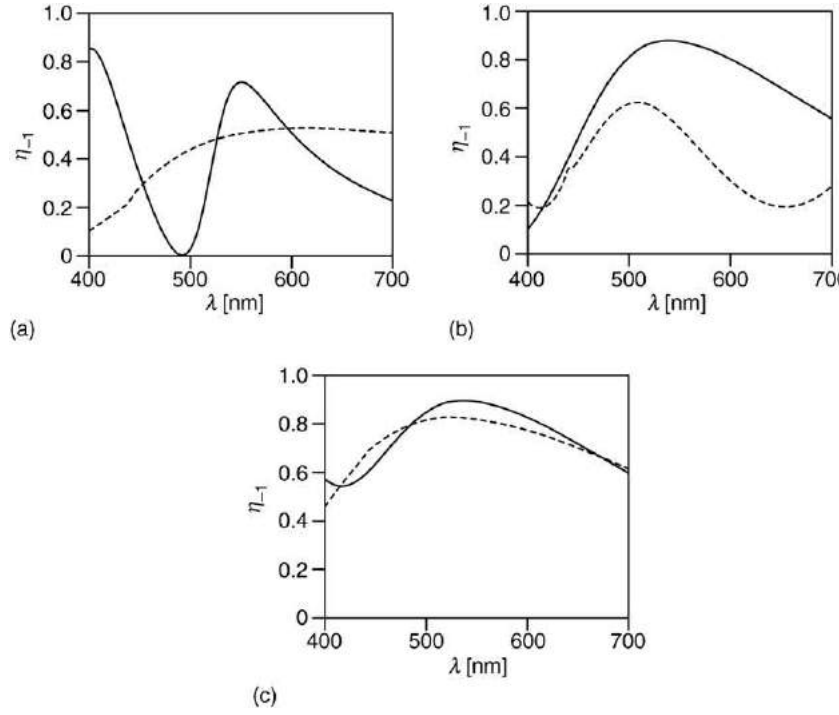


Fig. 5 Spectral diffraction efficiency curves of spectroscopic gratings in TE polarization. (a) Binary; (b) sinusoidal; and (c) triangular reflection gratings illuminated at Bragg angle. Solid lines: $d=\lambda_0$. Dashed lines: $d=2\lambda_0$.

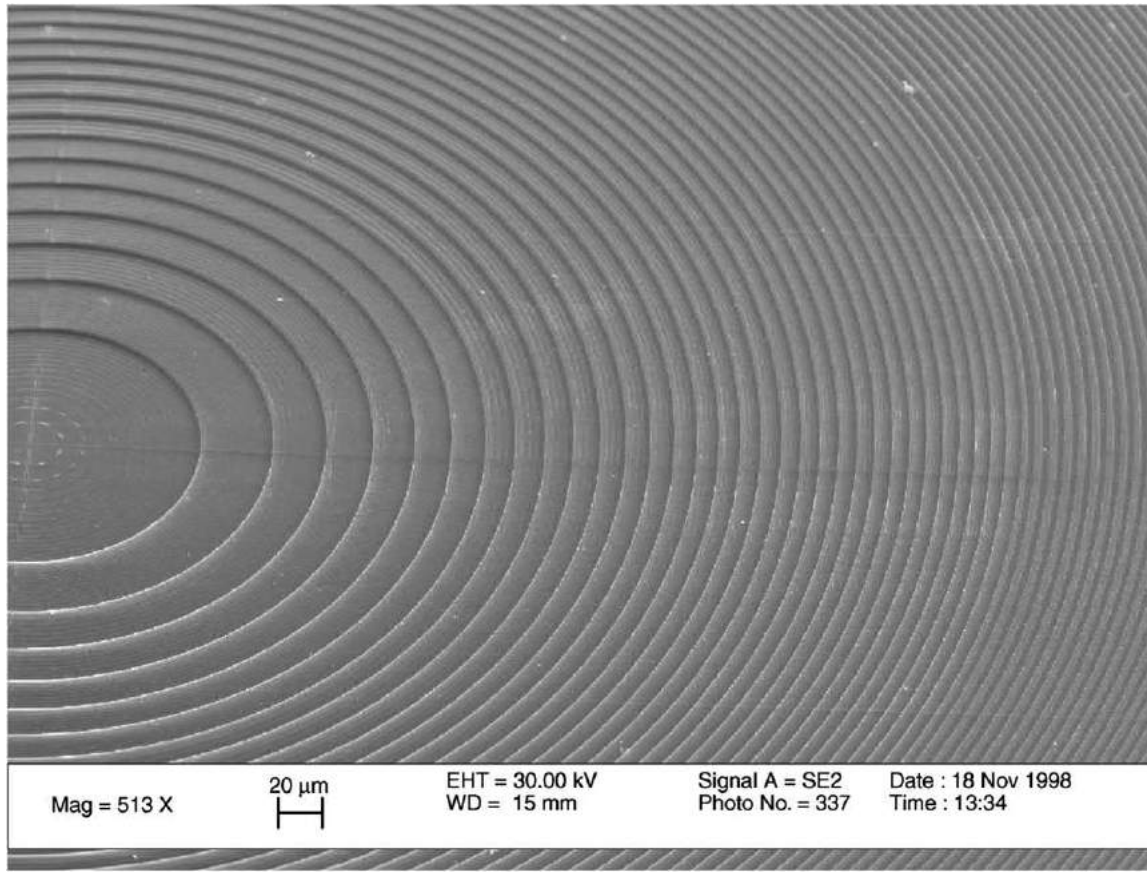


Fig. 6 A scanning electron micrograph of a part of a diffractive lens.

In Fig. 5(a) binary gratings with one groove (width $d/2$) per period is considered. Parametric optimization gives $H=422$ nm if $d=1\lambda$, and $H=168$ nm if $d=2\lambda$. Fig. 5(b) illustrates the results for sinusoidal gratings with $H=345$ nm if $d=1\lambda$, and $H=520$ nm if $d=2\lambda$. Finally, triangular gratings with $H=431$ nm if $d=1\lambda$, and $H=316$ nm if $d=2\lambda$, are considered. The triangular grating profile is seen to be the best choice in general.

Diffractive Lenses

Diffractive lenses are gratings with locally varying period and groove orientation. Fig. 6 illustrates the central part of a diffractive microlens fabricated by lithographic methods. It is seen to consist of concentric zones with radially decreasing width.

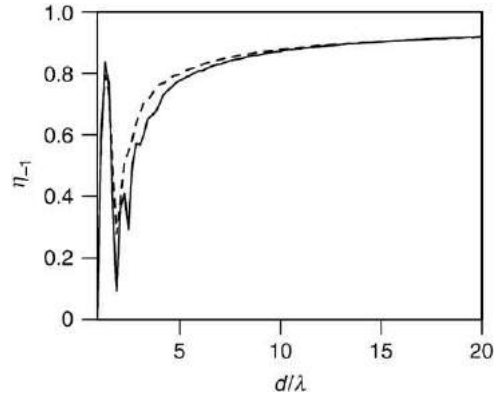


Fig. 7 Diffraction efficiency versus wavelength for triangular transmission gratings illuminated at normal incidence. Solid lines: TE polarization. Dashed lines: TM polarization.

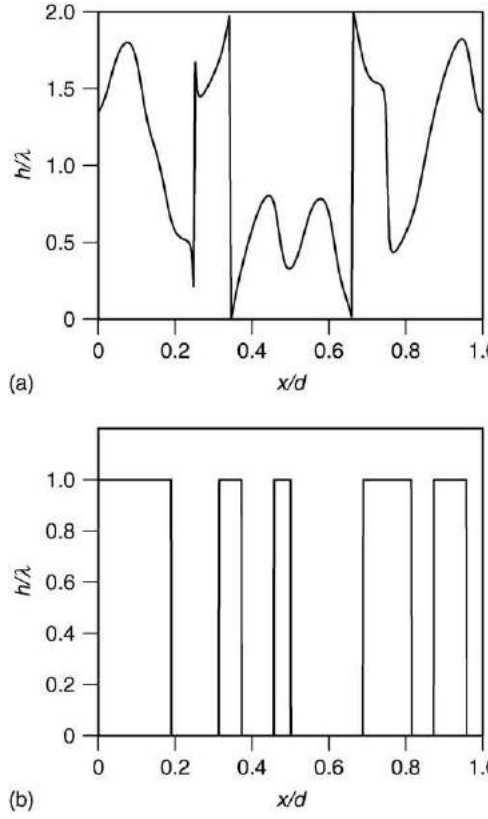


Fig. 8 Surface-relief profiles of (a) continuous and (b) binary 1→8 beamsplitter gratings.

The phase transmission function of an ideal diffractive focusing lens can be obtained by a simple optical path calculation, and is given by

$$\phi(r) = \frac{2\pi}{\lambda_0}\left(f - \sqrt{f^2 + r^2}\right) \approx -\frac{\pi}{\lambda_0 f} r^2 \tag{11}$$

where f is the focal length, r is the radial distance from the optical axis, and the approximate expression is valid in the paraxial region $r \ll f$ Here λ_0 is the design wavelength, which determines the modulation depth of the locally blazed grating profile:

$$H = \frac{\lambda_0}{n(\lambda_0) - 1} \tag{12}$$

and $n(\lambda_0)$ is the refractive index of the dielectric material at $\lambda = \lambda_0$.

The zone boundaries r_n are defined by the condition $\phi(r_n) = -2\pi n$, and therefore we obtain from **Eq. (11)**:

$$r_n = \sqrt{2nf\lambda_0 + (n\lambda_0)^2} \approx \sqrt{2nf\lambda_0} \tag{13}$$

Obviously, when the numerical aperture of the lens is large, the zone width in the outer regions of the lens is reduced to values of the order of λ. **Fig. 7** illustrates the local diffraction efficiency in different part of a diffractive lens: we use the order $m = -1$ of a locally triangular diffraction grating of the type shown in **Fig. 1(c)** with $\theta = 0$, assuming that $n_I = n_2 = 1.5$ (fused silica) and $n_1 = n_{III} = 1$. The efficiency η_{-1} is calculated as a function of λ using rigorous diffraction theory for both TE and TM polarization (the electric vector points in the groove direction in the TE case and is perpendicular to it in the TM case). It is seen that the local efficiency of the lens becomes poor in regions where the local grating period is small, and thus diffractive lenses operate best in the paraxial region (it is, however, possible to obtain improved non-paraxial performance by optimizing the local grating profile).

A factor that limits the use of diffractive lenses with polychromatic light is their large chromatic aberration: in the paraxial limit the focal length depends on the wavelength according to the formula:

$$f(\lambda) = \frac{\lambda_0}{\lambda} f \tag{14}$$

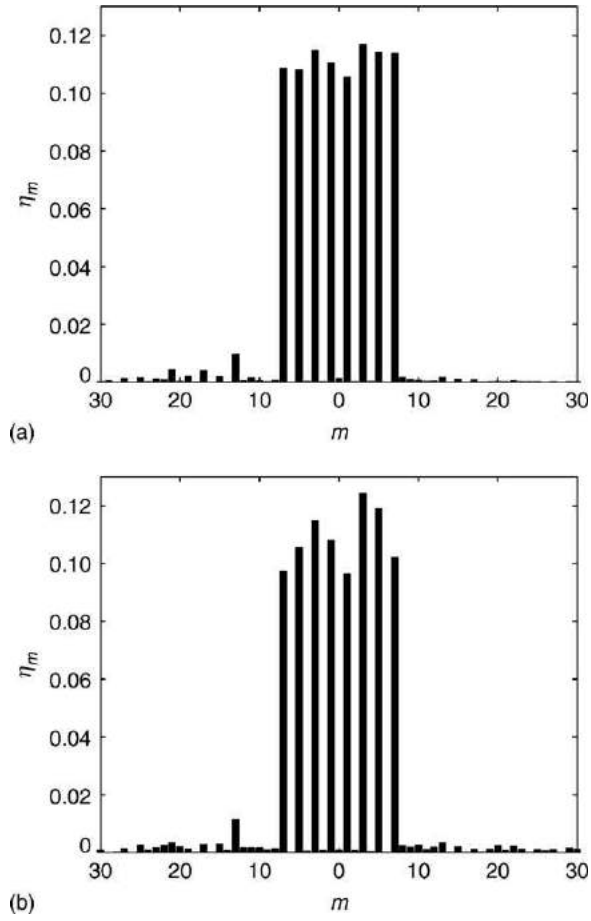


Fig. 9 Diffraction pattern produced by a continuous-profile beamsplitter grating when (a) $d = 200\lambda$ and (b) $d = 50\lambda$.

However, since the dispersion of a grating has an opposite sign compared to the diffraction of a prism, a weak positive diffractive lens can be combined with a relatively strong positive refractive lens to form a single-element achromatic lens. Such a hybrid lens is substantially thinner and lighter than a conventional all-refractive achromatic lens made of crown and flint glasses.

Beam Splitter Gratings

Multiple beamsplitters, also known as array illuminators, are gratings with sophisticated periodic structure that are capable of transforming an incident plane wave into a set of diffraction orders with a specified distribution of diffraction efficiencies. Most often one wishes to obtain a one- or two-dimensional array of diffracted plane waves with equal efficiency. This goal can be achieved by appropriate design of the grating profile, which may be of binary, multilevel, or continuous form.

Fig. 8 illustrates binary and continuous grating profiles capable of dividing one incident plane wave into eight diffracted plane waves (central odd-numbered orders of the grating) with equal efficiency in view of TEA. Fig. 9 shows the distribution of energy among the different diffraction orders in and around the array produced by the continuous profile, evaluated by rigorous theory for two different grating periods, while Fig. 10 gives corresponding results for the binary grating.

The fraction of incident energy that ends up in the desired eight orders is 72% for the binary grating and 96% for the continuous grating according to TEA, making the latter more attractive from a theoretical viewpoint. However, the continuous profile is more difficult to fabricate accurately than the binary one, especially for small values of the ratio d/λ. Moreover, the array uniformity produced by the continuous profile degrades faster than that of the binary profile when the period is reduced.

An example of the structure of a biperiodic grating that is capable of forming a large, shaped array of diffraction orders with equal efficiency is presented in Fig. 11. Fig. 12 shows the far-field diffraction pattern produced by this element: the shape of Finland with the location of the town of Joensuu highlighted by placing the zero diffraction order on the appropriate position on the map (the apparent nonuniformity in the array is mostly due to the finite resolution of the CCD detector).

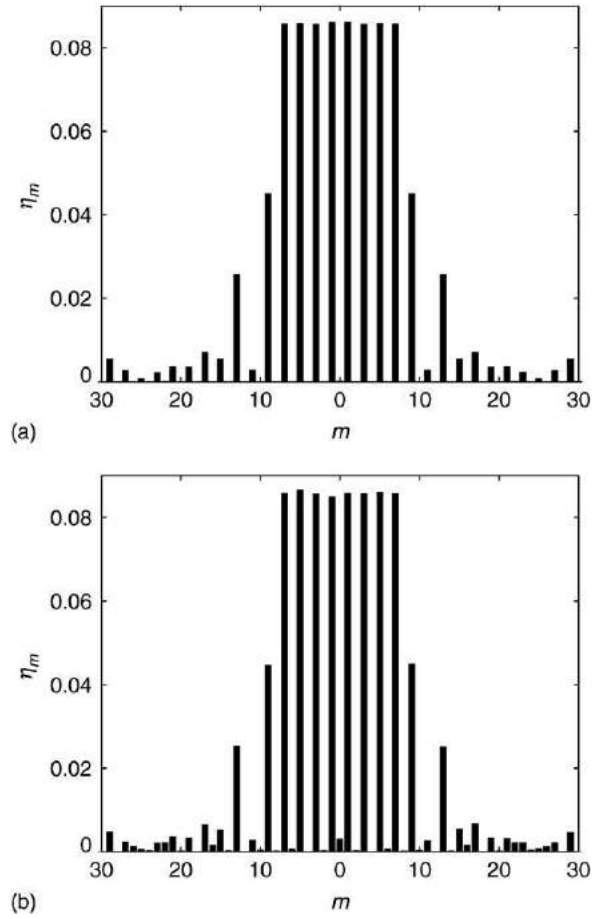


Fig. 10 Diffraction pattern produced by a binary beamsplitter grating grating when (a) $d = 200\lambda$ and (b) $d = 50\lambda$.

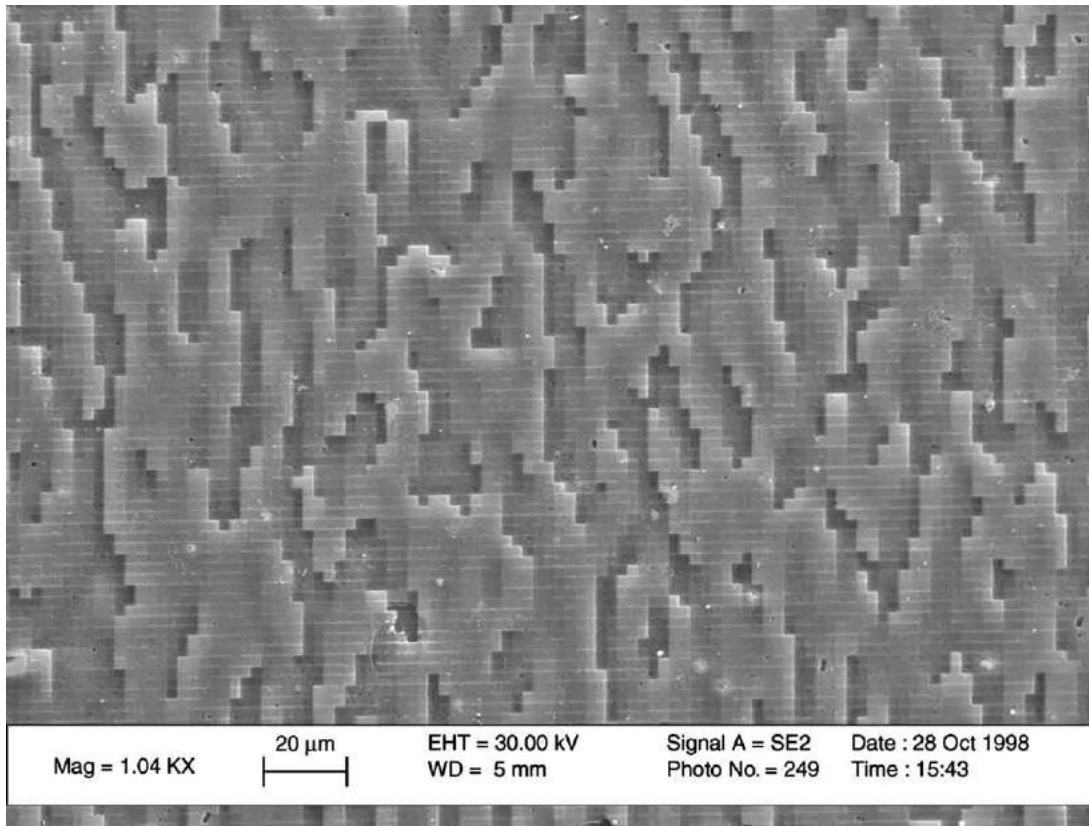


Fig. 11 Scanning electron micrograph of a sophisticated beamsplitter grating.

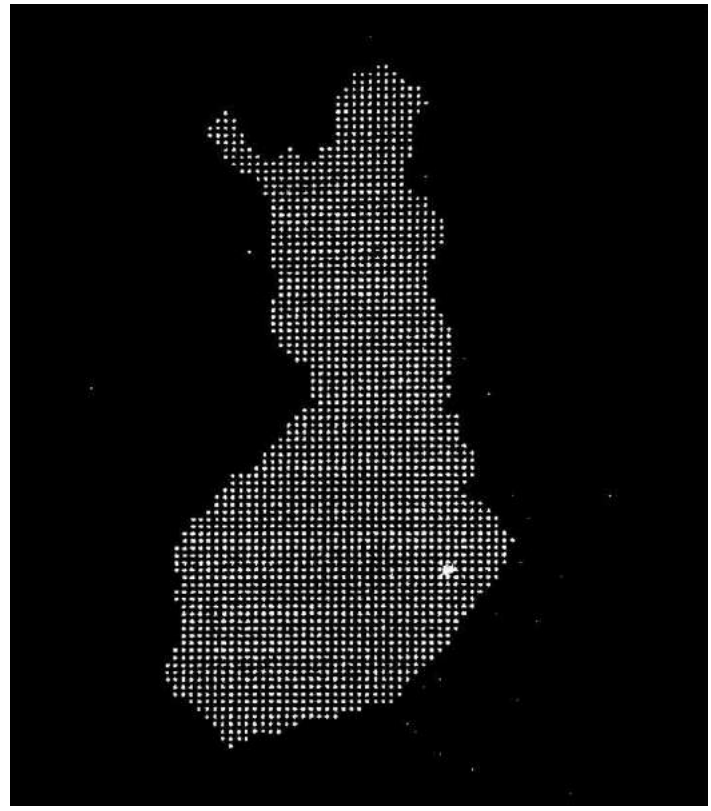


Fig. 12 The far-field diffraction pattern produced by the grating in Fig. 11.

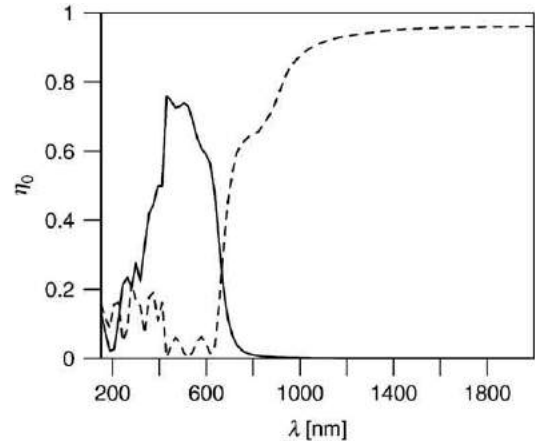


Fig. 13 Zero order spectral transmission (solid line) and reflection (dashed line) of an aluminium inductive grid filter with 400 nm period.

Inductive Grid Filters

As a final example we consider a scaled-down version of a device well known from the door of a microwave oven: one can see in through the holes pierced in the metallic screen in the oven door, but the microwaves are trapped inside because their wavelengths are large in comparison with the dimensions of the holes. Another example of these inductive grid filters is found in large radio telescope mirrors for centimeter-scale wavelengths, which are wide-grid structures rather than smoothly curved metal surfaces.

Fig. 13 illustrates the spectral transmission and reflection curves of an inductive grid filter of the type illustrated in Fig. 3(b). We assume that $d_x = d_y = d = 400$ nm, $c_x = c_y = c = 310$ nm, $H = 500$ nm, and that the grid is a self-supporting structure ($n_{\mathrm{I}} = n_{\mathrm{III}} = 1$) made of aluminum. Only the zero reflected and transmitted orders propagate when $\lambda > d$, but other orders (with rather low efficiency) appear for smaller wavelengths. We see that the filter reflects infrared radiation effectively, but transmits a substantial part of radiation at visible and shorter wavelength regions.

See also: Fraunhofer Diffraction

Further Reading

Chandezon, J., Maystre, D., Raoult, G., 1980. A new theoretical method for diffraction gratings and its numerical application. Journal of Optics 11, 235–241.
Gaylord, T.K., Moharam, M.G., 1985. Analysis and applications of optical diffraction by gratings. Proceedings of IEEE 73, 894–937.
Herzig, H.P., 1997. Micro-optics: Elements, Systems and Applications. London: Taylor & Francis.
Hutley, M.C., 1982. Diffraction Gratings. London: Academic Press.
Li, L., 2001. Mathematical reflections on the fourier modal method in grating theory. In: Bao, G., Cowsar, L., Masters, W. (Eds.), Mathematical Modelling in Optical Science. Philadelphia: SIAM, pp. 111–139.
Li, L., Chandezon, J., Granet, G., Plumey, J.-P., 1999. Rigorous and efficient grating-analysis method made easy for optical engineers. Applied Optics 38, 304–313.
Petit, R. (Ed.), 1980. Electromagnetic Theory of Gratings. Berlin: Springer-Verlag.
Rayleigh, J.W.S., 1907. On the dynamical theory of gratings. Proceeding of the Royal Society A 79, 399.
Solymar, L., Cooke, D.J., 1981. Volume Holography and Volume Gratings. London: Academic Press.
Turunen, J., Kuittinen, M., Wyrowski, F., 2000. Diffractive optics: electromagnetic approach. In: Wolf, E. (Ed.), Progress in Optics XL. Amsterdam: Elsevier.
Turunen, J., Wyrowski, F. (Eds.), 2000. Industrial and Commercial Applications of Diffractive Optics. Berlin: Akademie Verlag.
van Renesse, R.L., 1993. Optical Document Security. Boston, MA: Artech House.

Fraunhofer Diffraction

BD Guenther, Duke University, Durham, NC, USA

Introduction

Geometrical optics does a reasonable job of describing the propagation of light in free space unless that light encounters an obstacle. Geometrical optics predicts that the obstacle would cast a sharp shadow. On one side of the shadow's edge would be a bright uniform light distribution, due to the incident light, and on the other side there would be darkness. Close examination of an actual shadow edge reveals, however, dark fringes in the bright region and bright fringes in the dark region. This departure from the predictions of geometrical optics was given the name diffraction by Grimaldi.

There is no substantive difference between diffraction and interference. The separation between the two subjects is historical in origin and is retained for pedagogical reasons. Since diffraction and interference are actually the same physical process, we expect the observation of diffraction to be a strong function of the coherence of the illumination. With incoherent sources, geometrical optics is usually all that is needed to predict the performance of an optical system and diffraction can be ignored except at dimensions on the scale of a wavelength. With light sources having a large degree of spatial coherence, it is impossible to neglect diffraction in any analysis.

Even though diffraction produced by most common light sources and objects is a small effect, it is possible to observe diffraction without special equipment. By viewing a light source through a slit formed by two fingers that are nearly touching, fringes can be observed. Diffraction is used by photographers and television cameramen to provide artistic highlights in an image containing small light sources. They accomplish this by placing a screen in front of the camera lens. Light from point sources in the field of view is diffracted by the screen and produces 'stars' in the image. A third place where diffraction effects are easily observed is when a mercury or sodium streetlight, a few hundred meters away, is viewed through a sheer curtain.

Fresnel originally developed an approximate scalar theory based on Huygens' principle which states that:

> Each point on a wavefront can be treated as a source of a spherical wavelet called a secondary wavelet or a Huygens' wavelet. The envelope of these wavelets, at some later time, is constructed by finding the tangent to the wavelets. The envelope is assumed to be the new position of the wavefront.

Rather than simply using the wavelets to construct an envelope, Fresnel's scalar theory assumes that the Huygens' wavelets interfere to produce a new wavefront.

A rigorous diffraction theory is based on Maxwell's equations and uses the boundary conditions associated with the obstacle to calculate a field scattered by the obstacle. The origins of this scattered field are currents induced in the obstacle by the incident field. The scattered field is allowed to interfere with the incident field to produce a resultant diffracted field. The application of the rigorous theory is very difficult and for most problems an approximate scalar theory developed by Kirchhoff and Sommerfeld is used.

Kirchhoff Theory of Diffraction

Kirchhoff's theory was based on the elastic theory of light but can be reformulated into a vector theory. Here we will limit our discussion to the scalar formulation. Given an incident wave, φ, we wish to calculate the optical wave at point P_0, in **Fig. 1**, located at $\mathbf{r}_0$, in terms of the wave's value on a surface, S, that we construct about the observation point.

To obtain a solution to the wave equation at the point, P_0, we select as a Green's function the one proposed by Kirchhoff (Box 1), a unit amplitude spherical wave, expanding about the point at $\mathbf{r}_0$, denoted by Ψ, and then apply Green's theorem (see Box 2). We will run into boundary condition problems as we proceed with the derivation of the solution. Sommerfeld removed the problems by assuming a slightly more complex Green's function (see Box 1). We will limit our discussion to the Kirchhoff Green's function.

Green's theorem requires that there be no sources (singularities) inside the surface S but our Green's function [1] is based on a source at r_0. We eliminate the source by constructing a small spherical surface S_ε, of radius ε, about $\mathbf{r}_0$, excluding the singularity at $\mathbf{r}_0$ from the volume of interest, shown in gray in **Fig. 1**. The surface integration that must be performed in Green's theorem is over the surface $S' = S + S_\varepsilon$.

Within the volume enclosed by S′, the Green's function, Ψ, and the incident wave, φ, satisfy the scalar Helmholtz equation so that the volume integral can be written as

$$\int\int_V\int (\Psi\nabla^2\varphi - \varphi\nabla^2\Psi)\mathrm{d}v = -\int\int_V\int (\Psi\varphi k^2 - \varphi\Psi k^2)\mathrm{d}v \tag{7}$$

The right side of (7) is identically equal to zero. This fact allows us to use [6] to produce a simplified statement of Green's theorem:

$$\int\int_{S'}\left(\Psi\frac{\partial\varphi}{\partial n} - \varphi\frac{\partial\Psi}{\partial n}\right)\mathrm{d}s = 0 \tag{8}$$

 doi:10.1016/B978-0-12-809283-5.00695-9

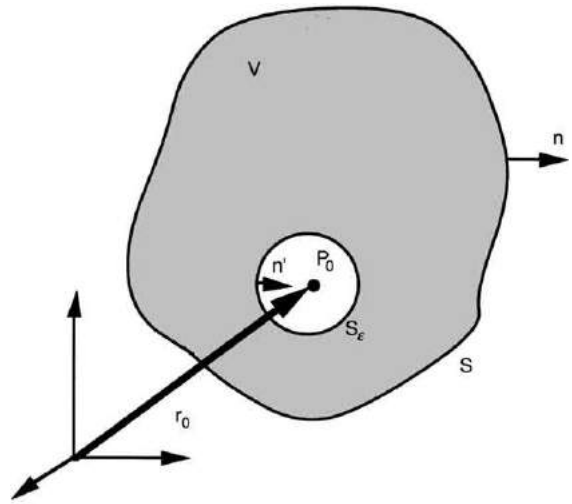


Fig. 1 Region of integration for solution of Kirchhoff's diffraction integral. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Box 1 Green's Function

1. **Kirchhoff Green's function**: Assume that the wave, Ψ, is due to a single point source at $\mathbf{r_0}$. At an observation point, $\mathbf{r}_1$, the Green's function is given by

$$\Psi(\mathbf{r}_1) = \frac{e^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}, \tag{1}$$

   where $r_{01} = |\mathbf{r}_{01}| = |\mathbf{r}_1 - \mathbf{r}_0|$ is the distance from the source of the Green's function, at $\mathbf{r}_0$, to the observation position, at $\mathbf{r}_1$.
2. **Sommerfeld Green's function**: Assume that there are two point sources, one at P_0 and one at P'_0 (see Fig. 2 where the source, P'_0, is positioned to be the mirror image of the source at P_0 on the opposite side of the screen, thus

$$r_{01} = r'_{01} \quad \cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) = -\cos(\hat{\mathbf{n}}, \mathbf{r}'_{01}) \tag{2}$$

   A Green's function that could be used is

$$\Psi' = \frac{e^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}} + \frac{e^{-i\mathbf{k}\cdot\mathbf{r}'_{01}}}{r'_{01}} \tag{3a}$$

   or

$$\Psi' = \frac{e^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}} + \frac{e^{-i\mathbf{k}\cdot\mathbf{r}'_{01}}}{r'_{01}} \tag{3b}$$

Box 2 Green's Theorem

To calculate the complex amplitude, $\tilde{E}(\mathbf{r})$ of a wave at an observation point defined by the vector $\mathbf{r}$, we need to use a mathematical relation known as Green's Theorem. Green's theorem states that, if φ and Ψ are two scalar functions that are well behaved, then

$$\int_S \int (\varphi \nabla\Psi \cdot \hat{\mathbf{n}} - \Psi \nabla\varphi \cdot \hat{\mathbf{n}}) ds = \int_V \int \int (\varphi \nabla^2\Psi - \Psi \nabla^2\varphi) dv \tag{4}$$

The vector identities

$$\nabla\Psi \cdot \hat{\mathbf{n}} = \frac{\partial\Psi}{\partial n} \quad \nabla\varphi \cdot \hat{\mathbf{n}} = \frac{\partial\varphi}{\partial n} \tag{5}$$

allow Green's theorem to be written as

$$\int_S \int \left[\varphi \frac{\partial\Psi}{\partial n} - \Psi \frac{\partial\varphi}{\partial n}\right] ds = \int_V \int \int [\varphi \nabla^2\Psi - \Psi \nabla^2\varphi] dv \tag{6}$$

This equation is the prime foundation of scalar diffraction theory but only the proper choice of Ψ, φ and the surface, S, allows direct application to the diffraction problem.

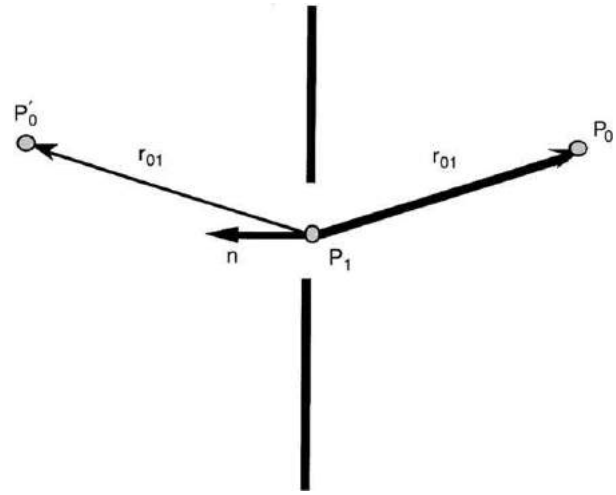


Fig. 2 Geometry for the Sommerfeld Green's function. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Because the integral over the combined surfaces, S and S_ε, is equal to zero, the integral over the surface S must equal the negative of the integral over the surface, S_ε:

$$-\int\int_{S_\varepsilon}\left(\Psi\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial\Psi}{\partial n}\right)\mathrm{d}s=\int\int_{S}\left(\Psi\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial\Psi}{\partial n}\right)\mathrm{d}s \tag{9}$$

To perform the surface integrals, we must evaluate the Green's function on the surface, S′. We therefore select $\mathbf{r}_1$, the vector defining the observation point, to be on either of the two surfaces that make up S′. The derivative, $\partial\Psi/\partial n$, to be evaluated is

$$\frac{\partial\Psi}{\partial n}=\frac{\partial\Psi}{\partial r}\frac{\partial r}{\partial n}=\left(-ik\frac{\mathrm{e}^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}-\frac{\mathrm{e}^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}^2}\right)\cos\left(\mathrm{n},\mathbf{r}_{01}\right) \tag{10}$$

where $\cos\left(\hat{\mathbf{n}},\mathbf{r}_{01}\right)$ is the cosine of the angle between the outward normal $\hat{\mathbf{n}}$ and the vector $\mathbf{r}_{01}$, the vector between points P_0 and P_1. (Note from **Fig. 1** that the outward normal on S_ε is inward, toward P_0, while the normal on S is outward, away from P_0.)

For example if $\mathbf{r}_1$ were on S_ε, then

$$\cos\left(\mathbf{n},\mathbf{r}_{01}\right)=-1, \tag{11}$$

$$\frac{\partial\Psi(\mathbf{r}_1)}{\partial n}=\left[\frac{1}{\varepsilon}+ik\right]\frac{\mathrm{e}^{-ik\varepsilon}}{\varepsilon}, \tag{12}$$

where ε is the radius of the small sphere we constructed around the point at P_0.

We now use **(10)** to rewrite the two integrals in **(9)**. The integral over the surface S_ε is

$$\int\int_{S_\varepsilon}\left(\Psi\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial\Psi}{\partial n}\right)\mathrm{d}s=\int\int_{S_\varepsilon}\left[\frac{\partial\varphi}{\partial n}\frac{\mathrm{e}^{-ik\varepsilon}}{\varepsilon}-\varphi\frac{\mathrm{e}^{-ik\varepsilon}}{\varepsilon}\left(\frac{1}{\varepsilon}+ik\right)\right]\times\varepsilon^2\sin\theta\mathrm{d}\theta\mathrm{d}\phi \tag{13}$$

while the integral over the surface S is

$$\int\int_{S}\left[\frac{\mathrm{e}^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial}{\partial n}\left(\frac{\mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}\right)\right]r^2\sin\ \theta\mathrm{d}\,\theta\mathrm{d}\phi \tag{14}$$

The omitted volume contained within the surface S_ε is allowed to shrink to zero by taking the limit as $\varepsilon\to0$. **Eq. (14)** will not be affected by taking the limit. The first and last terms of the right side of **(13)** go to zero as $\varepsilon\to0$ because they contain $\varepsilon\mathrm{e}^{-ik\varepsilon}$. The second term in **(13)** contains $\mathrm{e}^{-ik\varepsilon}$ which goes to 1 as $\varepsilon\to0$. Therefore, in the limit as $\varepsilon\to0$, **(9)** becomes

$$\varphi(\mathbf{r}_0)=\frac{1}{4\pi}\int\int_{S}\left[\frac{\mathrm{e}^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial}{\partial n}\left(\frac{\mathrm{e}^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}\right)\right]\mathrm{d}s \tag{15}$$

which is sometimes called the integral theorem of Kirchhoff. By carefully selecting the surface of integration in **(15)**, we can calculate the diffraction field observed at P_0 produced by an aperture in an infinite opaque screen.

Assume that a source at P_2 in **Fig. 3** produces a spherical wave that is incident on an infinite, opaque screen from the left. To find the field at P_0 in **Fig. 3**, we apply **(15)** to the surface $S_1+S_2+\Sigma$, where S_1 is a plane surface adjacent to the screen, Σ is that portion of S_1 in the aperture and S_2 is a large spherical surface, of radius R, centered on P_0, see **Fig. 3**. The first question to address is

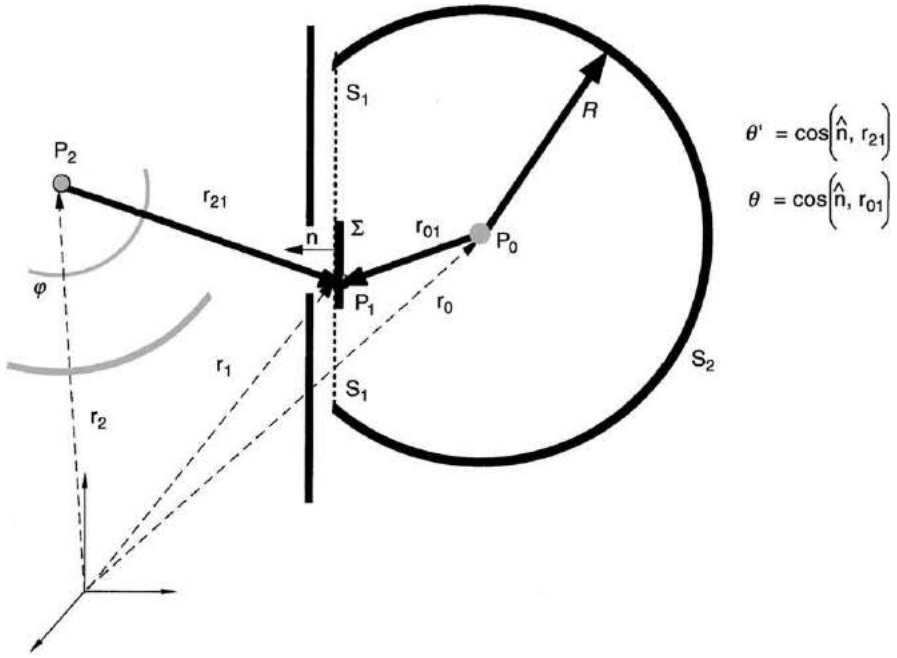


Fig. 3 Integration surface for diffraction calculation. P_2 is the source and P_0 is the point where the field is to be calculated. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

how to calculate the contribution from S_2 or better yet how do we show we can ignore contributions from S_2. As R increases, Ψ and ϕ will decrease as $1/R$. However, the area of integration increases as R^2 so the $1/R$ fall-off is not a sufficient reason for neglecting the contribution of the integration over S_2.

On the surface S_2, the Green's function and its derivative are

$$\Psi = \frac{e^{-ikR}}{R} \tag{16}$$

$$\frac{\partial \Psi}{\partial n}\left(-ik - \frac{1}{R}\right)\frac{e^{-ikR}}{R} \approx -ik\Psi \tag{17}$$

where the approximate equality is for large R. Thus, the integral over S_2 for large R is

$$\int\int_{S_2}\left[\Psi\frac{\partial\varphi}{\partial n} + \varphi(ik\Psi)\right]ds \quad = \int\int_{S_2}\Psi\left(\frac{\partial\varphi}{\partial n} + ik\varphi\right)R^2 \sin\,\theta d\theta\, d\phi \tag{18}$$

The quantity $R\Psi = e^{-ik\cdot R}$ is finite as $R\to\infty$ so for the integral to vanish, we must have

$$\lim_{R\to\infty} R\left(\frac{\partial\varphi}{\partial n} + ik\varphi\right) = 0 \tag{19}$$

This requirement is called the Sommerfeld radiation condition and is satisfied if $\varphi\to 0$ as fast as $1/R$ (for example, a spherical wave). Since the illumination of the screen will be a spherical wave, or at least a linear combination of spherical waves, we should expect the integral over S_2 to make no contribution (it is exactly zero).

Another way to insure that surface S_2 makes no contribution is to assume that the light turns on at time t_0. At a later time, t, when we desire to know the field at P_0, the radius of S_2 is $R > c\,(t - t_0$ which physically means that the light has not had time to reach S_2. In this situation, there can be no contribution from S_2. This is not a perfect solution because when the wave is of finite duration it can no longer be considered monochromatic.

We should expect the major contribution in the integral over S_1 to come from the portion of the surface in the aperture, Σ. We make the following assumptions about the incident wave, ϕ (these assumptions are known as St. Venaut's hypothesis or Kirchhoff's boundary conditions)

1. We assume that, in the aperture, ϕ and $\partial\phi/\partial n$ have the values they would have if the screen were not in place.
2. On the portion of S_1 not in the aperture, ϕ and $\partial\phi/\partial n$ are identically zero.

The Kirchhoff boundary conditions allow the screen to be neglected, reducing the problem to an integration only over the aperture. It is surprising that the assumptions about the contribution of the surface S_2 and the screen yield very accurate results

(as long as polarization is not important, the aperture is large with respect to the wavelength, and we don't get too close to the aperture).

Mathematically the Kirchhoff boundary conditions are incorrect. The two Kirchhoff boundary conditions imply that the field is zero everywhere behind the screen, except in the aperture, which makes Ψ and $\partial\psi/\partial n$ discontinuous on the boundary of the aperture. Another problem with the boundary conditions is that Ψ and $\partial\Psi/\partial n$ are known only if the problem has already been solved.

The Sommerfeld Green's functions will remove the inconsistencies of the Kirchhoff theory. If we use the Green's function [3a] then Ψ vanishes in the aperture while $\partial\Psi/\partial n$ can assume the value required by Kirchhoff's boundary conditions. If we use the Green's functions [3b] then the converse holds, i.e. $\partial\Psi/\partial n$ is zero in the aperture. This improved Green's function makes the problem harder with little gain in accuracy so we will retain the Kirchhoff formalism.

As a result of the above discussion, the surface integral **(15)** is only performed over the surface, Σ, in the aperture. The evaluation of the Green's function on this surface can be simplified by noting that normally $r_{01} \gg \lambda$, i.e., $k \gg 1/r_{01}$, thus on $\sum$,

$$\frac{\partial \Psi}{\partial n} = \cos(\mathbf{n}, \mathbf{r}_{01})\left(-ik - \frac{1}{r_{01}}\right)\frac{e^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}} \approx -ik\cos(\hat{\mathbf{n}}, \mathbf{r}_{01})\frac{e^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}} \tag{20}$$

Substituting this approximate evaluation of the derivative into **(15)** yields

$$\varphi(\mathbf{r}_0) = \frac{1}{4\pi}\int\int_{\sum} \frac{e^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}\left[\frac{\partial\varphi}{\partial n} + ik\varphi\cos(\hat{\mathbf{n}}, \mathbf{r}_{01})\right]ds \tag{21}$$

The source of the incident wave is a point source located at P_2, with a position vector, $\mathbf{r}_2$, measured with respect to the coordinate system and a distance $|\mathbf{r}_{21}|$ away from P_1, a point in the aperture (see **Fig. 3**). The incident wave is therefore a spherical wave of the form

$$\varphi(\mathbf{r}_{21}) = A\frac{e^{-i\mathbf{k}\cdot\mathbf{r}_{21}}}{r_{21}} \tag{22}$$

which fills the aperture. Here also we will assume that $r_{21} \gg \lambda$ so that the derivative of the incident wave assumes the same form as **(20)**. Then **(21)** can be written

$$\varphi(\mathbf{r}_0) = \frac{iA}{\lambda}\int\int_{\sum} \frac{e^{-i\mathbf{k}\cdot(\mathbf{r}_{21}+\mathbf{r}_{01})}}{r_{21}r_{01}} \times \left[\frac{\cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) - \cos(\hat{\mathbf{n}}, \mathbf{r}_{21})}{2}\right]ds \tag{23}$$

This relationship is called the Fresnel–Kirchhoff diffraction formula. It is symmetric with respect to r_{01} and r_{21}, making the problem identical when the source and measurement point are interchanged.

A physical understanding of **(23)** can be obtained if it is rewritten as

$$\varphi(\mathbf{r}_0) = \int\int_{\sum} \Phi(\mathbf{r}_{21})\frac{e^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}ds \tag{24}$$

The field at P_0 is due to the sum of an infinite number of secondary Huygens sources in the aperture Σ. The secondary sources are point sources radiating spherical waves of the form

$$\Phi(\mathbf{r}_{21})\frac{e^{-i\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}} \tag{25}$$

with amplitude $\Phi(r_{21})$, defined by

$$\Phi(\mathbf{r}_{21})\frac{i}{\lambda}\left[A\frac{e^{-i\mathbf{k}\cdot\mathbf{r}_{21}}}{r_{21}}\right]\left[\frac{\cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) - \cos(\hat{\mathbf{n}}, \mathbf{r}_{21})}{2}\right] \tag{26}$$

The imaginary constant, i, causes the wavelets from each of these secondary sources to be phase shifted with respect to the incident wave. The obliquity factor, in the amplitude **(26)**

$$\frac{1}{2}[\cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) - \cos(\hat{\mathbf{n}}, \mathbf{r}_{21})] \tag{27}$$

causes the secondary sources to have a forward directed radiation pattern.

If we had used the Sommerfeld Green's function, the only modification to our result would be a change in the obliquity factor to $\cos(\mathbf{n},\mathbf{r}_{01})$. In our discussions below we will assume that the angles are all small, allowing the obliquity factor to be replaced by 1 so that in the remaining discussion the choice of Green's function has no impact.

We will assume the source of light is at infinity, $z_1 = \infty$ in **Fig. 3**, so that the aperture is illuminated by a plane wave, traveling parallel to the z-axis. With this assumption

$$\theta' = \cos(\mathbf{n}, \mathbf{r}_{21}) \approx -1 \tag{28}$$

We will also make the paraxial approximation which assumes that the viewing position is close to the z-axis, leading to

$$\theta = \cos(\mathbf{n}, \mathbf{r}_{01}) \approx 1 \tag{29}$$

With these assumptions Eq. (24) becomes identical to the Huygens–Fresnel integral discussed in the section on Fresnel diffraction:

$$\tilde{E}(\mathbf{r}_0) = \frac{\mathrm{i}}{\lambda}\int\int_{\sum}\frac{\tilde{E}_i(\mathbf{r})}{R}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}}\mathrm{d}s \tag{30}$$

The modern interpretation of the Fresnel–Kirchhoff (or now in the small-angle approximation, the Fresnel–Huygens) integral is to view it as a convolution integral. By considering free space to be a linear system, the result of propagation can be calculated by convolving the incident (input) wave with the impulse response of free space (our Green's function),

$$\frac{\mathrm{i}}{\lambda}\frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}}}{R} \tag{31}$$

The job of calculating diffraction has only begun with the derivation of the Fresnel–Kirchhoff integral. In general, an analytic expression for the integral cannot be found because of the difficulty of performing the integration over R. There are two approximations that will allow us to obtain analytic expressions of the Fresnel–Kirchhoff integral. In both approximations, all dimensions are assumed to be large with respect to the wavelength. In one approximation, the viewing position is assumed to be far from the obstruction; the resulting diffraction is called Fraunhofer diffraction and will be discussed here. The second approximation, which leads to Fresnel diffraction, assumes that the observation point is nearer the obstruction, to be quantified below.

Fraunhofer Approximation

In Fraunhofer diffraction, we require that the source of light and the observation point, P_0, be far from the aperture so that the incident and diffracted wave can be approximated by plane waves. A consequence of this requirement is that the entire waveform passing through the aperture contributes to the observed diffraction.

The geometry to be used in this derivation is shown in Fig. 4. The wave incident normal to the aperture, Σ, is a plane wave and the objective of the calculation is to find the departure of the transmitted wave from its geometrical optical path. The calculation will provide the light distribution, transmitted by the aperture, as a function of the angle the light is deflected from the incident direction. We assume that diffraction makes only a small perturbation on the predictions of geometrical optics. The deflection angles encountered in this derivation are, therefore, small, as assumed above, and we will be able to use the paraxial approximation.

The distance from a point P, in the aperture, to the observation point P_0, of Fig. 4, is

$$R^2 = (x-\xi)^2 + (y-\eta)^2 + Z^2 \tag{32}$$

From Fig. 4 we see R_0 is the distance from the center of the screen to the observation point, P_0,

$$R_0^2 = \xi^2 + \eta^2 + Z^2 \tag{33}$$

The difference between these two vectors is

$$R_0^2 - R^2 = (R_0 - R)(R_0 + R) = \xi^2 + \eta^2 + Z^2 - Z^2 - (x^2 - 2x\xi + \xi^2) - (y^2 - 2y\eta + \eta^2) = 2(x\xi + y\eta) - (x^2 + y^2) \tag{34}$$

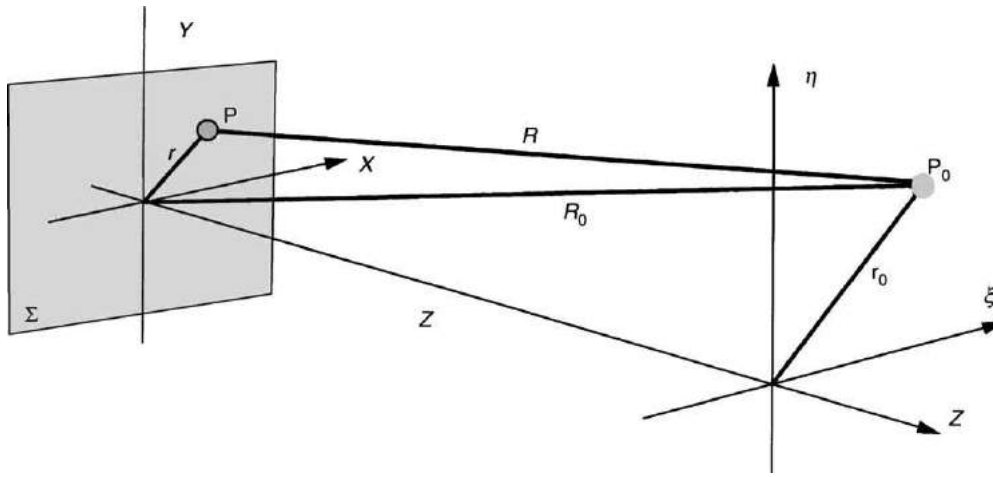


Fig. 4 Geometry for Fraunhofer diffraction. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The equation for the position of point P in the aperture can be written in terms of (34):

$$r = R_0 - R = \frac{R_0^2 - R^2}{R_0 + R}, = [2(x\xi + y\eta) - (x^2 + y^2)]\frac{1}{R_0 + R} \tag{35}$$

The reciprocal of $(R_0 + R)$ can be written as

$$\frac{1}{R_0 + R} = \frac{1}{2R_0 + R - R_0} = \frac{1}{2R_0}\left(1 + \frac{R - R_0}{2R_0}\right)^{-1} \tag{36}$$

Now using (36)

$$R_0 - R = \left[\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0}\right]\left[1 - \frac{R_0 - R}{2R_0}\right]^{-1} = \left[\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0}\right]\left[1 - \frac{r}{2R_0}\right]^{-1} \tag{37}$$

If the diffraction integral (30) is to have a finite (nonzero) value, then

$$k|R_0 - R| \ll kR_0 \tag{38}$$

This requirement insures that all the Huygens's wavelets, produced over the aperture from the center out to the position r, will have similar phases and will interfere to produce a nonzero amplitude at P_0. This requirement results in

$$\frac{1}{1 - \frac{r}{2R_0}} \approx 1 \tag{39}$$

From this equation we are tempted to state that the aperture dimension must be small relative to the observation distance. However, for the exponent to provide a finite contribution in the integral, it is kr that is small, not the aperture dimension, r.

Using the approximation for $(R_0 - R)$, we obtain the diffraction integral

$$\tilde{E}_P = \frac{i\alpha e^{-i\mathbf{k}\cdot\mathbf{R}_0}}{\lambda R_0}\int\int_{\sum} f(x,y)e^{+ik}\left(\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0}\right)dx\,dy \tag{40}$$

The change in the amplitude of the wave due to the change in r, as we move across the aperture, is neglected, allowing R in the denominator of the Huygens–Fresnel integral to be replaced by R_0 and moved outside of the integral. We have introduced the complex transmission function, $f(x,y)$, of the aperture to allow a very general aperture to be treated. If the aperture function described the variation in absorption of the aperture as a function of position, as would be produced by a photographic negative, then $f(x,y)$ would be a real function. If the aperture function described the variation in transmission of a biological sample, it might be entirely imaginary.

The argument of the exponent in (40) is

$$ik\left(\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y2}{2R_0}\right) = i2\pi\left(\frac{x\xi + y\eta}{\lambda R_0} - \frac{x^2 + y^2}{2\lambda R_0}\right) \tag{41}$$

If the observation point, P_0, is far from the screen, we can neglect the second term and treat the phase variation across the aperture as a linear function of position. This is equivalent to assuming that the diffracted wave is a collection of plane waves. Mathematically, the second term in (41) can be neglected if

$$\frac{x^2 + y^2}{2\lambda R_0} \ll 1 \tag{42}$$

This is called the far-field approximation and the theory yields Fraunhofer diffraction. If the quadratic term of (41) is on the order of 2π then the fraction is

$$\frac{x^2 + y^2}{2\lambda R_0} \approx \Theta(1) \tag{43}$$

and we must retain the quadratic term and the theory yields Fresnel diffraction.

We define spatial frequencies in the x and y direction by

$$\begin{aligned} \omega_x &= \frac{2\pi}{\lambda}\sin\theta_x = -\frac{2\pi\xi}{\lambda R_0} \\ \omega_y &= \frac{2\pi}{\lambda}\sin\theta_y = -\frac{2\pi\eta}{\lambda R_0} \end{aligned} \tag{44}$$

We define the spatial frequencies with a negative sign to allow equations involving spatial frequencies to have the same form as those involving temporal frequencies. The negative sign is required because ωt and $\mathbf{k}\cdot\mathbf{r}$ appear in the phase of the wave with opposite signs and we want the Fourier transform in space and in time to have the same form.

With the variables defined in (44), the integral becomes

$$\tilde{E}_P(\omega_x, \omega_y) = \frac{i\alpha e^{-i\mathbf{k}\cdot\mathbf{R}_0}}{\lambda R_0}\int\int_{\sum} f(x,y)e^{-i(\omega_x x + \omega_y y)}dx\,dy \tag{45}$$

The result of our derivation is that the Fraunhofer diffraction field, $\tilde{E}_P$ can be calculated by performing the two-dimensional Fourier transform of the aperture's transmission function.

Definition of Fourier Transform

In this discussion the function $F(\omega)$ is defined as the *Fourier transform* of $f(t)$:

$$\mathcal{F}\{f(t)\} \equiv F(\omega) \equiv \int_{-\infty}^{\infty} f(\tau)\mathrm{e}^{-\mathrm{i}\omega\tau}\mathrm{d}\tau \tag{46}$$

The transformation from a temporal to a frequency representation given by (46) does not destroy information; thus, the inverse transform can also be defined

$$\mathcal{F}^{-1}\{F(\omega)\} = f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}\omega \tag{47}$$

By assuming that the illumination of an aperture is a plane wave and that the observation position is in the far field, the diffraction of an aperture is found to be given by the Fourier transform of the function describing the amplitude transmission of the aperture. The amplitude transmission of the aperture, $f(x,y)$, may thus be interpreted as the superposition of mutually coherent plane waves.

Diffraction by a Rectangular Aperture

We will now use Fourier transform theory to calculate the Fraunhofer diffraction pattern from a rectangular slit and point out the reciprocal relationship between the size of the diffraction pattern and the size of the aperture.

Consider a rectangular aperture with a transmission function given by

$$f(x,y) = \begin{cases} 1 & |x| \le x_0, |y| \le y_0 \\ 0 & \text{all other } x \text{ and } y \end{cases} \tag{48}$$

Because the aperture is two-dimensional, we need to apply a two-dimensional Fourier transform but it is not difficult because the amplitude transmission function is separable, in x and y. The diffraction amplitude distribution from the rectangular slit is simply one-dimensional transforms carried out individually for the x and y dimensions:

$$\tilde{E}_P \frac{\mathrm{i}\alpha}{\lambda R_0}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}\int_{-x_0}^{x_0} f(x)\mathrm{e}^{-\mathrm{i}\omega_x x}\mathrm{d}x \int_{-y_0}^{y_0} f(y)\mathrm{e}^{-\mathrm{i}\omega_y y}\mathrm{d}y \tag{49}$$

Since both $f(x$ and $f(y)$ are defined as symmetric functions, we need only calculate the cosine transforms to obtain the diffracted field:

To calculate the Fourier transform defined by (46) we can rewrite the transform as

$$F(\omega) = \int_{-\infty}^{\infty} f(\tau)\cos\,\omega\tau\mathrm{d}\tau - \mathrm{i}\int_{-\infty}^{\infty} f(\tau)\sin\omega\tau\mathrm{d}\tau \tag{50}$$

If $f(\tau)$ is a real, even function then the Fourier transform can be obtained by simply calculating the cosine transform:

$$\int_{-\infty}^{\infty} f(\tau)\cos\,\omega\tau\mathrm{d}\tau \tag{51}$$

$$\tilde{E}_P = \mathrm{i}\frac{4x_0y_0\alpha}{\lambda R_0}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}\frac{\sin\omega_x x_0}{\omega_x x_0}\frac{\sin\omega_y y_0}{\omega_y y_0} \tag{52}$$

The intensity distribution of the Fraunhofer diffraction produced by the rectangular aperture is

$$I_P = I_0\frac{\sin^2\omega_x x_0}{(\omega_x x_0)^2}\frac{\sin^2\omega_y y_0}{(\omega_y y_0)^2} \tag{53}$$

The maximum intensities in the x and y directions occur at

$$\omega_x x_0 = \omega_y y_0 = 0 \tag{54}$$

The area of this rectangular aperture is defined as $A = 4x_0y_0$, resulting in an expression for the maximum intensity of

$$I_0 = \left[\frac{\mathrm{i}4x_0y_0\alpha\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}}{\lambda R_0}\right]^2 = \frac{A^2\alpha^2}{\lambda^2 R_0^2} \tag{55}$$

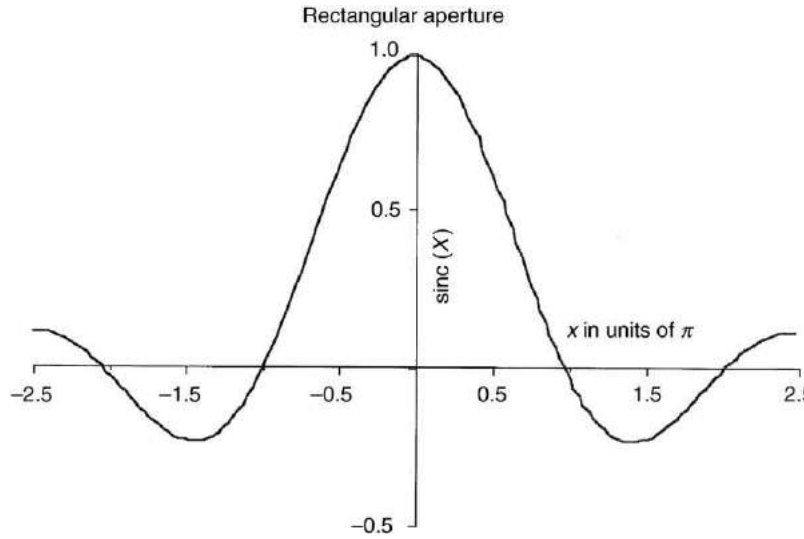


Fig. 5 The amplitude of the wave diffracted by a rectangular slit. This sinc function describes the light wave's amplitude that would exist in both the *x* and *y* directions. The coordinate would have to be scaled by the dimension of the slit. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The minima of (53) occur when $\omega_x x_0 = n\pi$ or when $\omega_y y_0 = m\pi$. The location of the zeros can be specified as a dimension in the observation plane or, using the paraxial approximation, in terms of the angle defined by (44)

$$\begin{aligned} \sin\theta_x \approx \theta_x \approx \frac{\xi}{R_0} &= \frac{n\lambda}{2x_0} \\ \sin\theta_y \approx \theta_y \approx \frac{\eta}{R_0} &= \frac{m\lambda}{2y_0} \end{aligned} \tag{56}$$

The dimensions of the diffraction pattern are characterized by the location of the first zero, i.e., when $n = m = 1$, and are given by the observation plane coordinates, ξ and η. The dimensions of the diffraction pattern are inversely proportional to the dimensions of the aperture. As the aperture dimension expands, the width of the diffraction pattern decreases until, in the limit of an infinitely wide aperture, the diffraction pattern becomes a delta function.

Fig. 5 is a theoretical plot of the amplitude of the diffraction wave from a rectangular slit.

Diffraction from a Circular Aperture

To obtain the diffraction pattern from a circular aperture we first convert from the rectangular coordinate system we have used to a cylindrical coordinate system. The new cylindrical geometry is shown in Fig. 6. At the aperture plane

$$\begin{aligned} x = s\cdot\cos\varphi \qquad & y = s\cdot\sin\varphi \\ f(x,y) = f(s,\varphi) \quad & \mathrm{d}x\,\mathrm{d}y = s\,\mathrm{d}s\mathrm{d}\varphi \end{aligned} \tag{57}$$

At the observation plane

$$\xi = \rho\cdot\cos\theta \quad \eta = \rho\cdot\sin\theta \tag{58}$$

In the new, cylindrical, coordinate system at the observation plane, the spatial frequencies are written as

$$\begin{aligned} \omega_x &= -\frac{k\xi}{R_0} = -\frac{k\rho}{R_0}\cos\theta \\ \omega_y &= -\frac{k\eta}{R_0} = -\frac{k\rho}{R_0}\sin\theta \end{aligned} \tag{59}$$

From Fig. 6 we see that the observation point, P, can be defined in terms of the angle ψ, where

$$\sin\psi = \frac{\rho}{R_0} \tag{60}$$

This allows an angular representation of the size of the diffraction pattern if it is desired.

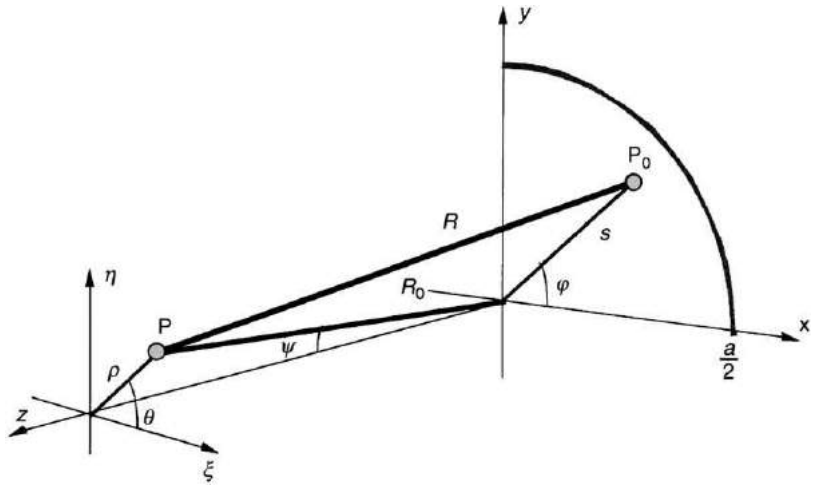


Fig. 6 Geometry for calculation of Fraunhofer diffraction from a circular aperture. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Using (57) and (59), we may write

$$\begin{aligned}\omega_x x + \omega_y y &= -\frac{ks\rho}{R_0}(\cos\theta\cos\varphi + \sin\theta\sin\varphi),\\ &= -\frac{ks\rho}{R_0}\cos(\theta-\varphi)\end{aligned} \tag{61}$$

The Huygens–Fresnel integral for Fraunhofer diffraction can now be written in terms of cylindrical coordinates:

$$\tilde{E}_P = \frac{i\alpha}{\lambda R_0} e^{-ik\cdot R_0}\int_0^{\frac{a}{2}}\int_0^{2\pi} f(s,\varphi) e^{-ik\frac{s\rho}{R_0}\cos(\theta-\varphi)} s\,ds\,d\varphi \tag{62}$$

We will use this equation to calculate the diffraction amplitude from a clear aperture of diameter a, defined by the equation

$$f(s,\varphi) = \begin{cases} 1 & s \le \dfrac{a}{2},\ \text{all } \varphi \\ 0 & s > \dfrac{a}{2} \end{cases} \tag{63}$$

The symmetry of this problem is such that

$$f(s,\varphi) = f(s)g(\varphi) = f(s) \tag{64}$$

$$\mathcal{F}\{f(s,\varphi)\} = \mathcal{F}\{f(s)\} = F(\rho) \tag{65}$$

$$F(\rho,\theta) = \int_0^{\infty} f(s)s\,ds\int_0^{2\pi} e^{-i\rho\cdot s\,\cos(\varphi-\theta)}d\varphi \tag{66}$$

$$F(\rho) = \int_0^{\infty} f(s)s\,ds\int_0^{2\pi} e^{-i\rho\cdot s\,\cos\varphi}d\varphi \tag{67}$$

The second integral in (67) belongs to a class of functions called the Bessel function defined by the integral

$$J_n(s\rho) = \frac{1}{2\pi}\int_0^{2\pi} e^{-i[s\rho\,\sin\varphi - n\varphi]}d\varphi \tag{68}$$

In (67) the integral corresponds to the $n=0$, zero-order, Bessel function. Using this definition, we can rewrite (67) as

$$F(\rho) = \int_0^{\infty} f(s)J_0(s\rho)s\,ds \tag{69}$$

This transform is called the Fourier–Bessel transform or the Hankel zero-order transform.

Using these definitions, the transform of the aperture function, $f(s)$, is

$$F(\rho) = \int_0^{\frac{a}{2}} J_0(s\rho)s\,ds \tag{70}$$

We use the identity

$$xJ_1(x) = \int_0^x \chi J_0(\chi)\,d\chi \tag{71}$$

to obtain

$$F(\rho) = \frac{a}{2\rho} J_1\left(\frac{\rho a}{2}\right) \tag{72}$$

We can now write the amplitude of the diffracted wave as

$$\tilde{E}_P = \frac{i\alpha}{\lambda} e^{-i\mathbf{k}\cdot\mathbf{R}_0} \left[\frac{\pi a}{k\rho} J_1\left(\frac{ka\rho}{2R_0}\right)\right] \tag{73}$$

A plot of the function contained within the bracket of (73) is given in Fig. 7. If we define

$$u = \frac{ka\rho}{2R_0} \tag{74}$$

then the spatial distribution of intensity in the diffraction pattern can be written in a form known as the Airy formula,

$$I = I_0 \left[\frac{2J_1(u)}{u}\right]^2 \tag{75}$$

where we have defined

$$I_0 = \left(\frac{\alpha A}{\lambda R_0}\right)^2 \tag{76}$$

with A representing the area of the aperture, $A = \pi\left(\frac{a}{2}\right)^2$.

The intensity pattern described by (75) is called the Airy pattern. The intensity at $u=0$ is the same as was obtained for a rectangular aperture of the same area, because, in the limit,

$$\lim_{u\to 0}\left[\frac{2J_1(u)}{u}\right] = 1 \tag{77}$$

For the Airy pattern, 84% of the total area is contained in the disk between the first zeros of (75). Those zeros occur at $u = \pm 1.22\,\pi$, which corresponds to a radius in the observation plane of

$$\rho = \frac{1.22(\lambda R_0)}{a} \tag{78}$$

91% of the light intensity is contained within the circle bounded by the second minimum at $u = 2.233\pi$. The intensities in the secondary maxima of the diffraction pattern of a rectangular aperture (53) are much larger than the intensities in the secondary maxima of the Airy pattern of a circular aperture of equal area. The peak intensities, relative to the central maximum, of the first three secondary maxima of a rectangular aperture are 4.7%, 1.6%, and 0.8%, respectively. For a circular aperture, the same quantities are 1.7%, 0.04%, and 0.02%, respectively.

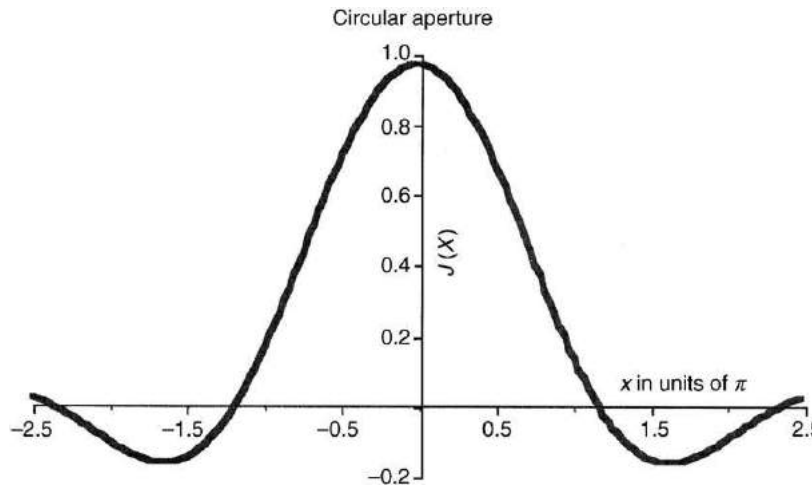


Fig. 7 Amplitude of a wave diffracted by a circular aperture. The observed light distribution is constructed by rotating this Bessel function around the optical axis. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Array Theorem

There is an elegant mathematical technique for handling diffraction from multiple apertures, called the array theorem. The theorem is based on the fact that Fraunhofer diffraction is given by the Fourier transform of the aperture function and utilizes the convolution integral.

The convolution of the functions $a(t)$ and $b(t)$ is defined as

$$g(t) = a(t) \otimes b(t) = \int_{-\infty}^{\infty} a(t)b(\tau - t)\,\mathrm{d}t \tag{79}$$

The Fourier transform of a convolution of two functions is the product of the Fourier transforms of the individual functions:

$$\mathcal{F}\{a(t) \otimes b(t)\} = \mathcal{F}\left\{\int_{-\infty}^{\infty} a(t)b(\tau - t)\,\mathrm{d}t\right\} = A(\omega)B(\omega) \tag{80}$$

We will demonstrate the array theorem for one dimension, where the functions represent slit apertures. The results can be extended to two dimensions in a straightforward way.

Assume that we have a collection of identical apertures, shown on the right of Fig. 8. If one of the apertures is located at the origin of the aperture plane, its transmission function is $\psi(x)$. The transmission function of an aperture located at a point, x_n, can be written in terms of a generalized aperture function, $\psi(x-\alpha)$, by the use of the sifting property of the delta function

$$\psi(x - x_n) = \int \psi(x - \alpha)\delta(\alpha - x_n)\,\mathrm{d}\alpha \tag{81}$$

The aperture transmission function representing an array of apertures will be the sum of the distributions of the individual apertures, represented graphically in Fig. 8 and mathematically by the summation

$$\Psi(x) = \sum_{n=1}^{N} \psi(x - x_n) \tag{82}$$

The Fraunhofer diffraction from this array is $\Phi(\omega_x)$, the Fourier transform of $\Psi(x)$,

$$\Phi(\omega_x) = \int_{-\infty}^{\infty} \Psi(x)\mathrm{e}^{-\mathrm{i}\omega_x x}\,\mathrm{d}x \tag{83}$$

which can be rewritten as

$$\Phi(\omega_x) = \sum_{n=1}^{N} \int_{-\infty}^{\infty} \psi(x - x_n)\mathrm{e}^{-\mathrm{i}\omega_x x}\,\mathrm{d}x \tag{84}$$

We now make use of the fact that $\psi(x-x_n)$ can be expressed in terms of a convolution integral. The Fourier transform of $\psi(x-x_n)$ is, from the convolution theorem (80), the product of the Fourier transforms of the individual functions that make up the convolution:

$$\Phi(\omega_x) \quad = \sum_{n=1}^{N} \mathcal{F}\{\psi(x-\alpha)\}\mathcal{F}\{\delta(\alpha - x_n)\} = \mathcal{F}\{\psi(x-\alpha)\}\sum_{n=1}^{N}\mathcal{F}\{\delta(\alpha - x_n)\} = \mathcal{F}\{\psi(x-\alpha)\}\mathcal{F}\left\{\sum_{n=1}^{N}\delta(\alpha - x_n)\right\} \tag{85}$$

The first transform in (85) is the diffraction pattern of the generalized aperture function and the second transform is the diffraction pattern produced by a set of point sources with the same spatial distribution as the array of identical apertures. We will call this second transform the array function.

To summarize, the array theorem states that the diffraction pattern of an array of similar apertures is given by the product of the diffraction pattern from a single aperture and the diffraction (or interference) pattern of an identically distributed array of point sources.

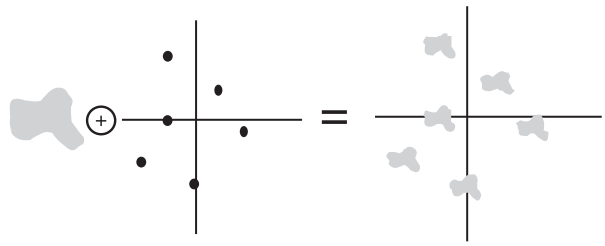


Fig. 8 The convolution of an aperture with an array of delta functions will produce an array of identical apertures, each located at the position of one of the delta functions. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

N Rectangular Slits

An array of N identical apertures is called a diffraction grating in optics. The Fraunhofer diffraction patterns produced by such an array have two important properties; a number of very narrow beams are produced by the array and the beam positions are a function of the wavelength of illumination of the apertures and the relative phase of the waves radiated by each of the apertures.

Because of these properties, arrays of diffracting apertures have been used in a number of applications.

- At radio frequencies, arrays of dipole antennas are used to both radiate and receive signals in both radar and radio astronomy systems. One advantage offered by diffracting arrays at radio frequencies is that the beam produced by the array can be electrically steered by adjusting the relative phase of the individual dipoles.
- An optical realization of a two-element array of radiators is Young's two-slit experiment and a realization of a two-element array of receivers is Michelson's stellar interferometer. Two optical implementations of arrays, containing more diffracting elements, are diffraction gratings and holograms. In nature, periodic arrays of diffracting elements are the origin of the colors observed on some invertebrates.
- Many solids are naturally arranged in three-dimensional arrays of atoms or molecules that act as diffraction gratings when illuminated by X-ray wavelengths. The resulting Fraunhofer diffraction patterns are used to analyze the ordered structure of the solids.

The array theorem can be used to calculate the diffraction pattern from N rectangular slits, each of width a and separation d. The aperture function of a single slit is equal to

$$\psi(x,y) = \begin{cases} 1 & |x| \le \dfrac{a}{2}, |y| \le y_0 \\ 0 & \text{all other } x \text{ and } y \end{cases} \tag{86}$$

The Fraunhofer diffraction pattern of this aperture has already been calculated and is given by **(52)**:

$$\mathcal{F}\{\psi(x,y)\} = K\frac{\sin\ \alpha}{\alpha} \tag{87}$$

where the constant K and the variable α are

$$\begin{aligned} K &= \frac{\mathrm{i}2ay_0\alpha}{\lambda R_0}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}\frac{\sin\ \omega_y y_0}{\omega_y y_0} \\ \alpha &= \frac{ka}{2}\sin\ \theta_x \end{aligned} \tag{88}$$

The array function is

$$A(x) = \sum_{n=1}^{N}\delta(x - x_n) \quad \text{where } x_n = (n-1)d \tag{89}$$

and its Fourier transform is given by

$$\mathcal{F}\{A(x)\} = \frac{\sin\ N\beta}{\sin\beta} \tag{90}$$

$$\beta = \frac{\mathrm{kd}}{2}\sin\ \theta_x$$

The Fraunhofer diffraction pattern's intensity distribution in the x-direction is thus given by

$$I_\theta = I_0 \underbrace{\frac{\sin^2\alpha}{\alpha^2}}_{\text{shape factor}} \underbrace{\frac{\sin^2 N\beta}{\sin^2\beta}}_{\text{grating factor}} \tag{91}$$

We have combined the variation in intensity in the y-direction into the constant I_0 because we assume that the intensity variation in the x-direction will be measured at a constant value of y.

A physical interpretation of **(91)** views the first factor as arising from the diffraction associated with a single slit; it is called the shape factor. The second factor arises from the interference between light from different slits; it is called the grating factor. The fine detail in the spatial light distribution of the diffraction pattern is described by the grating factor and arises from the coarse detail in the diffracting object. The coarse, overall light distribution in the diffraction pattern is described by the shape factor and arises from the fine detail in the diffracting object.

Young's Double Slit

The array theorem makes the analysis of Young's two-slit experiment a trivial exercise. This application of the array theorem will demonstrate that the interference between two slits arises naturally from an application of diffraction theory. The result of this analysis will support a previous assertion that interference describes the same physical process as diffraction and the division of the two subjects is an arbitrary one.

The intensity of the diffraction pattern from two slits is obtained from (91) by setting $N=2$:

$$I_\theta = I_0 \frac{\sin^2\alpha}{\alpha^2} \cos^2\beta \tag{92}$$

The sinc function describes the energy distribution of the overall diffraction pattern, while the cosine function describes the energy distribution created by interference between the light waves from the two slits. Physically, α is a measure of the phase difference between points in one slit and β is a measure of the phase difference between similar points in the two slits. Zeros in the diffraction intensity occur whenever $\alpha = n\pi$ or whenever $\beta = \frac{1}{2}(2n+1)\pi$. Fig. 9 shows the interference maxima, from the grating factor, under the central diffraction peak, described by the shape factor. The number of interference maxima contained under the central maximum is given by

$$\frac{2d}{a} - 1 \tag{93}$$

In Fig. 9 three different slit spacings are shown with the ratio d/a equal to 3, 6, and 9, respectively.

The Diffraction Grating

In this section we will use the array theorem to derive the diffraction intensity distribution of a large number of identical apertures. We will discover that the positions of the principal maxima are a function of the illuminating wavelength. This functional relationship has led to the application of a diffraction grating to wavelength measurements.

The diffraction grating normally used for wavelength measurements is not a large number of diffracting apertures but rather a large number of reflecting grooves cut in a surface such as gold or aluminum. The theory to be derived also applies to these

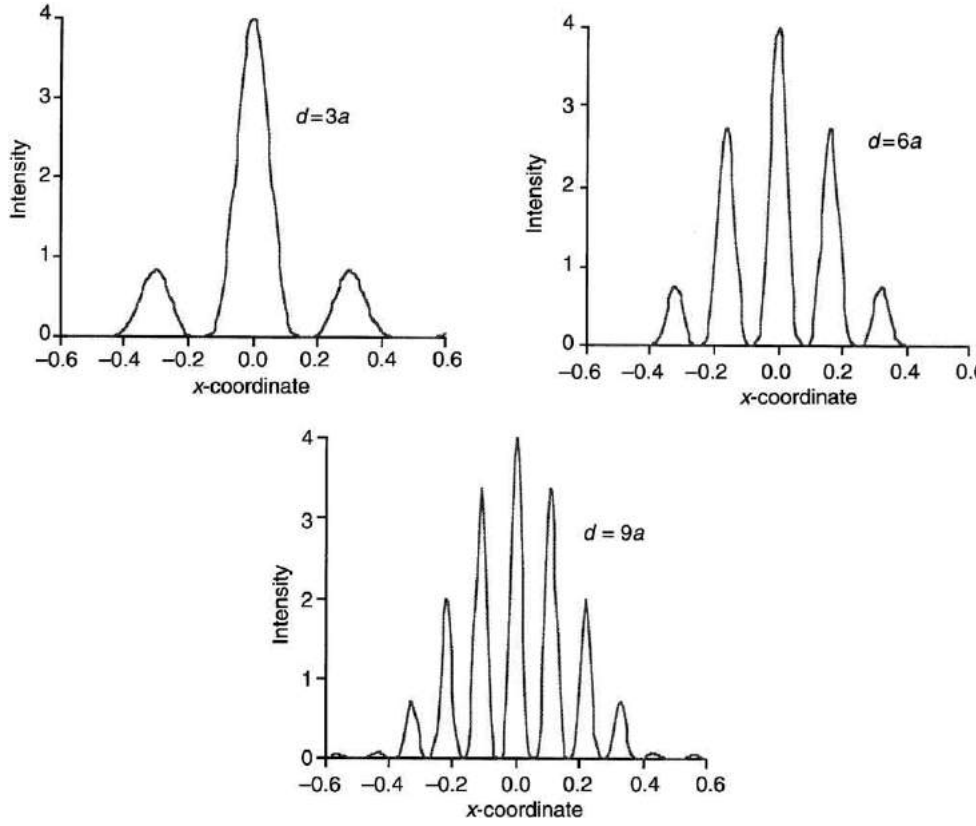


Fig. 9 The number of interference fringes beneath the main diffraction peak of a Young's two-slit experiment with rectangular apertures. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

reflection gratings but a modification must be made to the theory. The shape of the grooves in the reflection grating can be used to control the fraction of light diffracted into a principal maximum and we will examine how to take this into account. A grating whose groove shape has been controlled to enhance the energy contained in a particular principal maximum is called a blazed grating. The use of special groove shapes is equivalent to the modification of the phase of individual elements of an antenna array at radio frequencies.

The construction of a diffraction grating for use in an optical device for measuring wavelength was first suggested by David Rittenhouse, an American astronomer, in 1785, but the idea was ignored until Fraunhofer reinvented the concept in 1819. Fraunhofer's first gratings were fine wires spaced by wrapping the wires in the threads of two parallel screws. He later made gratings by cutting (*ruling*) grooves in gold films deposited on the surface of glass. H.A. Rowland made a number of well designed ruling machines, which made possible the production of large-area gratings. Following a suggestion by Lord Rayleigh, Robert Williams Wood (1868–1955) developed the capability to control the shape of the individual grooves.

If N is allowed to assume values much larger than 2, the appearance of the interference fringes, predicted by the grating factor, changes from a simple sinusoidal variation to a set of narrow maxima, called principal maxima, surrounded by much smaller, secondary maxima. To calculate the diffraction pattern we must evaluate **(91)** when N is a large number.

Whenever $N\beta = m\pi$, $m = 0,1,2,\ldots$, the numerator of the second factor in **(91)** will be zero, leading to an intensity that is zero, $I_\theta = 0$. The denominator of the second factor in **(91)** is zero when $\beta = l\pi$, $l = 0,1,2,\ldots$ If both conditions occur at the same time, the ratio of m and N is equal to an integer, $m/N = l$, and instead of zero intensity, we have an indeterminate value for the intensity, $I_\theta = 0/0$. To find the actual value for the intensity, we must apply L'Hospital's rule

$$\lim_{\beta \to l\pi} \frac{\sin N\beta}{\sin \beta} = \lim_{\beta \to l\pi} \frac{N \cos N\beta}{\cos \beta} = N \tag{94}$$

L'Hospital's rule predicts that whenever

$$\beta = l\pi = \frac{m}{N}\pi \tag{95}$$

where l is an integer, a principal maximum in the intensity will occur with a value given by

$$I_{\theta P} = N^2 I_0 \frac{\sin^2 \alpha}{\alpha^2} \tag{96}$$

Secondary maxima, much weaker than the principal maxima, occur when

$$\beta = \left(\frac{2m+1}{2N}\right)\pi \quad m = 1, 2, \ldots \tag{97}$$

(When $m = 0$, m/N is an integer, thus the first value that m can have in **(97)** is $m = 1$.) The intensity of each secondary maximum is given by

$$I_{\theta S} = I_0 \frac{\sin^2 \alpha}{\alpha^2} \frac{\sin^2 N\beta}{\sin^2 \beta} = I_0 \frac{\sin^2 \alpha}{\alpha^2} \left[\frac{\sin^2\left(\frac{2m+1}{2}\right)\pi}{\sin^2\left(\frac{2m+1}{2N}\right)\pi}\right] = I_0 \frac{\sin^2 \alpha}{\alpha^2} \left[\frac{1}{\sin\left(\frac{2m+1}{2N}\right)\pi}\right]^2 \tag{98}$$

The quantity $(2m+1)/2N$ is a small number for large values of N, allowing the small angle approximation to be made:

$$I_{\theta S} \approx I_0 \frac{\sin^2 \alpha}{\alpha^2} \left[\frac{2N}{\pi(2m+1)}\right]^2 \tag{99}$$

The ratio of the intensity of a secondary maximum and a principal maximum is given by

$$\frac{I_{\theta S}}{I_{\theta P}} = \left[\frac{2}{\pi(2m+1)}\right]^2 \tag{100}$$

The strongest secondary maximum occurs for $m = 1$ and, for large N, has an intensity that is about 4.5% of the intensity of the neighboring principal maximum.

The positions of principal maxima occur at angles specified by the grating formula.

$$\beta = \frac{\mathrm{kd} \sin \theta}{2} = l\pi = \frac{m}{N}\pi \tag{101}$$

The angular position of the principal maxima (the diffracted angle θ_d) is given by the Bragg equation

$$\sin \theta_\mathrm{d} = \frac{l\lambda}{d} \tag{102}$$

where l is called the grating order. This simple relationship between the angle and the wavelength can be used to construct a device to measure wavelength, the grating spectrometer.

The model we have used to obtain this result is based on a periodic array of identical apertures. The transmission function of this array is a periodic square wave. If, for the moment, we treat the grating as infinite in size, we discover that the principal maxima in the diffraction pattern correspond to the terms of the Fourier series describing a square wave.

The zero order, $m = 0$, corresponds to the temporal average of the periodic square wave and has an intensity proportional to the spatially averaged transmission of the grating. Because of its equivalence to the temporal average of a time-varying signal, the zero-order principal maximum is often called the dc term.

The first-order principal maximum corresponds to the fundamental spatial frequency of the grating and the higher orders correspond to the harmonics of this frequency.

The dc term provides no information about the wavelength of the illumination. Information about the wavelength of the illuminating light can only be obtained by measuring the angular position of the first or higher order principal maxima.

Grating Spectrometer

The curves shown in Fig. 10 display the separation of the principal maxima as a function of sin θ_d. The angular separation of principal maxima can be converted to a linear dimension by assuming a distance, r, from the graing to the observation plane. In the lower right-hand curve of Fig. 10, a distance of 2 meters was assumed. Grating spectrometers are classified by the size in meters of R used in their design. The larger r, the easier it is to resolve wavelength differences. For example, a 1 meter spectrometer is a higher-resolution instrument than a 1/4 meter spectrometer.

The fact that the grating is finite in size causes each of the orders to have an angular width that limits the resolution with which the illuminating wavelength can be measured. To calculate the resolving power of the grating, we first determine the angular width of a principal maximum. This is accomplished by measuring the angular change, of the principal maximum's position, when β changes from $\beta = l\pi = m\pi/N$ to $\beta = (m+1)\pi/N$, i.e., $\Delta\beta = \pi/N$. Using the definition of β

$$\beta = \frac{kd \sin \theta_d}{2} \tag{103}$$

$$\Delta\beta = \frac{\pi d \cos \theta_d \Delta\theta}{\lambda}$$

and the angular width is

$$\Delta\theta = \frac{\lambda}{Nd \cos \theta_d} \tag{104}$$

The derivative of the grating formula gives

$$\Delta\lambda = \frac{d}{l} \cos \theta_d \Delta\theta \tag{105}$$

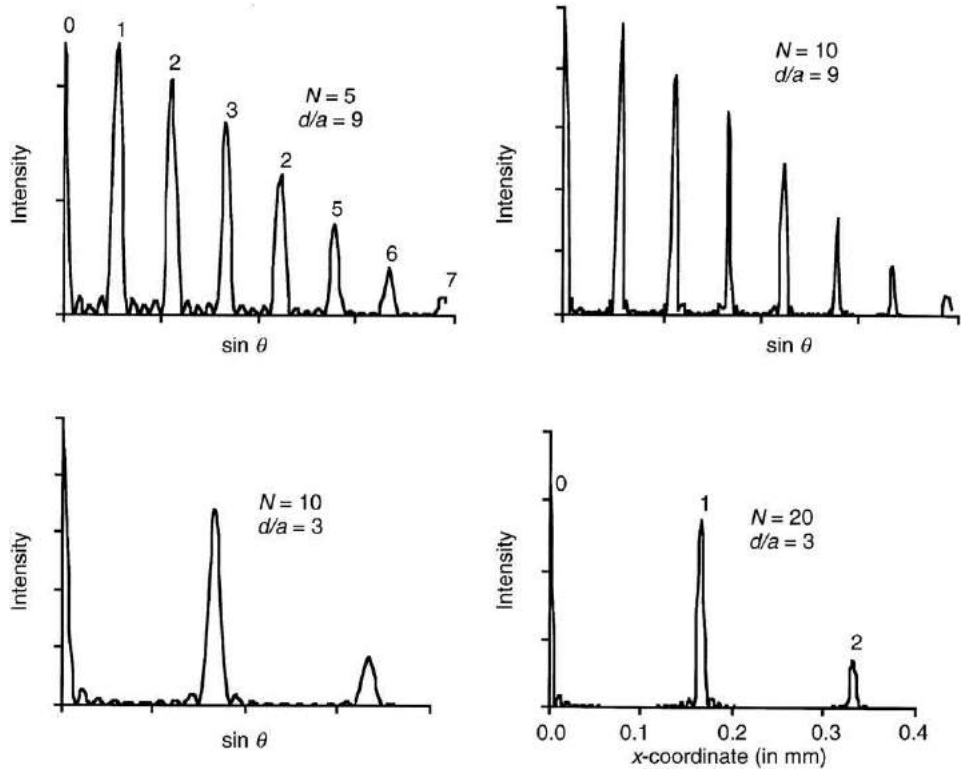


Fig. 10 Decrease in the width of the principal maxima of a transmission grating with an increasing number of slits. The various principal maxima are called orders, numbering from zero, at the origin, out to as large as seven in this example. Also shown is the effect of different ratios, *d*/*a*, on the number of visible orders. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

$$\Delta\theta = \frac{l\Delta\lambda}{d\cos\theta_d}$$

Equating this spread in angle to **Eq. (104)** yields

$$\frac{l\Delta\lambda}{d\cos\theta} = \frac{\lambda}{Nd\cos\theta} \tag{106}$$

The resolving power of a grating is therefore

$$\frac{\lambda}{\Delta\lambda} = Nl \tag{107}$$

The improvement of resolving power with N can be seen in **Fig. 10**. A grating 2 inches wide and containing 15 000 grooves per inch would have a resolving power in second order ($l=2$) of 6×10^4. At a wavelength of 600 nm this grating could resolve two waves, differing in wavelength by 0.01 nm.

The diffraction grating is limited by overlapping orders. If two wavelengths, λ and $\lambda+\Delta\lambda$, have successive orders that are coincident, then from **Eq. (102)**

$$(l+1)\lambda = l(\lambda+\Delta\lambda) \tag{108}$$

The minimum wavelength difference for which this occurs is defined as the free spectral range of the diffraction grating

$$(\Delta\lambda)_{SR} = \frac{\lambda}{l} \tag{109}$$

Blazed Gratings

We have been discussing amplitude transmission gratings. Amplitude transmission gratings have little practical use because they waste light. The light loss is from a number of sources.

1. Light is diffracted simultaneously into both positive and negative orders (the positive and negative frequencies of the Fourier transform). The negative diffraction orders contain redundant information and waste light.
2. In an amplitude transmission grating, light is thrown away because of the opaque portions of the slit array.
3. The width of an individual aperture leads to a shape factor for a rectangular aperture of$\mathrm{sinc}^2\alpha = (\sin^2\alpha)/\alpha^2$, which modulates the grating factor and causes the amplitude of the orders to rapidly decrease. This can be observed in **Fig. 10**, where the second order is very weak. Because of the loss in intensity at higher orders, only the first few orders ($l=1,2$ or) are useful. The shape factor also causes a decrease in diffracted light intensity with increasing wavelength for higher-order principal maxima.
4. The location of the maximum in the diffracted light, i.e., the angular position for which the shape factor is a maximum, coincides with the location of the principal maximum due to the zero-order interference. This zero-order maximum is independent of wavelength and is not of much use.

One solution to the problems created by transmission gratings would be the use of a grating that modified only the phase of the transmitted wave. Such gratings would operate using the same physical processes as a microwave phased array antenna, where the location of the shape factor's maximum is controlled by adding a constant phase shift to each antenna element. The construction of an optical transmission phase grating with a uniform phase variation across the aperture of the grating is very difficult. For this reason, a second approach, based on the use of reflection gratings, is the practical solution to the problems listed above.

By tilting the reflecting surface of each groove of a reflection grating, **Fig. 11**, the position of the shape factor's maximum can be controlled. Problems 1, 3, and 4 are eliminated because the shape factor maximum is moved from the optical axis out to some angle with respect to the axis. The use of reflection gratings also removes Problem 2 because all of the incident light is reflected by the grating.

Robert Wood, in 1910, developed the technique of producing grooves of a desired shape in a reflective grating by shaping the diamond tool used to cut the grooves. Gratings, with grooves shaped to enhance their performance at a particular wavelength, are said to be blazed for that wavelength. The physical properties on which blazed gratings are based can be understood by using **Fig. 11**. The groove faces can be treated as an array of mirror surfaces. The normal to each of the groove faces makes an angle θ_B with the normal to the grating surface. We can measure the angle of incidence and the angle of diffraction with respect to the grating normal or with respect to the groove normal, as shown in **Fig. 11**. From **Fig. 11**, we can write a relationship between the angles

$$\theta_i = \varphi_i - \theta_B \quad -\theta_d = -\varphi_d + \theta_B \tag{110}$$

(The sign convention used defines positive angles as those measured in a counterclockwise rotation from the normal to the surface. Therefore, θ_B is a negative angle.) The blaze angle provides an extra degree of freedom that will allow independent adjustment of the angular location of the principal maxima of the grating factor and the zero-order, single-aperture, diffraction maximum. To see how this is accomplished, we must determine, first, the effect of off-axis illumination of a diffraction grating.

Off-axis illumination is easy to incorporate into the equation for the diffraction intensity from an array. To include the effect of an off-axis source, the phase of the illuminating wave is modified by changing the incident illumination from a plane wave of amplitude, E, traveling parallel to the optical axis, to a plane wave with the same amplitude, traveling at an angle θ_i to the optical

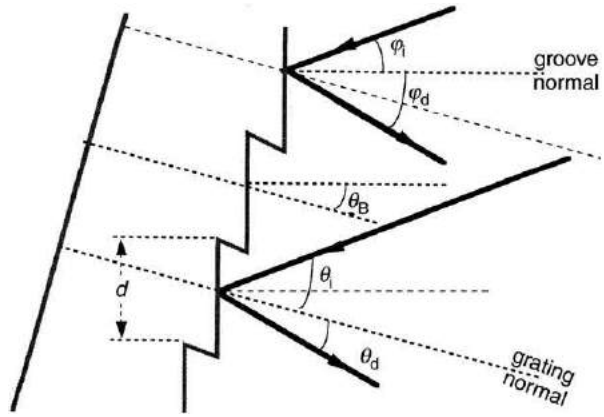


Fig. 11 Geometry for a blazed reflection grating. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

axis: $\tilde{E}e^{-ikx\sin\theta_i}$. (Because we are interested only in the effects in a plane normal to the direction of propagation, we ignore the phase associated with propagation along the z-direction, $kz\cos\theta_i$.) The off-axis illumination results in a modification of the parameter for single-aperture diffraction from

$$\alpha = \frac{ka}{2}\sin\theta_d \tag{111}$$

to

$$\alpha = \frac{ka}{2}(\sin\theta_i + \sin\theta_d) \tag{112}$$

and for multiple aperture interference from

$$\beta = \frac{kd}{2}\sin\theta_d \tag{113}$$

to

$$\beta = \frac{kd}{2}(\sin\theta_i + \sin\theta_d) \tag{114}$$

The zero-order, single-aperture, diffraction peak occurs when $\alpha = 0$. If we measure the angles with respect to the groove face, this occurs when

$$\alpha = \frac{ka}{2}(\sin\varphi_i + \sin\varphi_d) = 0 \tag{115}$$

The angles are therefore related by

$$\sin\varphi_i = -\sin\varphi_d \tag{116}$$

$$\varphi_i = -\varphi_d$$

We see that the single-aperture, diffraction maximum (the shape factor's maximum) occurs at the same angle that reflection from the groove faces occurs. We can write this result in terms of the angles measured with respect to the grating normal

$$\theta_i = -(\theta_d + 2\theta_B) \tag{117}$$

The blaze condition requires the single aperture diffraction maximum to occur at the lth principal maximum, for wavelength λ_B. At that position

$$l\pi = \frac{2\pi d}{\lambda_B}(\sin\theta_i + \sin\theta_d) \tag{118}$$

$$l\lambda_B = 2d\sin\frac{1}{2}(\theta_i + \theta_d)\cos\frac{1}{2}(\theta_i - \theta_d)$$

For the special geometrical configuration called the Littrow condition, where $\theta_i = \theta_d$, we find that (117) leads to the equation

$$l\lambda_B = 2d\sin\theta_B \tag{119}$$

By adjusting the blaze angle, the single-aperture diffraction peak can be positioned on any order of the interference pattern. Typical blaze angles are between 15° and 30° but gratings are made with larger blaze angles.

Further Reading

Born, M., Wolf, E., 1970. Principles of Optics. New York: Pergamon Press.
Guenther, R.D., 1990. Modern Optics. New York: Wiley.
Haus, H.A., 1984. Waves and Fields in Optoelectronics. Englewood Cliffs, NJ: Prentice-Hall.
Hecht, E., 1998. Optics, 3rd edn Reading, MA: Addison-Wesley.
Jackson, J.D., 1962. Classical Electrodynamics. New York: Wiley.
Klein, M.V., 1970. Optics. New York: Wiley.
O'Neill, E.L., 1963. Introduction to Statistical Optics. Reading, MA: Addison-Wesley.
Rossi, B., 1957. Optics. Reading, MA: Addison-Wesley.
Rubinowicz, A., 1984. In: Wolf, E. (Ed.), Progress in Optics,Tahe Miyamoto–Wolf diffraction wave, vol. IV. North-Holland: Amsterdam, p. 201.
Sommerfeld, A., 1964. Optics. New York: Academic Press.
Stone, J.M., 1963. Radiation and Optics. New York: McGraw-Hill.

Fresnel Diffraction

BD Guenther, Duke University, Durham, NC, USA

Fresnel was a civil engineer who pursued optics as a hobby, after a long day of road construction. Based on his own very high-quality observations of diffraction, Fresnel used the wave propagation concept developed by Christiaan Huygens (1629–1695) to develop a theoretical explanation of diffraction. We will use a descriptive approach to obtain the Huygens–Fresnel integral by assuming an aperture can be described by N pinholes which act as sources for Huygens' wavelets. The interference between these sources will lead to the Huygens–Fresnel integral for diffraction.

Huygens' principle views wavefronts as the product of wavelets from various luminous points acting together. To apply Huygens' principle to the propagation of light through an aperture of arbitrary shape, we need to develop a mathematical description of the field from an array of Huygens' sources filling the aperture. We will begin by obtaining the field from a pinhole that is illuminated by a plane wave,

$$\tilde{\mathrm{E}}_{\mathrm{i}}(\mathrm{r},t)=\tilde{\mathrm{E}}_{\mathrm{i}}(\mathrm{r})\mathrm{e}^{\mathrm{i}\omega t}$$

Following the lead of Fresnel, we will use theory of interference to combine the fields from two pinholes and then generalize to N pinholes. Finally by letting the areas of each pinhole approach an infinitesimal value, we will construct an arbitrary aperture of infinitesimal pinholes. The result will be the Huygens–Fresnel integral.

We know that the wave obtained after propagation through an aperture must be a solution of the wave equation,

$$\nabla^2\tilde{\mathrm{E}}=\mu\varepsilon\frac{\partial^2\tilde{\mathrm{E}}}{\partial t^2}$$

We will be interested only in the spatial variation of the wave so we need only look for solutions of the Helmholtz equation

$$(\nabla^2+k^2)\tilde{\mathrm{E}}=0$$

The problem is further simplified by replacing this vector equation with a scalar equation,

$$(\nabla^2+k^2)\tilde{E}(x,y,z)=0$$

This replacement is proper for those cases where $\mathbf{n}E(x,y,z)$ [where $\mathbf{n}$ is a unit vector] is a solution of the vector Helmholtz equation. In general, we cannot substitute $\mathbf{n}E$ for the electric field $\mathbf{E}$ because of Maxwell's equation

$$\nabla\cdot\mathrm{E}=0$$

Rather than working with the magnitude of the electric field, we should use the scalar amplitude of the vector potential. We will neglect this complication and assume the scalar, E, is a single component of the vector field, $\mathbf{E}$. A complete solution would involve a scalar solution for each component of $\mathbf{E}$.

The pinhole is illuminated by a plane wave, and the wave that leaves the pinhole will be a spherical wave which is written in complex notation as

$$\tilde{E}(\mathrm{r})\mathrm{e}^{\mathrm{i}\omega t}=A\frac{\mathrm{e}^{-\mathrm{i}\delta}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathrm{r}}}{r}\mathrm{e}^{\mathrm{i}\omega t}$$

The complex amplitude

$$\tilde{E}(\mathrm{r})=A\frac{e^{-\mathrm{i}\delta}e^{-\mathrm{i}\mathbf{k}\cdot\mathrm{r}}}{r} \tag{1}$$

is a solution of the Helmholtz equation. The field at P_0, in Fig. 1, from two pinholes: one at P_1, located a distance $\mathrm{r}_{01}=|\mathrm{r}_0-\mathrm{r}_1|$ from P_0, and one at P_2, located a distance $\mathrm{r}_{02}=|\mathrm{r}_0-\mathrm{r}_2|$ from P_0, is a generalization of Young's interference. The complex amplitude is

$$\tilde{E}(\mathrm{r}_0)=\frac{A_1}{r_{01}}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathrm{r}_{01}}+\frac{A_2}{r_{02}}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathrm{r}_{02}} \tag{2}$$

We have incorporated the phase δ_1 and δ_2 into the constants A_1 and A_2 to simplify the equations.

The light emitted from the pinholes is due to a wave, E_{i}, incident onto the screen from the left. The units of E_{i} are per unit area so to obtain the amount of light passing through the pinholes, we must multiply E_{i} by the areas of the pinholes. If $\Delta\sigma_1$ and $\Delta\sigma_2$ are the areas of the two pinholes, respectively, then

$$A_1\propto\tilde{E}_{\mathrm{i}}(\mathrm{r}_1)\Delta\sigma_1\quad A_2\propto\tilde{E}_{\mathrm{i}}(\mathrm{r}_2)\Delta\sigma_2$$

$$\tilde{E}(\mathrm{r}_0)=C_1\frac{\tilde{E}_{\mathrm{i}}(\mathrm{r}_1)}{r_{01}}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathrm{r}_{01}}\Delta\sigma_1+C_2\frac{\tilde{E}_{\mathrm{i}}(\mathrm{r}_2)}{r_{02}}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathrm{r}_{02}}\Delta\sigma_2 \tag{3}$$

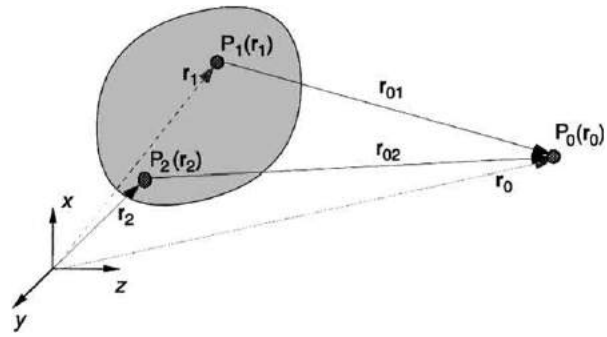


Fig. 1 Geometry for application of Huygens' principle to two pinholes. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

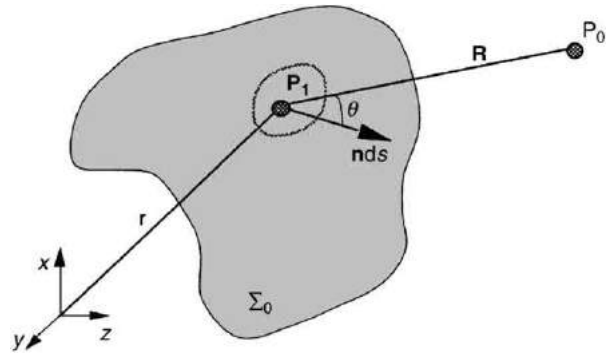


Fig. 2 The geometry for calculating the field at P_0 using (9). Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

where C_i is the constant of proportionality. The constant will depend on the angle that $\mathbf{r}_{0i}$ makes with the normal to $\Delta\sigma_i$. This geometrical dependence arises from the fact that the apparent area of the pinhole decreases as the observation angle approaches 90°.

We can generalize (3) to N pinholes

$$\tilde{E}(\mathrm{r}_0) = \sum_{j=1}^{N} C_j \frac{\tilde{E}_\mathrm{i}(\mathrm{r}_j)}{r_{0j}} \mathrm{e}^{-\mathrm{ik}\cdot\mathrm{r}_{0j}} \Delta\sigma_j \tag{4}$$

The pinhole's diameter is assumed to be small compared to the distances to the viewing position but large compared to a wavelength. In the limit as $\Delta\sigma_j$ goes to zero, the pinhole becomes a Huygens source. By letting N become large, we can fill the aperture with these infinitesimal Huygens' sources and convert the summation to an integral.

It is in this way that we obtain the complex amplitude, at the point P_0, see Fig. 2, from a wave exiting an aperture, $\sum_0$, by integrating over the area of the aperture. We replace $\mathbf{r}_{0j}$, in (4), by $\mathbf{R}$, the position of P_1, the infinitesimal Huygens' source of area, ds, measured with respect to the observation point, P_0. We also replace $\mathbf{r}_j$ by $\mathbf{r}$, the position of the infinitesimal area, ds, with respect to the origin of the coordinate system. The discrete summation (4) becomes the integral

$$\tilde{E}(\mathrm{r}_0) = \int\int_A C(\mathrm{r}) \frac{\tilde{E}_\mathrm{i}(\mathrm{r})}{R} \mathrm{e}^{-\mathrm{ik}\cdot\mathrm{R}} \mathrm{ds} \tag{5}$$

This is the Fresnel integral. The variable $C(\mathbf{r})$ depends upon θ, the angle between $\mathbf{n}$, the unit vector normal to the aperture, and $\mathbf{R}$, shown in Fig. 2. We now need to determine how to treat $C(\mathbf{r})$.

The Obliquity Factor

When using Huygens' principle, a problem arises with the spherical wavelet produced by each Huygens' source. Part of the spherical wavelet propagates backward and results in an envelope propagating toward the light source, but such a wave is not observed in nature. Huygens neglected this problem. Fresnel required the existence of an obliquity factor to cancel the backward wavelet but was unable to derive its functional form. When Kirchhoff placed the theory of diffraction on a firm mathematical foundation, the obliquity factor was generated quite naturally in the derivation. In this intuitive derivation of the Huygens–Fresnel

integral, we will present an argument for treating the obliquity factor as a constant at optical wavelengths. We will then derive the constant value it must be assigned.

The obliquity factor, $C(\mathbf{r})$, in (5), causes the amplitude per unit area of the transmitted light to decrease as the viewing angle increases. This is a result of a decrease in the effective aperture area with viewing angle. When Kirchhoff applied Green's theorem to the scalar diffraction problem, he found the obliquity factor to have an angular dependence given by

$$\frac{\cos(\hat{\mathrm{n}}, \mathrm{R}) - \cos(\hat{\mathrm{n}}, \mathrm{r_s})}{2} \tag{6}$$

which includes the geometrical effect of the incident wave arriving at the aperture, at an angle with respect to the normal to the aperture from a source located at a position $\mathbf{r}_s$, with respect to the aperture. If the source is at infinity, then the incident wave can be treated as a plane wave, incident normal to the aperture, and we may simplify the angular dependence of the obliquity factor to

$$\frac{1 + \cos\theta}{2}$$

where $\theta \rightarrow (\hat{\mathrm{n}}, \mathrm{R})$ is the angle between the normal to the aperture and the vector $\mathbf{R}$. This is the configuration shown in Fig. 2. The obliquity factor provides an explanation of why it is possible to ignore the backward-propagating wave that occurs in application of Huygens' principle. For the backward wave, $\theta = \pi$, and the obliquity factor is zero.

The obliquity factor increases the difficulty of working with the Fresnel integral and it is to our benefit to be able to treat it as a constant. We can neglect the angular contribution of the obliquity factor by making an assumption about the resolving power of an optical system operating at visible wavelengths.

Assume we are attempting to resolve two stars that produce plane waves at the aperture of a telescope with an aperture diameter, a. The wavefronts from the two stars make an angle ψ with respect to each other, Fig. 3:

$$\tan\psi \approx \psi = \frac{\Delta x}{a}$$

The smallest angle, ψ, that can be measured is determined by the smallest length, Δx, that can be measured. We know we can measure a fraction of wavelength with an interferometer but, without an interferometer, we can measure a length no smaller than λ, leading to the assumption that $\Delta x \geq \lambda$. This reasoning leads to the assumption that the smallest angle we can measure is

$$\psi \geq \frac{\lambda}{a} \tag{7}$$

The resolution limit established by the above reasoning places a limit on the minimum separation that can be produced at the back focal plane of the telescope. The minimum distance on the focal plane between the images of star 1 and 2 is given by

$$d = f\psi \tag{8}$$

From (7)

$$d = \frac{\lambda f}{a}$$

The resolution limit of the telescope can also be expressed in terms of the cone angle produced when the incident plane wave is focused on the back focal plane of the lens. From the geometry of Fig. 3, the cone angle is given by

$$\tan\theta = \frac{a}{2f}$$

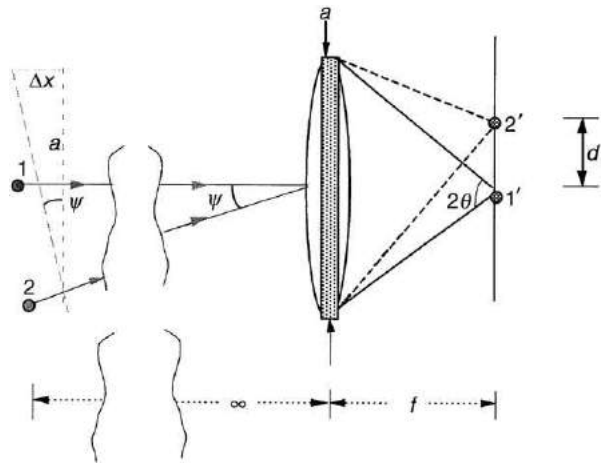


Fig. 3 Telescope resolution. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The separation between the stars 1 and 2 at the back focal plane of the lens in Fig. 3 is thus

$$d = \frac{\lambda}{2\tan\theta} \tag{9}$$

If we assume that the minimum d is 3λ (this is four times the resolution of a typical photographic film at visible wavelengths), then the largest obliquity angle that should be encountered in a visible imaging system is

$$\tan\theta = \frac{\lambda}{2d} = \frac{\lambda}{6\lambda} = \frac{1}{6} = 0.167$$

$$\theta = 9.5^{\circ} \approx 10^{\circ}$$

yielding a value for $\cos\theta = 0.985$ and $(1 + cos\ \theta)/2 = 0.992$. The obliquity factor has only changed by 0.8% over the angular variation of 0° to 10°. The obliquity factor as a function of angle is shown in Fig. 4 for an incident plane wave. The obliquity factor undergoes a 10% change when θ varies from 0° to about 40°; therefore, the obliquity factor can safely be treated as a constant in any optical system that involves angles less than 40°.

While we have shown that it is safe to ignore the variability of C, we still have not assigned a value to the obliquity factor. To find the proper value for C, we will compare the result obtained using (5) with the result predicted by using geometric optics.

We illuminate the aperture $\sum_0$ in Fig. 5 with a plane wave of amplitude α, traveling parallel to the z-axis. Geometrical optics (equivalently the propagation of an infinite plane wave) predicts a field at P_0, on the z-axis, a distance z_0 from the aperture, given by

$$\tilde{E}_{\text{geom}} = \alpha e^{-ikz_0} \tag{10}$$

The area of the infinitesimal at P_1 (the Huygens source) is

$$ds = r\,dr\,d\phi$$

Based on our previous argument, the obliquity factor can be treated as a constant, C, that can be removed from under the integral. The incident wave is a plane wave whose value at $z = 0$ is $E(r) = \alpha$. Using these parameters, the Fresnel integral (5) can be written as

$$\tilde{E}(z_0) = C\alpha \int\int \frac{e^{-ik\cdot R}}{R} r\,dr\,d\phi \tag{11}$$

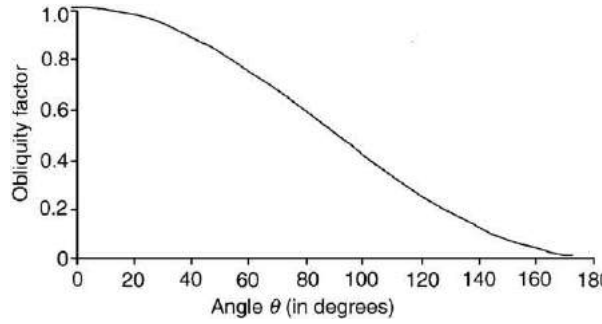


Fig. 4 The obliquity factor as a function of the angle defined in Eq. (6) for an incident plane wave. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

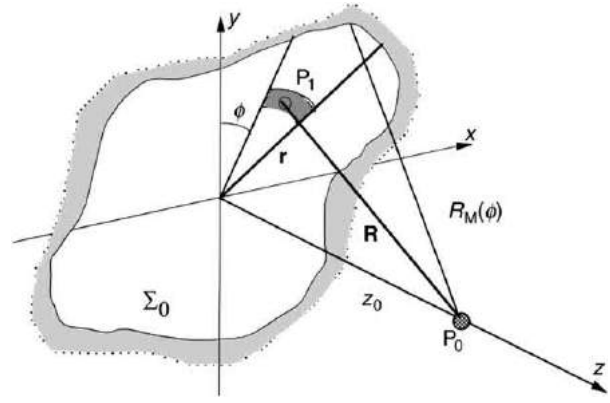


Fig. 5 Geometry for evaluating the constant in the Fresnel integral. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

From the geometry in Fig. 5, the distance from the Huygens source to the observation point is

$$z_0^2 + r^2 = R^2$$

where z_0 is a constant equal to the distance from the observation plane to the aperture plane. The variable of integration can be written in terms of R:

$$r\,\mathrm{d}r = R\,\mathrm{d}R$$

The limits of integration over the aperture extend from $R = z_0$ to the maximum value of R, $R = R_M(\phi)$.

$$\tilde{E}(z_0) = C\alpha \int_0^{2\pi} \int_{z0}^{R_M(\phi)} \mathrm{e}^{-i\mathrm{k}R}\,\mathrm{d}R\,\mathrm{d}\phi \tag{12}$$

The integration over R can now be carried out to yield

$$\tilde{E}(z_0) = \underbrace{\frac{C\alpha}{\mathrm{i}k}\mathrm{e}^{-\mathrm{i}kz_0}\int_0^{2\pi}\mathrm{d}\phi}_{\text{Geometrical Optics}} - \underbrace{\frac{C\alpha}{\mathrm{i}k}\int_0^{2\pi}\mathrm{e}^{-\mathrm{i}kR_M}\,\mathrm{d}\phi}_{\text{Diffraction}} \tag{13}$$

The first integration in (13) is easy to perform and contains the amplitude at the observation point due to geometric optics. The second term may be interpreted as interference of the waves diffracted by the boundary of the aperture. An equivalent statement is that the second term is the interference of the waves scattered from the aperture's boundary, an interpretation of diffraction that was first suggested by Young.

The aperture is irregular in shape, at least on the scale of a wavelength, thus in general, kR_M will vary over many multiples of 2π as we integrate around the aperture. For this reason, we should be able to ignore the second integral in (13) if we confine our attention to the light distribution on the z-axis. After neglecting the second term, we are left with only the geometrical optics component of (13):

$$\tilde{E}(z_0) = \frac{2\pi C}{\mathrm{i}k}\alpha\mathrm{e}^{-\mathrm{i}kz_0} \tag{14}$$

For (14) to agree with the prediction of geometric optics (10) the constant C must be equal to

$$C = \frac{\mathrm{i}k}{2\pi} = \frac{\mathrm{i}}{\lambda} \tag{15}$$

The Huygens–Fresnel integral can be written, using the value for C just derived,

$$\tilde{E}(\mathrm{r}_0) = \frac{\mathrm{i}}{\lambda}\int\int_{\sum}\frac{\tilde{E}_\mathrm{i}(\mathrm{r})}{R}\mathrm{e}^{-\mathrm{i}k\cdot\mathrm{R}}\mathrm{d}s \tag{16}$$

The job of calculating diffraction has only begun with the derivation of the Huygens–Fresnel integral (16). Rigorous solutions exist for only a few idealized obstructions. To allow discussion of the general properties of diffraction, it is necessary to use approximate solutions. To determine what type approximations we can make let's look at the light propagation path shown in Fig. 6.

For the second term in (13) to contribute to the field at point P_0, in Fig. 6, the phase of the exponent that we neglected in (13) must not vary over 2π when the integration is performed over the aperture, i.e.,

$$\Delta = \mathrm{k}\cdot(\mathrm{R}_1 - \mathrm{R}_1' + \mathrm{R}_2 - \mathrm{R}_2') < 2\pi$$

From the geometry of Fig. 6, the two paths from the source, S, to the observation point, P_0, are

$$R_1 + R_2 = \sqrt{z_1^2 + (x_1 + b)^2} + \sqrt{z_2^2 + (x_2 + b)^2}$$

$$R_1' + R_2' = \sqrt{z_1^2 + x_1^2} + \sqrt{z_2^2 + x_2^2}$$

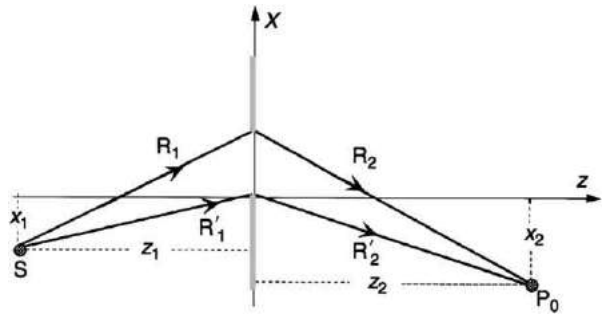


Fig. 6 One-dimensional slit used to establish Fresnel approximations. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

By assuming that the aperture, of width b, is small compared to the distances z_1 and z_2, the difference between these distances, Δ, can be rewritten as a binomial expansion of a square root

$$\sqrt{1+b} = 1 + \frac{b}{2} - \frac{b^2}{8} + \ldots \tag{17}$$

thus

$$\Delta \approx k\left(\frac{x_1}{z_1} + \frac{x_2}{z_2}\right)b + \frac{k}{2}\left(\frac{1}{z_1} + \frac{1}{z_2}\right)b^2 + \ldots \tag{18}$$

If we assume that the second term of this expansion is small, i.e.,

$$\frac{k}{2}\left(\frac{1}{z_1} + \frac{1}{z_2}\right)b^2 \ll 2\pi$$

$$\frac{b^2}{2}\frac{1}{z_1} + \frac{1}{z_2} < \lambda$$

we see that the phase of the wavefront in the aperture is assumed to have a quadratic dependence upon aperture coordinates. Diffraction predicted by this assumption is called Fresnel diffraction.

To discover the details of the approximation consider **Fig. 7**. In the geometry of **Fig. 7**, the Fresnel integral **(16)** becomes

$$\tilde{E}_{P_0} = \frac{i\alpha}{\lambda}\int\int_{\sum} f(x,y)\frac{e^{-ik\cdot(R+R')}}{RR'}\,dx\,dy \tag{19}$$

As we mentioned in our discussion of **(13)** the integral that adds diffractive effects to the geometrical optics field is nonzero only when the phase of the integrand is stationary. For Fresnel diffraction, we can insure that the phase is nearly constant if R does not differ appreciably from Z, or R' from Z'. This is equivalent to stating that only the wave in the aperture, around a point in **Fig. 7** labeled **S**, called the stationary point, will contribute to E_P. The stationary point, **S**, is the point in the aperture plane where the line, connecting the source and observation positions, intersects the plane. Physically, only light propagating over paths nearly equal to the path predicted by geometrical optics (obtained from Fermat's principle) will contribute to E_{P_0}.

The geometry of **Fig. 7** allows the distances R and R' to be written as

$$R^2 = (x-\xi)^2 + (y-\eta)^2 + Z^2 \tag{20}$$

$$R'2 = (x_s - x)^2 + (y_s - y)^2 + Z'2$$

To solve these expressions for R and R', we apply the binomial expansion **(17)** and retain the first two terms:

$$R + R' \approx Z + Z' + \left[\frac{(x-\xi)^2}{2Z} + \frac{(x_s-x)^2}{2Z'}\right] + \left[\frac{(y-\eta)^2}{2Z} + \frac{(y_s-y)^2}{2Z'}\right] \tag{21}$$

In the denominator of **(19)**, R is replaced by Z and R' by Z'. (By making this replacement, we are implicitly assuming that the amplitudes of the spherical waves are not modified by the differences in propagation distances over the area of integration. This is a reasonable approximation because the two propagation distances are within a few hundred wavelengths of each other). In the exponent, we must use **(21)**, because the phase changes by a large fraction of a wavelength, as we move from point to point in the aperture.

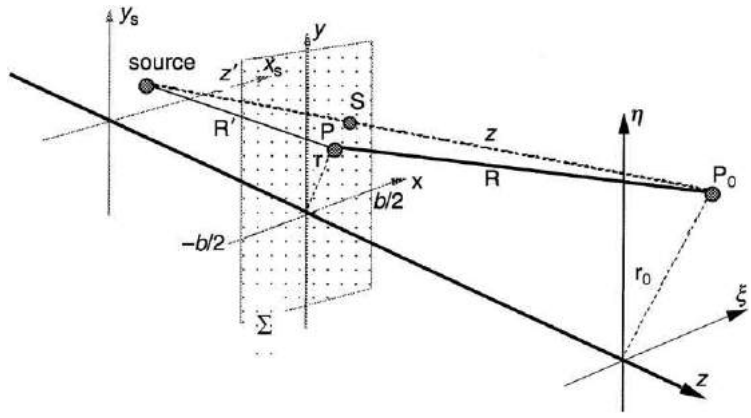


Fig. 7 Geometry for Fresnel diffraction. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

If the wave, incident on the aperture, were a plane wave we can assume $R' = \infty$ and (19) becomes

$$\tilde{E}_{P_0} = \frac{i\alpha}{\lambda}\frac{e^{-ikZ}}{Z}\int\int_{\sum} f(x,y)e^{-\frac{ik}{2Z}[(x-\xi)^2+(y-\eta)^2]}dx\,dy \tag{22}$$

The physical interpretation of (22) states that when a plane wave illuminates the obstruction, the field at point P_0 is a spherical wave, originating at the aperture, a distance Z away from P_0:

$$\frac{e^{-ikZ}}{Z}$$

The amplitude and phase of this spherical wave are modified by the integral containing a quadratic phase dependent on the obstruction's spatial coordinates.

By defining three new parameters,

$$\rho = \frac{ZZ'}{Z+Z'} \quad \text{or} \quad \frac{1}{\rho} = \frac{1}{Z} + \frac{1}{Z'} \tag{23a}$$

$$x_0 = \frac{Z'\xi + Zx_s}{Z+Z'} \tag{23b}$$

$$y_0 = \frac{Z'\eta + Zy_s}{Z+Z'} \tag{23c}$$

the more general expression for Fresnel diffraction of a spherical wave can be placed in the same format as (22). The newly defined parameters x_0 and y_0 are the coordinates, in the aperture plane, of the stationary point, **S**. The parameters in (23) can be used to express, after some manipulation, the spatial dependence of the phase, in the integrand of (19)

$$R + R' = Z + Z' + \frac{(\xi - x_s)^2 + (\eta - y_s)^2}{2(Z+Z')} + \left[\frac{(x-x_0)^2 + (y-y_0)^2}{2\rho}\right]$$

A further simplification to the Fresnel diffraction integral can be made by obtaining an expression for the distance, D, between the source and the observation point:

$$D = Z + Z' + \frac{(\xi - x_s)^2 + (\eta - y_s)^2}{2(Z+Z')} \tag{24}$$

We use this definition of D to also write

$$\frac{1}{ZZ'} = \frac{Z+Z'}{ZZ'}\frac{1}{Z+Z'} \approx \frac{1}{\rho D}$$

Using the parameters we have just defined, we may rewrite (19) as

$$\tilde{E}_{P_0} = \frac{i\alpha}{\lambda\rho D}e^{-ikD}\int\int f(x,y)e^{-\frac{ik}{2\rho}[(x-x_0)^2+(y-y_0)^2]}dxdy \tag{25}$$

With the use of the variables defined in (23), the physical significance of the general expression of the Huygens–Fresnel integral can be understood in an equivalent fashion as was (22). At point P_0, a spherical wave

$$\frac{e^{-ikD}}{D}$$

originating at the source, a distance D away, is observed, if no obstruction are present. Because of the obstruction, the amplitude and phase of the spherical wave are modified by the integral in (25). This correction to the predictions of geometrical optics is called Fresnel diffraction.

Mathematically, (25) is an application of the method of stationary phase, a technique developed in 1887 by Lord Kelvin to calculate the form of a boat's wake. The integration is nonzero only in the region of the critical point we have labeled **S**. Physically, the light distribution, at the observation point, is due to wavelets from the region around **S**. The phase variations of light, coming from other regions in the aperture, are so rapid that the value of the integral over those spatial coordinates is zero. The calculation of the integral for Fresnel diffraction is complicated because, if the observation point is moved, a new integration around a different stationary point in the aperture must be performed.

Rectangular Apertures

If the aperture function, $f(x,y)$, is separable in the spatial coordinates of the aperture, then we can rewrite (25) as

$$\tilde{E}_{P_0} = \frac{i\alpha}{\lambda\rho D}e^{-ikD}\int_{-\infty}^{\infty} f(x)e^{-ig(x)}dx\int_{-\infty}^{\infty} f(y)e^{-ig(y)}dy$$

$$\tilde{E}_{P_0} = A[C(x) - iS(x)][C(y) - iS(y)]$$

A represents the spherical wave from the source, a distance D away,

$$A = \frac{i\alpha}{2D} e^{-ikD}$$

If we treat the aperture function as a simple constant, C and S are integrals of the form

$$C(x) = \int_{x_1}^{x_2} \cos[g(x)]dx \tag{26a}$$

$$S(x) = \int_{x_1}^{x_2} \sin[g(x)]dx \tag{26b}$$

The integrals, $C(x)$ and $S(x)$, have been evaluated numerically and are found in tables of Fresnel integrals. To use the tabulated values for the integrals, **(26)** must be written in a general form

$$C(w) = \int_0^w \cos\left(\frac{\pi u^2}{2}\right) du \tag{27}$$

$$S(w) = \int_0^w \sin\left(\frac{\pi u^2}{2}\right) du \tag{28}$$

The variable u is an aperture coordinate, measured relative to the stationary point, $\mathbf{S}(x_0,y_0)$, in units of

$$\sqrt{\frac{\lambda\rho}{2}}$$

$$u = \sqrt{\frac{2}{\lambda\rho}}(x - x_0) \quad \text{or} \quad u = \sqrt{\frac{2}{\lambda\rho}}(y - y_0) \tag{29}$$

The parameter w in **(27)** and **(28)** specifies the location of the aperture edge relative to the stationary point **S**. The parameter w is calculated through the use of **(29)** with x and y replaced by the coordinates of the aperture edge.

The Cornu Spiral

The plot of $S(w)$ versus $C(w)$, shown in **Fig. 8**, is called the Cornu spiral in honor of M. Alfred Cornu (1841–1902) who was the first to use this plot for graphical evaluation of the Fresnel integrals.

To use the Cornu spiral, the limits w_1 and w_2 of the aperture are located along the arc of the spiral. The length of the straight line segment drawn from w_1 to w_2 gives the magnitude of the integral. For example, if there were no aperture present, then, for the x-dimension, $w_1 = -\infty$ and $w_2 = \infty$. The length of the line segment from the point $(-1/2,1/2)$ to $(1/2,1/2)$ would be the value of E_{P_0}, i.e., $\sqrt{2}$. An identical value is obtained for the y-dimension, so that

$$\tilde{E}_{P_0} = 2A = \frac{\alpha e^{-ikD}}{D}$$

This is the spherical wave that is expected when no aperture is present.

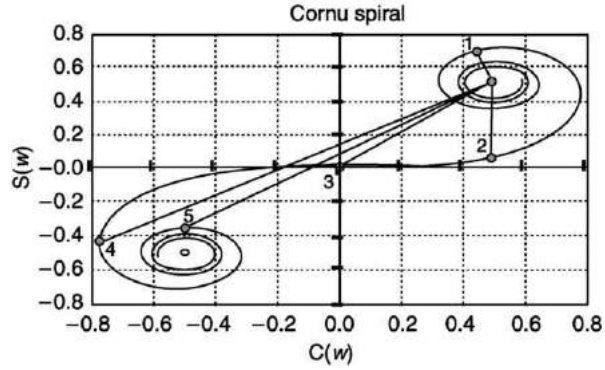


Fig. 8 Cornu spiral obtained by plotting the values for $C(w)$ versus $S(w)$ from a table of Fresnel integrals. The numbers indicated along the arc of the spiral are values of w. The points 1–5 correspond to observation points shown in **Fig. 9** and are used to calculate the light intensity diffracted by a straight edge. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

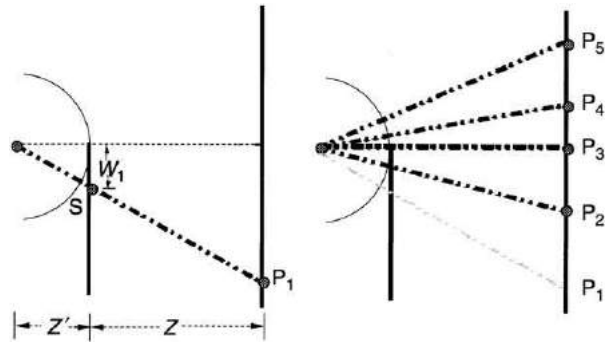


Fig. 9 A geometrical construct to determine w_1 for the stationary point S associated with five different observation points. The values of w are then used in Fig. 8 to calculate the intensity of light diffracted around a straight edge. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

As an example of the application of the Fresnel integrals, assume the diffracting aperture is an infinitely long straight edge reducing the problem to one dimension:

$$f(x,y) = \begin{Bmatrix} 1 & x \geq x_1 \\ 0 & x < x_1 \end{Bmatrix}$$

The value of E_{P_0}, in the y-dimension is from above $\sqrt{2}$ since $w_1 = -\infty$ and $w_1 = \infty$. In the x-dimension, the value of w at the edge is

$$w_{x_1} = \sqrt{\frac{2}{\lambda\rho}}(x_1 - x_0)$$

We will assume that the straight edge blocks the negative half-plane so that $x_1 = 0$. The straight edge is treated as an infinitely wide slit with one edge located at infinity so that $w_2 = \infty$ in the upper half-plane. When the observation point is moved, the coordinate x_0 of **S** changes and the origin of the aperture's coordinate system moves. New values for the w's must therefore be calculated for each observation point.

Fig. 9 shows the assumed geometry. A set of observation points, P_1–P_5, on the observation screen are selected. The origin of the coordinate system in the aperture plane is relocated to a new position (given by the position of **S**) when a new observation point is selected. The value of w_1, the position of the edge with respect to **S**, must be recalculated for each observation point. The distance from the origin to the straight edge, w_1, is positive for P_1 and P_2, zero for P_3, and negative for P_4 and P_5.

Fig. 8 shows the geometrical method used to calculate the intensity values at each of the observation points in Fig. 9 using the Cornu spiral. The numbers labeling the straight line segments in Fig. 8 are associated with the labels of the observation points in Fig. 9.

To obtain an accurate calculation of Fresnel diffraction from a straight edge, a table of Fresnel integrals provides the input to the following equation

$$I_P = I_0\left\{\left[\frac{1}{2} - C(w_1)\right]^2 + \left[\frac{1}{2} - S(w_1)\right]^2\right\}$$

where $I_0 = 2A^2$. Table 1, shows the values extracted from the table of Fresnel integrals to find the relative intensity at various observation points. The result obtained by using either method for calculating the light distribution in the observation plane, due to the straight edge in Fig. 9, is plotted in Fig. 10. The relative intensities at the observation points depicted in Fig. 9 are labeled on the diffraction curve of Fig. 10.

Fresnel Zones

Fresnel used a geometrical construction of zones to evaluate the Huygens–Fresnel integral. The Fresnel zone is a mathematical construct, serving the role of a Huygens source in the description of wave propagation. Assume that at time t, a spherical wavefront from a source at P_1 has a radius of R'. To determine the field at the observation point, P_0, due to this wavefront, a set of concentric spheres of radii, Z, $Z+(\lambda/2)$, $Z+2(\lambda/2)$,…,$Z+j(\lambda/2)$,… are constructed, where Z is the distance from the wavefront to the observation point on the line connecting P_1 and P_0 (see Fig. 11). These spheres divide the wavefront into a number of zones, $\zeta_1,\zeta_2,\ldots,\zeta_j\ldots$, called Fresnel zones, or half-period zones.

We treat each zone as a circular aperture illuminated from the left by a spherical wave of the form

$$\tilde{E}_j(\mathrm{R}') = \frac{A e^{-\mathrm{ik}\cdot\mathrm{R}'}}{R'} = \frac{A e^{-\mathrm{ikR}'}}{R'}$$

Table 1 Fresnel integrals for straight edge

w_1	$C(w_1)$	$S(w_1)$	I_P/I_0	P_j
∞	0.5	0.5	0	
2.0	0.4882	0.3434	0.01	
1.5	0.4453	0.6975	0.021	P_1
1.0	0.7799	0.4383	0.04	
0.5	0.4923	0.0647	0.09	P_2
0	0	0	0.25	P_3
−0.5	−0.4923	−0.0647	0.65	
−1.0	−0.7799	−0.4383	1.26	P_4
−1.2	−0.7154	−0.6234	1.37	
−1.5	−0.4453	−0.6975	1.16	
−2.0	−0.4882	−0.3434	0.84	P_5
−2.5	−0.4574	−0.6192	1.08	
$-\infty$	−0.5	−0.5	1.0	

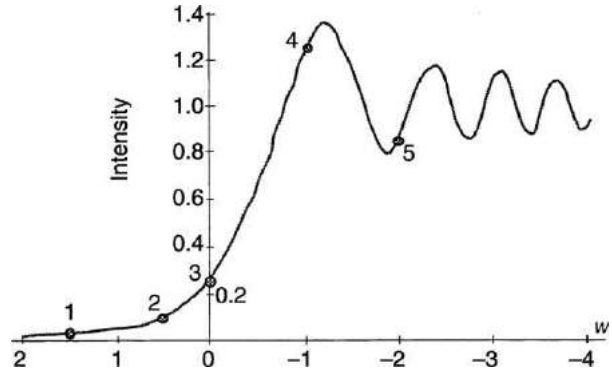


Fig. 10 Light diffracted by a straight edge with its edge located at $w=0$. The points labeled 1–5 were found using the construction in Fig. 9 to obtain w. This was then used in Fig. 8 to find a position on the Cornu spiral. The length of the lines shown in Fig. 8 from the (1/2,1/2) point to the numbered positions led to the intensities shown. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

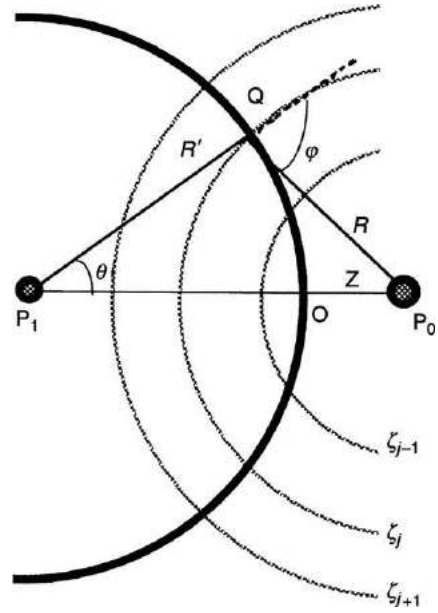


Fig. 11 Construction of Fresnel zones observed from a position P_0 on a spherical wave originating from a source at point P_1. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

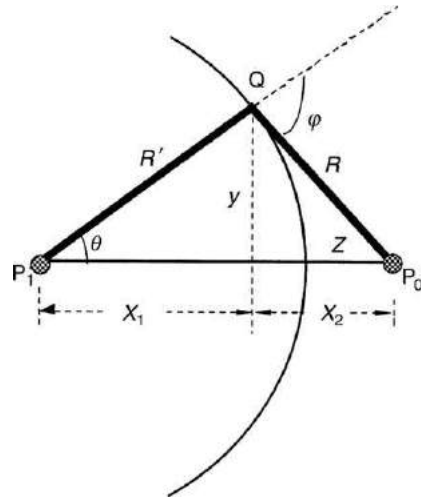


Fig. 12 Geometry for finding the relationship between R and R' yielding **Eq. (32)**. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

R' is the radius of the spherical wave. The field at P_0 due to the jth zone is obtained by using **(5)**

$$\tilde{E}_j(P_0) = \frac{A}{R'} e^{-ik\cdot R'} \int\int_{\zeta_j} C(\varphi) \frac{e^{-ik\cdot R}}{R} ds \tag{30}$$

For the integration over the jth zone, the surface element is

$$ds = R' 2\sin\theta d\theta d\phi \tag{31}$$

The limits of integration extend over the range

$$Z + (j-1)\frac{\lambda}{2} \leq R \leq Z + j\frac{\lambda}{2}$$

The variable of integration is R; thus, a relationship between R' and R must be found. This is accomplished by using the geometrical construction shown in **Fig. 12**.

A perpendicular is drawn from the point Q on the spherical wave to the line connecting P_1 and P_0, in **Fig. 12**. The distance from the source to the observation point is $x_1 + x_2$ and the distance from the source to the plane of the zone is the radius of the incident spherical wave, R'. The distance from P_1 to P_0 can be written

$$x_1 + x_2 = R' + Z$$

The distance from the observation point to the zone is

$$\begin{aligned} R^2 &= x_2^2 + y^2 = [(R' + Z) - x_1]^2 + (R'\sin\theta)^2 \\ &= R' 2 + (R' + Z)^2 - 2R'(R' + Z)\cos\theta \end{aligned}$$

The derivative of this expression yields

$$R\, dR = R'(R' + Z)\sin\theta d\theta \tag{32}$$

Substituting **(32)** into **(31)** gives

$$ds = \frac{R'}{R' + Z} R dR d\varphi \tag{33}$$

The integration over ϕ is accomplished by rotating the surface element about the P_1P_0 axis. After integrating over ϕ between the limits of 0 and 2π, we obtain

$$\tilde{E}_j(P_0) = \frac{2\pi A}{R' + Z} e^{-ik\cdot R'} \int_{Z+(j-1)\frac{\lambda}{2}}^{Z+\frac{j\lambda}{2}} e^{-ik\cdot R} C(\varphi) dR \tag{34}$$

We assume that R', $Z \gg \lambda$ so that the obliquity factor is a constant over a single zone, i.e., $C(\varphi) = C_j$. The obliquity factor is not really a constant as we earlier assumed but rather a very slowly varying parameter of j, changing by less than two parts in 10^4 when 'j' changes by 500 (for $Z = 1$ meter and $\lambda = 500$ nm). In fact it only changes by 2 parts in 10^{-7} across one zone. We are safe in applying the assumption that the obliquity factor is a constant over a zone and this allows the integral in **(34)** to be calculated:

$$\tilde{E}_j(P_0) = \frac{2\pi i C_j A}{k(R' + Z)} e^{-ik(R' + Z)} e^{-ikj\frac{\lambda}{2}} \left[1 - e^{-ik\frac{\lambda}{2}}\right] \tag{35}$$

Using the identity $k\lambda = 2\pi$ and the definition for the distance between the source and observation point **(24)** modified for this geometry, $D = R' + Z$ **(35)** can be simplified to

$$\tilde{E}_j(\mathrm{P}_0) = 2\mathrm{i}\lambda(-1)^j \frac{C_j A}{D} \mathrm{e}^{-\mathrm{ikD}} \tag{36}$$

The physical reasons for the behavior predicted by **(36)** are quite easy to understand. The distance from P_0 to a zone changes by only $\lambda/2$ as we move from zone to zone and the area of a zone is almost a constant, independent of the zone number; thus, the amplitudes of the Huygens wavelets from each zone should be approximately equal. The alternation in sign, from zone to zone, is due to the phase change of the light wave from adjacent zones because the propagation paths for adjacent zones differ by $\lambda/2$.

To find the total field strength at P_0 due to N zones, the collection of Huygens' wavelets is added:

$$E(\mathrm{P}_0) = \sum_{j=1}^{N} \tilde{E}_j(\mathrm{P}_0) = \frac{2\mathrm{i}\lambda A}{D} e^{-\mathrm{ikD}} \sum_{j=1}^{N} (-1)^j C_j \tag{37}$$

To evaluate the sum, the elements of the sum are regrouped and rewritten as

$$-\sum_{j=1}^{N} (-1)^j C_j = \frac{C_1}{2} + \left(\frac{C_1}{2} - C_2 + \frac{C_3}{2}\right) + \left(\frac{C_3}{2} - C_4 + \frac{C_5}{2}\right) + \ldots$$

Because the C's are very slowly varying functions of j, even out to 500 zones, we are justified in setting the quantities in parentheses equal to zero. With this approximation, the summation can be set equal to one of two values, depending upon whether there is an even or odd number of terms in the summation

$$-\sum_{j=1}^{N} (-1)^j C_j = \left\{ \begin{array}{ll} \frac{1}{2}(C_1 + C_N) & N \text{ odd} \\ \frac{1}{2}(C_1 - C_N) & N \text{ even} \end{array} \right\}$$

For very large N, the obliquity factor approaches zero, $C_N \to 0$, as was demonstrated in **Fig. 4**. Thus, the theory has led us to the conclusion that the total field produced by an unobstructed wave is equal to one half the contribution from the first Fresnel zone, i.e., $E = E_1/2$. Stating this result in a slightly different way, we obtain a surprising result – the contribution from the first Fresnel zone is twice the amplitude of the unobstructed wave!

The zone construction can be used to analyze the effect of the obstruction of all or part of a zone. For example, by constructing a circular aperture with a diameter equal to the diameter of the first Fresnel zone, we have just demonstrated that it is possible to produce an intensity at the point P_0 equal to four times the intensity that would be observed if no aperture were present. To analyze the effects of a smaller aperture, we subdivide a half-period zone into a number of subzones such that there is a constant phase difference between each subzone. The individual vectors form a curve know as the vibrational curve. **Fig. 13** shows a vibrational curve produced by the addition of waves from nine subzones of the first Fresnel zone. The vibrational curve in **Fig. 13** is an arc with the appearance of a half-circle. If the radius of curvature of the arc were calculated, we would discover that it is a constant, except for the contribution of the obliquity factor,

$$\rho = \frac{\lambda R'}{R' + Z} C(\varphi)$$

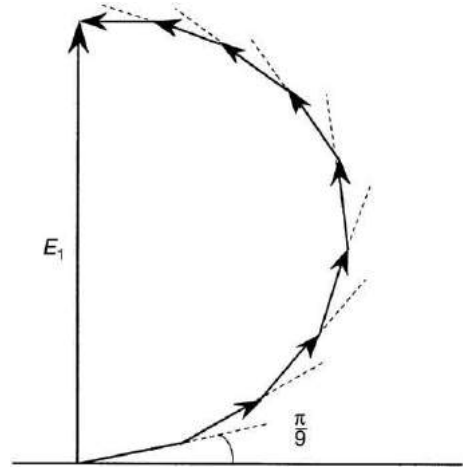


Fig. 13 Vector addition of waves from nine subzones constructed in the first Fresnel zone. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

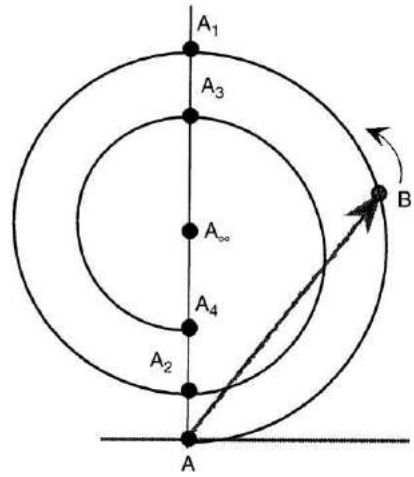


Fig. 14 Vibration curve for determining the Fresnel diffraction from a circular aperture. The change of the diameter of the half-circles making up the spiral has been exaggerated for easy visualization. The actual changes are one part in 10^{-5}. The position of point B is determined by the aperture size. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Because the obliquity factor for a single zone is a constant to 2 parts in 10^{-7}, the radius of curvature of the vibration curve can be considered a constant over a single zone. If we let the number of subzones approach infinity, the vibrational curve becomes a semicircle whose chord is equal to the wavelet produced by zone one, i.e., E_1. If we subdivide additional zones, and add the subzone contributions, we create other half-circles whose radii decrease at the same rate as the obliquity factor. The vibrational curve for the total wave is a spiral, constructed of semicircles which converges to a point halfway between the first half-circle, see **Fig. 14**. The length of the vector from the start to the end of the spiral is $E=E_1/2$, as we derived above. When this same construction technique is applied to a rectangular aperture, the vibrational curve generated is the Cornu spiral.

Circular Aperture

The vector addition technique, described in **Fig. 13**, can be used to evaluate Fresnel diffraction, at a point P_0, from a circular aperture and yields the intensity distribution along the axis of symmetry of the circular aperture. The zone concept will also allow a qualitative description of the light distribution normal to this axis.

To develop a quantitative estimate of the intensity at an observation point on the axis of symmetry of the circular aperture, we construct a spiral, **Fig. 14**, to represent the Fresnel zones of a spherical wave incident on the aperture. The point B on the spiral shown in **Fig. 14** corresponds to the portion of the spherical wave unobstructed by the screen. The length of the chord, AB, represents the amplitude of the light wave at the observation point P_0. As the diameter of the aperture increases, B moves along the spiral, in a counterclockwise fashion, away from A. The first maximum occurs when B reaches the point labeled A_1 in **Fig. 14**; the aperture has then uncovered the first Fresnel zone. At this point, the amplitude is twice what it would be with no obstruction. Four times the intensity!

If the aperture's diameter continues to increase, B reaches the point labeled A_2 in **Fig. 14** and the amplitude is very nearly zero; two zones are now exposed in the aperture. Further maxima occur when an odd number of zones are in the aperture and further minima when an even number of zones are exposed. **Fig. 15** shows an aperture containing four exposed Fresnel zones. The amplitude at the observation point would correspond to the chord drawn from A to A_4 in **Fig. 14**.

The aperture diameter can be fixed and the observation point P_0 can move along, or perpendicular to, the axis of symmetry of the circular aperture. As P_0 is moved away from the aperture, along the symmetry axis, i.e., as Z increases, the radius of the Fresnel zones increase without limit. For small values of n, the radius of the nth zone can be approximated by

$$r_n \approx \sqrt{nZ\lambda} \tag{38}$$

At Z_{max}, the light intensity is a maximum, given by the chord length from A to A_1 in **Fig. 14**. If $\lambda=500$ nm and $a=0.5$ mm then this maximum occurs when $Z=0.5$ m

$$Z_{max} = \frac{a^2}{\lambda} \tag{39}$$

If we start at Z_{max} and move toward the aperture, along the axis, as Z decreases in value, a point will be reached when the intensity on the axis becomes a minimum. The value of Z where the first minimum in intensity is observed is equal to

$$Z_{min} = \frac{a^2}{2\lambda}$$

In **Fig. 14** the chord would extend from A to A_2.

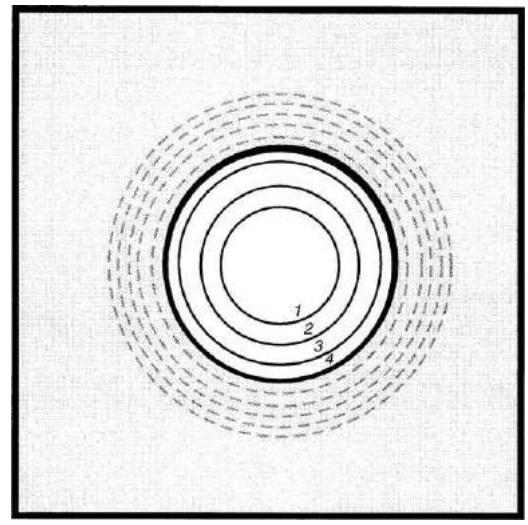


Fig. 15 Aperture with four Fresnel zones exposed. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

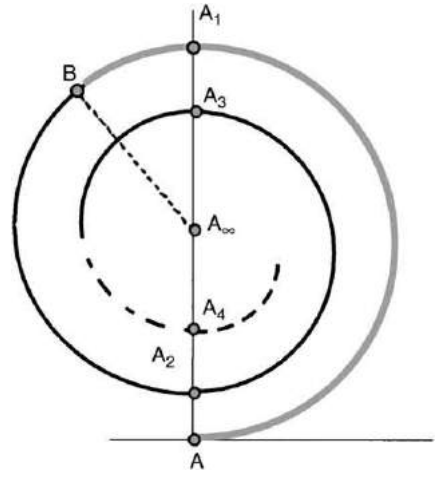


Fig. 16 Vibration curve for opaque disk. The shaded region makes no contribution because it is associated with the portion of the wave obstructed by the opaque object. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

As the observation point, P_0, is moved along the axis toward the aperture, and Z assumes values less than Z_{min}, the point B in Fig. 14 spirals inward, toward the center of the spiral and the intensity cycles through maximum and minimum values. The cycling of the intensity, as the observation point moves toward the aperture, will not continue indefinitely. At some point, the field on the axis will approach the field observed without the aperture because the distance between Z_{max} and Z_{min} shrinks to the wavelength of light making the intensity variation unobservable.

For values of Z that exceed Z_{max} of (39), the aperture radius, a, will be smaller than the radius of the first zone, and Fraunhofer diffraction will be observed because the aperture contains only one zone and the phase in the aperture is a constant.

Opaque Screen

If the screen containing a circular aperture, of radius a, is replaced by an opaque disk of radius a, the intensity distribution on the symmetry axis, behind the disk, is found to be equal to the value that would be observed with no disk present. This prediction was first derived by Poisson, to demonstrate that wave theory was incorrect; however, experimental observation supported the prediction and verified the theory.

We construct a spiral, shown in Fig. 16, similar to the one for the circular aperture in Fig. 14. The point B on the spiral represents the edge of the disk. The shaded portion of the spiral from A to B does not contribute because that portion of the wave is covered by the disk and the zones, associated with that portion of the wave, cannot be seen from the observation point.

The amplitude at P_0 is the length of the chord from B to A_∞ shown in Fig. 16. If the observation point moves toward the disk, then B moves along the spiral toward A_∞. There is always intensity on the axis for this configuration, though it slowly decreases until it reaches zero when the observation point reaches the disk; this corresponds to point B reaching point A_∞ on the spiral. Physically, zero intensity occurs when the disk blocks the entire light wave. There are no maxima or minima observed as the disk diameter, a, increases or as the observation distance changes. If the observation point is moved perpendicular to the symmetry axis, a set of concentric bright rings are observed. The origin of these bright rings can be explained using Fresnel zones, in a manner similar to the one used to explain the bright rings observed in a Fresnel diffraction pattern from a circular aperture.

Zone Plate

In the construction of Fresnel zones, each zone was assumed to produce a Huygens wavelet, out of phase with the wavelets produced by its nearest neighbors. If every other zone were blocked, then there would be no negative contributions to (37). The intensity on-axis would be equal to the square of the sum of the amplitudes produced by the in-phase zones – exceeding the intensity of the incident wave. An optical component made by the obstruction of either the odd or the even zones could therefore be used to concentrate the energy in a light wave.

The boundaries of the opaque zones, used to block out-of-phase wavelets, are seen from (38) to increase as the square root of the integers. An array of opaque rings, constructed according to this prescription is called a zone plate, see Fig. 17.

A zone plate will perform like a lens with focal length

$$f = \pm \frac{r_1^2}{\lambda} \tag{40}$$

The zone plate, shown in Fig. 17, will act as both a positive and negative lens. What we originally called the source can now be labeled the object point, O, and what we called the observation point can now be labeled the image point, I. The light passing through the zone plate is diffracted into two paths, labeled C and D in Fig. 17. The light waves, labeled C, converge to a real image point I. For these waves the zone plate performs the role of a positive lens with a focal length given by the positive value of (40). The light waves, labeled D in Fig. 17, appear to originate from the virtual image point labeled I. For these waves the zone plate performs the role of a negative lens with a focal length given by the negative value of (40).

The zone plate will not have a single focus, as is the case for a refractive optical element, but rather will have multiple foci. As we move toward the zone plate, from the first focus, given by (40), the effective Fresnel zones will decrease in diameter. The zone plate will no longer obstruct out-of-phase Fresnel zones and the light intensity on the axis will decrease. However, additional maxima, of the on-axis intensity, will be observed at values of Z for which the first zone plate opening contains an odd number of zones. These positions can also be labeled as foci of the zone plate; however, the intensity at each of these foci will be less than the intensity at the primary focus.

Lord Rayleigh suggested that an improvement of the zone plate design would result if, instead of blocking every other zone, we shifted the phase of alternate zones by 180°. The resulting zone plate, called a phase-reversal zone plate, would, more efficiently, utilize the incident light. R.W. Wood was the first to make such a zone plate. Holography provides an optical method of

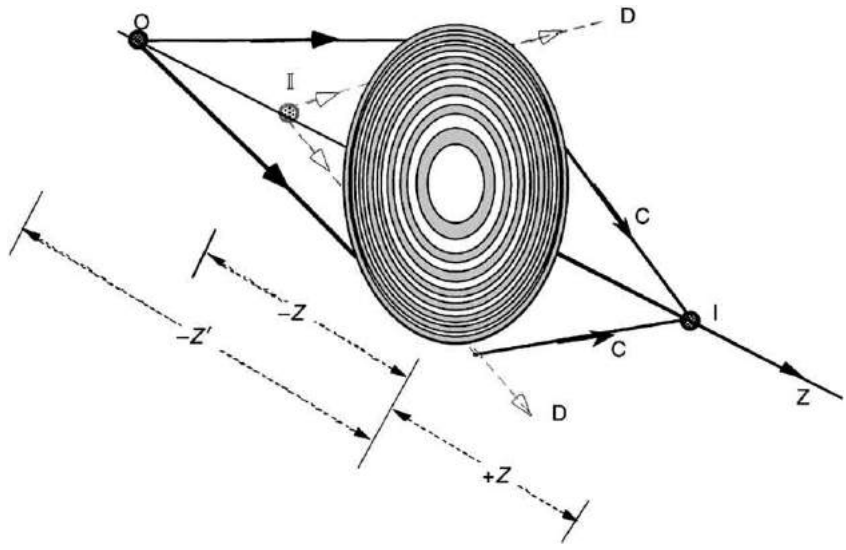


Fig. 17 A zone plate acts as if it were both a positive and a negative lens. Light from the object, O, is diffracted into waves traveling in both the D and the C directions. The light traveling in the C direction produces a real image of O at I. The light traveling in the D direction appears to originate from a virtual image at I. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

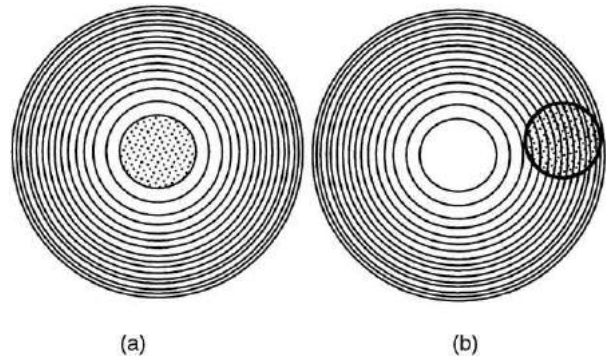


Fig. 18 (a) A set of Fresnel zones has been constructed about the optical path taken by some light ray. If the optical path of the ray is varied over the cross-hatched area shown in the figure, then the optical path length does not change. This cross-hatched area is equal to the first Fresnel zone and is described as the neighborhood of the light ray. (b) The neighborhood defined in (a) is moved so that it surrounds an incorrect optical path for a light ray. We see that this region of space would contribute no wave amplitude at the observation point because of the destructive interference between the large number of partial zones contained in the neighborhood. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

constructing the phase-reversal zone plates in the visible region of the spectrum. Semiconductor lithography has been used to produce zone plates in the x-ray region of the spectrum.

The resolving power of a zone plate is a function of the number of zones contained in the plate. When the number of zones exceeds 200, the zone plate's resolution approaches that of a refractive lens.

Fermat's Principle

The Fresnel zone construction provides physical insight into the interpretation of Fermat's principle which states that, if the optical path length of a light ray is varied in a neighborhood about the true path, there is no change in path length. By constructing a set of Fresnel zones about the optical path of a light ray, we can discover the physical meaning of a neighborhood.

The rules for constructing a Fresnel zone require that all rays passing through a given Fresnel zone have the same optical path length. The true path will pass through the center of the first Fresnel zone constructed on the actual wavefront. A neighborhood must be the area of the first Fresnel zone, for it is over this area that the optical path length does not change. **Fig. 18(a)** shows the neighborhood, about the true optical path, as a cross-hatched region equal to the first Fresnel zone.

Light waves do travel over wrong paths but we do not observe them because the phase differences for those waves that travel over the 'wrong' paths are such that they destructively interfere. By moving the neighborhood, defined in the previous paragraph as the first Fresnel zone, to a region displaced from the true optical path, we can use the zone construction to see that this statement is correct. In **Fig. 18(b)** the neighborhood is constructed about a ray that is improper according to Fermat's principle. We see that this region of space would contribute no energy at the observation point because of destructive interference between the large number of partial zones contained in the neighborhood.

This leads us to another interpretation of diffraction. If an obstruction is placed in the wavefront so that we block some of the normally interfering wavelets, we see light not predicted by Fermat's principle, i.e., we see diffraction.

Further Reading

Backer, B.B., Copson, E.J., 1969. The Mathematical Theory of Huygens' Principle. London: Oxford University Press.
Born, M., Wolf, E., 1970. Principles of Optics. New York: Pergamon Press.
Guenther, R.D., 1990. Modern Optics. New York: Wiley.
Haus, H.A., 1984. Waves and Fields in Optoelectronics. Englewood Cliffs, NJ: Prentice-Hall.
Hecht, E., 1998. Optics, 3rd edn. Reading, MA: Addison-Wesley.
Klein, M.V., 1970. Optics. New York: Wiley.
Rubinowicz, A., 1984. In: Wolf, E. (Ed.), Tahe Miyamoto–Wolf diffraction wave,Progress in Optics, vol. IV. .
Sayanagi, I., 1967. Pinhole Imagery. Journal of the Optical Society of America. 57, 1091–1099.
Sears, F.W., 1949. Optics. Reading, MA: Addison-Wesley.
Sommerfeld, A., 1964. Optics. New York: Academic Press.
Stegliani, D.J., Mittra, R., Semonin, R.G., 1967. Resolving power of a zone plate. Journal of the Optical Society of America 57, 610–613.
Stone, J.M., 1963. Radiation and Optics. New York: McGraw-Hill.

Early History of Quantum Electronics

CH Townes, University of California, Berkeley, CA, USA

All the ideas essential to making a laser were known before 1930, but there was no operating laser before 1960. So why didn't the laser come sooner? There are several reasons. One important impediment to the laser invention was that a combination of ideas from quantum mechanics and from electrical engineering was needed, and these two fields were not well mixed in the early days. Another is that, while some physicists recognized that amplification could occur if there was a population inversion of states, they did not consider the coherence of such amplifications and did not recognize its importance or usefulness. It was just something that in principle could happen, but was not very interesting. What was required was a combination of physics and electrical engineering thinking and recognition of the coherence of such amplification. In addition the importance of such a development had to be visualized and recognized in a way that it led to very devoted work towards its achievement.

It is striking that lasers grew out of the study of microwave spectra of molecules, for which engineering and quantum mechanics were both important and which helped orient thinking in the appropriate direction. That this origin was not accidental is convincingly demonstrated by the fact that three independent ideas for amplification by stimulated emission were generated about 1950. They were from Joe Weber, at the University of Maryland, from Nikolai Basov and Alexander Prokhovov at the Soviet Academy, and myself at Columbia University. Weber primarily wanted to point out the possibility, but didn't try to do it. In addition, his numbers were a bit off and no practical system was suggested. Basov and Prokhovov actually worked towards microwave amplification using a beam of molecules, as did I.

That stimulated amplification was recognized early, but not thought through, is illustrated by the fact that Prof. Richard Tolman, a theoretical physicist, wrote a discussion in 1924 of the net absorption of light by molecules, pointing out in particular that induced emission counteracted absorption and noting that if there were more molecules in the upper than in the lower state there could be 'negative absorption'. But, he wrote, 'This would usually be very small'. The Russian physicist Vitaly Ginsburg wrote me, after the maser and laser had appeared, that his professor, S.M. Levi, had been well aware of such effects back in the 1930s and had told him 'create an overpopulation at higher atomic levels and you will obtain an amplifier; the whole trouble is that it is difficult to create a substantial overpopulation of levels'.

The German physicist F.G. Houtermans said to me that in 1932, when told by a colleague of an unusual light intensity in a gaseous discharge, he had thought it might be a 'photon avalanche', i.e. multiplication of photons by stimulated emission.

Another Russian physicist, V.A. Fabricant, wrote a thesis in 1939 in which he discussed absorption and emission of light radiation in a gas and looked for 'negative absorption', or amplification. He did not discuss coherence or a resonant cavity, and was not able to achieve any amplification so his work was quickly forgotten. None of these early mentions of stimulated emission proposed how to actually get amplification, that it would be useful, nor clearly noted its coherence. Tolman did, however, write in 1927 that, 'We should expect radiation induced by an external field to be coherent with the radiation associated with that field'. I know of no proposal to actually make use of such amplification, until those made by microwave spectroscopists in the early 1950s.

Another clear indicator that even in the 1950s physicists and engineers did not think amplification by stimulated emission was particularly interesting nor useful is that during the 2½ years when Jim Gordon, Herb Zeiger, and I were working on trying to obtain microwave amplification (the maser), a large number of scientists visited my laboratory, saw what we were doing, but no one bothered to also try to obtain such amplification. After the maser worked, it hit the newspapers and then very quickly became a popular and intense field of interest for a number of physicists.

My own drive to produce oscillators by stimulated emission came from my strong interest in obtaining sources of waves shorter than the few millimeters wavelength which could be produced by electronic devices, in order to extend the high resolution study of molecular spectra to wavelengths shorter than microwaves, down into the far infrared. My students and I worked on several possible schemes for producing waves shorter than those produced by klystrons or magnetrons – frequency multiplication by nonlinearities, electronic beams passing over surfaces of solid materials with resonances, and anything else I could think of. None worked very well.

In early 1950, I was asked by the Office of Naval Research to form and chair a committee which would examine possible research towards obtaining wavelengths down to one millimeter and shorter. I chose outstanding scientists and engineers in a variety of fields which might touch on this problem. We met together, and visited many pertinent laboratories and individuals interested in such research. Nothing very promising seemed to turn up. But we wanted, of course, to at least provide a report summarizing the situation as we saw it. Our last meeting was April 26, 1951, in Washington, D.C. Worrying over our lack of success, I woke up early before the meeting. Breakfast was not ready, so I walked over to nearby Franklin Park, sat on a bench in front of beautiful blooming azaleas, and puzzled over why neither I nor the Committee had found any promising solutions.

I went over all the ideas I had had previously. Molecules, of course, can produce high frequencies. But I had previously concluded, I thought wisely and rigorously, that one could not obtain intense radiation from them because radiation intensity was a function of temperature, and the temperature could not be very high without destroying the molecules. Suddenly I realized that they did not need to have a temperature in the usual sense – they need not be in temperature equilibrium. There could be

more molecules in an upper than a lower excitation state, which could in principle produce indefinitely intense radiation. I pulled a piece of paper out of my pocket and wrote down the equations and numbers for such a case, using a molecule beam sent into a resonant cavity and ammonia molecules with which I was very familiar. My equations said one could get enough excited molecules, low enough loss in a resonant cavity, and it would work! Why hadn't I thought of it before?!

Back at Columbia, where I worked, and about 4 months later, the graduate student Jim Gordon agreed to work on trying to obtain such an oscillator using a beam of ammonia. I assured him that if it didn't work he could modify the experiment to do interesting spectroscopy and thus complete a thesis. But we both thought he had a good chance of making it work. And a young post doc working with me, Herb Zeiger, joined the effort.

Actually, I had first thought of obtaining stimulated emission from a molecular beam back in 1948, but simply as a demonstration of physical principles rather than as a useful amplification. Also, about a year later a young post doc, J.W. Trischka, working on molecular beams with Professors Rabi and Kusch, had also thought of demonstrating stimulated emission. We talked about it together, and he decided it wasn't worthwhile just demonstrating the effect because it wouldn't really prove any new physics. Neither of us at that time had recognized the real point and the possibility of useful amplification, which is one of the reasons mentioned above that the idea was delayed as long as it was.

It is perhaps important to emphasize again and to illustrate how out-of-the-way the use of stimulated emission was at that time for physicists, and how distant stimulated emission was from engineers. While we were working on the ammonia beam maser, Prof. L.H. Thomas, an outstanding physicist known for the 'Thomas Effect', frequently would run into me in the hallway at Columbia University and say that I didn't understand, and that my proposed ammonia oscillator could not work. However, I never got a clear explanation from him of what it was I didn't understand. And after we had been working on the ammonia maser system for about 2 years (a more-or-less normal time for a student thesis project), the Physics Department Chairman, Prof. Polycarp Kusch, and the previous chairman, Prof. I.I. Rabi, came into my office to object. They were excellent physicists, both received Nobel Prizes, and both were experts on molecular beams. They sat down in my office and said 'Look, Charlie, that is not going to work. We know it won't work and you know it won't work. You are wasting departmental money, and must stop!' Other people had also questioned what I was doing, frequently in particular whether the stimulated radiation would be coherent. So I had already thought over the situation many times. I had full notes from the quantum mechanics course I took as a student back in 1938, and could derive from them a proof of coherence. I felt also that I knew the quantitative numbers, such as the molecular beam intensity and the possible 'Q', or loss, in the cavity resonator, very well. I was also by then an Associate Professor, and the department chairman could not fire me simply because of stupidity. I replied to Rabi and Kusch 'No, I believe it has a reasonable chance of working and I'm going to continue!' Annoyed, they stomped out of the room.

About 2 months after the Rabi/Kusch incident, Jim Gordon dashed into the classroom where I was lecturing and said loudly 'It's working!' Most of the class then went up to the lab to see the new device. Rabi and Kusch were not against me; even though they were outstanding physicists, they probably just didn't quite comprehend the device. A couple of months after its successful operation, Kusch more or less apologized by saying 'Well, I should have realized that you probably know more about what you are doing than I do.'

After the maser successfully operated, there were other incidences showing the lack of appropriate focus of physicists on stimulated emission. I was a friend of Aage Bohr, the son of Niels Bohr, and in that connection was visiting him in Denmark. Niels Bohr asked me what research I was presently doing, so I told him about our new oscillator, the maser, and its remarkably pure frequency. He looked at me and said 'Oh no, that's not possible. You must be misunderstanding something.' I emphasized again what it was really doing, but he still seemed not to believe it could function that way. I presumed he was thinking in terms of the uncertainty principle and the finite time of passage of a molecule through the cavity, though I never was quite sure just why he felt it impossible. A similar thing happened shortly after that at a cocktail party in Princeton, where John Von Neumann asked the same question – what was my research at the moment? After telling him about the maser oscillator and the purity of frequency he reacted very similarly 'Oh no,' he said, 'That can't be right. You must be misunderstanding something.' After arguing a little more, he left to get another drink. Fifteen minutes later he came back, saying 'Hey, you are right!' He had understood, and wanted to talk much more about the maser and of possibly using excited semiconductors. Only after his death did I learn from his notebooks that he had considered exciting electrons in semiconductors with neutrons from a reactor, and had written Edward Teller about whether some experiments might be done with this to obtain intense light. However, he had not considered coherence, and Teller was apparently not interested enough to respond, so the matter was dropped.

The above account reminds me a bit of the amusing comments of Arthur Clark on 'Change'. He writes:

'People go through four states before any revolutionary development:

1. It's nonsense, don't waste my time
2. It's interesting, but not important
3. I always said it was a good idea
4. I thought of it first'.

It's clear that the world of physics was thinking very little in the direction of masers or lasers, and that many preconceptions as well as lack of interest stood in the way. My own experience with engineering at Bell Telephone Labs during World War II, in designing radar and electronic systems, plus my intense interest in obtaining short-wave oscillators, were clearly important in bringing me to the right ideas.

As masers became very interesting to the physics community and the field grew rapidly, my engineering experiences continued to be of help. I was well acquainted with the theoretical examination of noise in vacuum tube amplifiers by various individuals at Bell Labs, and recognized that the maser could provide much more sensitive amplification than could common electronic amplifiers, where discrete electronic charges produce the basic noise. On sabbatical in France, I worked on electron spin masers with Jean Combrisson and Arnold Honig who had the appropriate equipment. And then in Japan, Koichi Shimoda, Hidetoshi Takahashi, and I wrote a theoretical paper on the basic quantum noise of maser (or laser) amplification. The theory showed that maser amplifiers of microwaves should be about 100 times more sensitive than the existing electronic ones.

After 2–3 years of maser experiments and development I felt I wanted to move on to the shorter wavelengths for which I had generated the maser idea. Although I had first tried the idea at microwave frequencies because that seemed the easiest way to test out the general idea and the result had been exciting, I still wanted those shorter wavelengths. I had not come up with any great ideas of just how to get to much shorter wavelengths, which is why I waited several years after the maser worked before moving on. However, in 1957, 3 years after the first successful operation of the maser, I decided it was high time to simply figure out what was the best way I could imagine to move on into the infrared region and do it. A number of physicists had concluded that of course one couldn't make masers work at much shorter wavelengths, certainly not in the visible region, because spontaneous emission becomes so much faster as the wavelength is shortened and adequate inversion of population was not practical. But that was intuition, not quantitative science. As I wrote down equations for what might be done to move towards shorter wavelengths using a model of atomic or molecular excitation by radiation and a reasonably high Q resonator, it became clear that it was quite practical to move on even into the visible region. That was exciting! Why hadn't I, or someone else, looked at it carefully and quantitatively before that time?

I was at the time consulting at Bell Labs, with the assignment to spend a day there every 2 weeks and just talk with the Lab's scientific personnel. Since my former post doc and now brother-in-law Arthur Schawlow was then at Bell Labs, I of course talked with him. On telling him of my ideas for an 'optical maser', using a resonant cavity and excitation of atoms with optical radiation, he said he too was interested in that, and we decided to work together to optimize a system. It was Art who then suggested use of a Fabry–Perot resonator rather than the cavity with large holes I had used for a model, and that was an excellent addition. Why I did not think of that is a mystery, but Art had done his thesis at the University of Toronto in Fabry–Perot spectroscopy, and that might have been why the thought came to him.

Since Art Schawlow was participating, I decided the patent for the new 'optical maser' should belong to the Bell Labs (I already claimed ownership of the basic maser patent, which covered all wavelengths). Hence we recognized that the new idea must be kept confidential until Bell Labs lawyers had worked out and applied for an appropriate patent. This delay in public information sheds some additional light on how the scientific and technical world was thinking at the time. I had written down my original ideas for an 'optical maser' in my notebook and had it witnessed by my student Joe Giordmaine at Columbia University. I had also talked with a Columbia student Gordon Gould because he had been doing research with an intense light source and I wanted to know how much intensity he had in order to be sure I could get enough excited atoms and provide an oscillator at these short wavelengths. Except for these two persons, neither Art Schawlow nor I told anyone outside of Bell Labs about the 'optical maser', or laser idea until after Bell Labs had properly prepared its patent, which was about 11 months after my first notebook entry. For that entire time, no one has produced a record of any thoughts about extending the maser to optical wavelengths except Gordon Gould, who entered some ideas in his notebook about 1 month after I talked with him about the possibility of an optical maser, and 2 months after my original notebook entry. His notes were later to be the source of a long patent case.

The striking observation is that no one outside of Bell Labs except Gordon Gould, to whom I had explained my ideas, seems to have written or noted down anything about extending masers to these shorter wavelengths during the 11 months from my recognition that it could well be done until after the Schawlow and Townes paper on how to do it became available. After that there was considerable excitement and a number of ideas.

It is also noteworthy that, because of low general interest and competition, I did not publish a paper on how a maser might be made, but waited for publication until we demonstrated its operation, about 3 years after the idea had arrived. But after the maser worked the field became exciting and competitive. Hence, Schawlow and I thought we surely should publish a theoretical paper establishing the idea before taking the time to make a laser work. And indeed, there was much competition to build the first laser.

I myself helped a couple of my graduate students start work on trying to build an 'optical maser' or laser. But at about that time (1958), I was urged to undertake a job in Washington as Vice President and Director of Research of the Institute for Defense Analysis, an organization put together largely by the presidents of several universities to try to help advise the government on matters of science and technology. Sputnik had gone up the year before, and everyone was worried about the position of the U.S. with respect to Communist Russia, which seemed to be ahead particularly in some areas important to defense. I decided I should try to help, and accepted a two-year appointment in Washington. I recognized that this seriously distracted my attention from developing a laser quickly, but knew there were many others working towards lasers and so such devices would certainly be developed and our doing it just at Columbia University was neither critically important nor highly probable.

The field of masers and 'optical masers' or lasers was becoming so exciting and a bit scrambled that The Office of Naval Research asked me if I would organize a meeting on the subject. I did, with the help of a committee of many distinguished people in the field or closely related science and technology. And it was in a meeting of the Committee that we christened the field with the name 'Quantum Electronics'. This first international meeting on the subject was at Shawanga Lodge in New York State in September 1959. It made an occasion for the Russians Basov and Prokhorov to visit the United States (and my lab and home), and

was about 6 months before Ted Maiman made the first working laser. There were many interesting discussions of masers and their coming operation at optical wavelengths.

It is significant to note that, while the maser grew out of basic research in universities (with Russian work at the Russian Academy) and industry had little to do with early masers, all the first lasers were made in industrial laboratories. This illustrates the sociology and the strengths and weaknesses of industrial and of academic laboratories. While masers (and from them lasers) originated from research on microwave spectroscopy of molecules, industrial laboratories believed the field of microwave spectroscopy had little to offer in the way of commercial results. Because of equipment available and the interest of physicists in industry, the field was initiated and pursued immediately after World War II in commercial laboratories – by myself at Bell Labs, my friends at the RCA Labs, at Westinghouse, and at General Electric. For lack of interest in industry, such work was soon shut down, except at Bell Labs, and it moved to universities. Bell Labs generously allowed me to continue such work, although they wanted me to do some 'more useful' engineering. Once the field had obvious commercial possibilities, commercial laboratories began to support it well and the clear importance of masers and lasers made industry very interested. Really interested industry can more easily put strong new resources and push harder on a field than can academic laboratories, where money has to be granted and professors have a variety of other assignments. The first laser was made to work, of course, by Ted Maiman at Hughes Research Laboratories. Ted had been a student of Willis Lamb at Stanford, and there worked on radio spectroscopy. The second type of laser, rather similar to Maiman's but using a different material, was made to work at the General Electric Laboratories by Peter Sorokin and Mirek Stevenson. Sorokin had been a student of Bloembergen at Harvard in microwave or radio spectroscopy and Stevenson was one of my students in microwave spectroscopy at Columbia University. The next type of laser, and one I particularly admire, was made by Ali Javan, William Bennett, and Don Herriott. Javan had been a student with me in microwave spectroscopy at Columbia University, Bill Bennett a student in the molecular beam group at Columbia working on radio spectroscopy, and Don Harriott was an optics specialist. All of these originators, with the exception of Herriott, had been working at universities in the field which originated the idea and were recently hired by industry. The next important laser, involving semiconductors, was invented by Robert Hall and collaborators at General Electric Labs. Note that every one of these early lasers was created in industry.

After the first laser was operated, my own students at Columbia quickly turned from trying to make a laser to using lasers to explore more physics, the normal university function. And I am delighted that masers and lasers have provided such excellent tools for research, as well as for commercial and medical applications.

After a few years of exploring new physics with lasers, I decided that since there were many excellent scientists doing such work, I should move into fields which it seemed to me were being relatively neglected. I moved to the University of California at Berkeley to look for molecules in interstellar space by microwave spectroscopy, and to do infrared astronomy. Very soon after initiating work at Berkeley, one of my students, Albert Cheung, not only discovered the first polyatomic molecules in space, he found powerful water masers. A while before our discovery of water and identification of its radiation as due to maser action, it had been deduced that some microwave radiation of OH must be due to maser action. And since then, many, many masers due to a wide variety of molecules have been found in astronomical sources as well as a few powerful lasers. Since deviations from thermal equilibrium are common in the very thin gases excited by powerful sources in space, we should have expected this, but didn't. And these masers in space could have been easily detected with radio technology available back in the 1930s if anyone had searched the microwave spectrum carefully.

Clearly, masers and lasers could have been discovered and used much sooner than they actually were. We simply were neither thinking nor looking in the right directions. And this raises the natural question – what important and more-or-less obvious ideas are we missing now because of our lack of imaginative exploration?

Further Reading

Basov, N., Prokhorov, A., 1954. Application of molecular beams to radio spectroscopic studies of rotation spectra of molecules. J. Experimental. Theoretical Physics U.S.S.R. 27, 431–438.

Bloembergen, N., 1956. Proposal for a new type of solid state maser. Physical Review 104, 324–327.

Cheung, A., Rank, D., Townes, C., Thornton, D., Welch, J., 1969. Detection of water in interstellar regions by its microwave radiation. Nature 121, 626–628.

Combrisson, J., Honig, A., Townes, C., 1956. Utilization de la resonance de spins electroniques pour realiser un oscillateur ou un amplificateur en hyperfrequencies. Comptes Rendues 242, 2451.

Gordon, J., Zeiger, H., Townes, C., 1954. Molecular microwave oscillator and new hyperfine spectrum of NH_3. The Physical Review 95, 282–284.

Hall, R., Fenner, G., Kingsley, G., Soltys, J., Carlson, R., 1962. Coherent light emission from GaAs junction. Physical Review Letters 9, 366–368.

Javan, A., Bennett, W., Herriot, D., 1961. Population inversion and continuous optical maser oscillation in a gas discharge containing a He–Ne mixture. Physical Review Letters 6, 106–110.

Knowles, S., Mayer, C., Cheung, A., Rank, D., Townes, C., 1969. Spectra, variability, size, and polarization of H_2O microwave emission sources in the galaxy. Science 163, 1055–1057.

Maiman, T., 1960. Stimulated optical radiation in ruby. Nature 187, 493–494.

Perkins, F., Gold, T., Salpeter, E., 1966. Maser action in interstellar OH. Ap. J. 145, 361–366.

Schawlow, A., Townes, C., 1958. Infrared and optical masers. Physical Review 112, 1940–1949.

Shimoda, K., Takahashi, H., Townes, C., 1957. Fluctuations in amplification of quanta with application to maser amplifiers. Journal of the Physical Society of Japan 12, 686–700.

Sorokin, P., Stevenson, M., 1960. Stimulated infrared emission from trivalent uranium. Physical Review Letters 5, 557.

Tolman, R., 1924. Weak quantization. The Physical Review 24, 287–295.

Tolman, R., 1927. Statistical mechanics with applications to physics and chemistry. The Chemical Catalog Co., New York. pp. 169.

Weber, J., 1953. Amplification of microwave radiation by substances not in thermal equilibrium. Trans. Inst. Radio Engineers Professional Group on Electron Devices 3, 1–4.

Lenses and Mirrors

A Nussbaum, University of Minnesota, Minneapolis, MN, USA

Introduction

Lenses are carefully shaped pieces of transparent material used to form images. The term 'transparent' does not necessarily mean that the lens is transmitting visible light. The element germanium, well-known in the transistor industry, is a light-gray opaque crystal which forms images with infrared radiation, and provides the optics for night-vision instruments. The lenses used in cameras, telescopes, microscopes, and other well-known applications, are made from special glasses or – more recently – from plastics. Hundreds of different optical glass formulas are listed in the manufacturers' catalogues. Plastics have come into use in recent years because it is now possible to mold them with high precision. The way in which lenses form images has been studied for several hundred years, resulting in an enormous and very complicated literature. In this article, we show how this theory can be simplified by introducing the restriction that lenses are perfect. This limitation provides an easy introduction to the image forming process.

The Cardinal Points and Planes of an Optical System

The location, size and orientation of an image are items of information which a designer needs to know. The nature of the image can be determined by a simple model of an optical system, based on two experimental observations. The first is well-known and is pictured in **Fig. 1**. Parallel rays of light passing through a double convex lens (a simple magnifying glass, for example) should meet at the focal point or focus, designated as F'. For rays from the right side of the lens, another such point F exists at an equal distance from the lens. The other experiment uses the lens to magnify (**Fig. 2(a)**). The amount of magnification decreases as the lens is brought closer to 'ISAAC NEWTON' (**Fig. 2(b)**), and when the lens touches the paper, object and image are approximately the same size. The locations of object and image corresponding to equality in size are called the unit or principal points H and H', respectively, and the set of four points F, F', H, and H' are the cardinal points. Therefore, the planes through these points normal to the axis of the lens are the cardinal planes. **Fig. 3** shows an object whose image we can now determine. The coordinate orientation in this figure may appear strange at first. It is customary in optics to designate the lens axis as OZ, and the other two axes form a right-handed system. The x-axis then lies in the plane of the figure; the y-axis, coming out of the paper and perpendicular to it, is not shown. The object, of height x, is placed to the left of the focal point F. The cardinal points occur in the order F, H, H', F', since we expect – based on the experiment corresponding to **Fig. 2** – that the unit points are very close to the lens. Another fact to be demonstrated later is that H and H' are actually inside the lens. Two rays are shown leaving the point P at the top of the object. The ray which is parallel to the axis strikes the unit plane through H and continues to the right, completely missing the lens. The ray then meets the unit plane through H' at a point which must be a distance x from the axis. To understand why, we recognize that if the object were relocated so as to lie on the object-space unit plane, the ray emerging at the other unit plane will be at a distance x from the axis; as required by the definition of these planes. We then assume that the ray is bent or refracted at the unit plane at H' and proceeds to, and beyond, the focal point F'. The second ray, from P through the focal point F, is easily traced since the lens is indifferent to the direction of the light. Assume temporarily that we know the location of the image point P'. Let a parallel ray leave this point and travel to the left. The procedure just given indicates that this ray will strike the unit plane at H', emerge at the other unit plane, be refracted so as to pass through F and reach the object point P. Then reversing the direction of this ray, it will start at

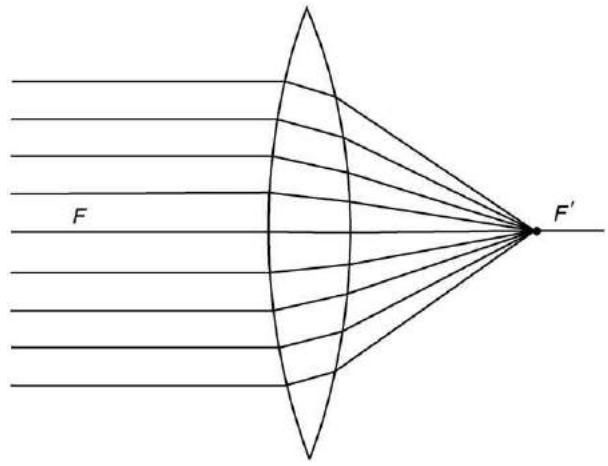


Fig. 1 Focal points of a double convex lens.

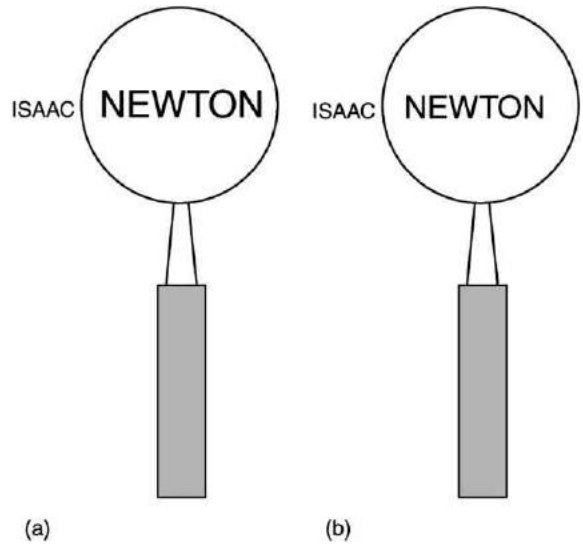


Fig. 2 Magnification by a double convex lens. The magnification in (a) is reduced when the lens is brought closer to the image (b).

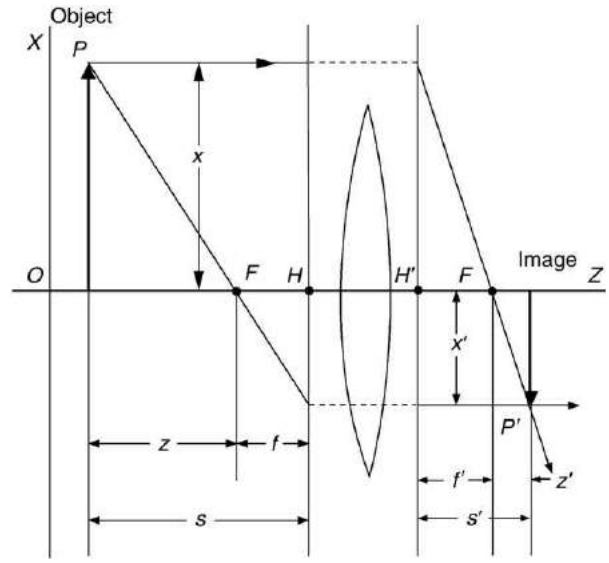


Fig. 3 Ray tracing using cardinal points and planes.

P, emerge at the unit plane H' and travel parallel to the axis, intersecting the other ray at P'. The distance between F and H is called the object space focal length f, with f' being the corresponding image space quantity. Simple geometry, making use of the similar triangles and the distances indicated in this diagram, can be used to derive the Newton lens equation or the equivalent Gauss lens equation. Detailed derivations of these two equations will be found in either of the sources listed in the Further Reading at the end of this article. It is not usually made clear in physics texts that these equations apply only to the special case of a single, thin lens, and we shall not apply them here. The image in this figure is real, inverted, and reduced. That is, it is smaller than the object, oriented in the opposite direction, and can be projected onto a screen located at the image position. This result can be verified by using an ordinary magnifier to form the image of a distant light source on a sheet of white paper.

Paraxial Matrices

To consider how light rays behave in optical systems, we introduce the index of refraction n of a material medium, defined as the ratio of the velocity of light c in a vacuum to its velocity v in that medium, or:

$$n = \frac{c}{v} \tag{1}$$

The velocity of light in air is approximately the same as it is in a vacuum. A light ray crossing the interface between air and a denser medium such as glass will be bent towards the normal to the surface. This is the phenomenon of refraction and the amount of bending is governed by Snell's law, which has the form:

$$n\sin\theta = n'\sin\theta' \tag{2}$$

where n and n' are the indices of refraction of the two media and θ and θ' are the angles as measured from the normal. **Fig. 4** shows a ray of light leaving an object point P and striking the first surface of a lens at point P_1. It is refracted there, and proceeds to the second surface. All rays shown lie in the z, x-plane or meridional plane, which passes through the symmetry axis OZ. The amount of refraction is specified by Snell's law, **Eq. (2)**, which we shall now simplify. The sine of an angle can be approximated by the value of the angle itself, if this value is small when expressed in radians (less than 0.1 rad), and Snell's law simplifies to

$$n_1\theta_1 = n_1'\theta_1' \tag{3}$$

This is called the paraxial form of Snell's law; the word paraxial means 'close to the axis'. It turns out, however, that the Snell's law angles are not convenient to work with, and we eliminate them with the terms:

$$\theta_1 = \alpha_1 + \phi, \quad \theta_1' = \alpha_1' + \phi \tag{4}$$

where α_1 is the angle which the incident ray makes with the axis OZ, α'_1 is the corresponding angle for the refracted ray, and ϕ is the angle which the radius r_1 (the line from C to P_1) of the lens surface makes at the center of curvature C. This particular angle can be specified as

$$\sin\phi = \frac{x_1}{r_1} \tag{5}$$

but using the paraxial approximation simplifies this to

$$\phi = \frac{x_1}{r_1} \tag{6}$$

Substituting for the angles gives

$$n_1'\alpha_1' = \frac{n_1 - n_1'}{r_1} + n_1\alpha_1 \tag{7}$$

Notice that the distance from the point P_1 to the axis is labeled as either x_1 or x'_1. This strange notation leads to the trivial relation

$$x_1 = x_1' \tag{8}$$

and this can be combined with **Eq. (7)** to obtain the matrix equation:

$$\begin{pmatrix} n_1'\alpha_1' \\ x_1' \end{pmatrix} = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix} \tag{9}$$

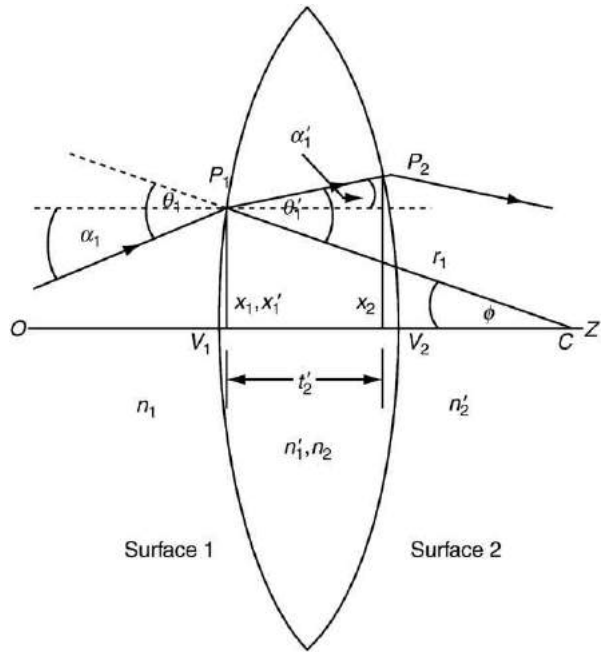


Fig. 4 Ray passing through a double convex lens.

where the constant k_1 is called the refracting power of surface 1 and is defined as

$$k_1 = \frac{n_1' - n_1}{r_1} \tag{10}$$

To review the procedure for matrix multiplication, the first element $n'_1\alpha'_1$ in the one-column product matrix is calculated by multiplying the upper left-hand corner of the 2×2 matrix with the first element of the one-column matrix to its right, obtaining 1 ($n_1\alpha_1$), and to this is added the product $-k_1x_1$ of the upper right-hand element in the 2×2 matrix and the second member of the one-column matrix. This square matrix is called the refraction matrix R_1 for surface 1, defined as

$$R_1 = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} \tag{11}$$

Next, we look at what happens to the ray as it travels from surface 1 to surface 2. As it goes from P_1 to P_2, its distance from the axis becomes

$$x_2 = x_1' + t_1' \tan\alpha_1' \tag{12}$$

or using the paraxial approximation for small angles:

$$x_2 = x_1' + t_1'\alpha_1' \tag{13}$$

Another approximation we shall use is to regard t'_1 as being equal to the distance between the lens vertices V_1 and V_2. Using the identity:

$$\alpha_2 = \alpha_1' \tag{14}$$

leads to a second matrix equation:

$$\begin{pmatrix} n_2\alpha_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t_1'/n_1' & 1 \end{pmatrix} \begin{pmatrix} n_1'\alpha_1' \\ x_1' \end{pmatrix} \tag{15}$$

where the 2×2 translation matrix T_{21} is defined as

$$T_{21} = \begin{pmatrix} 1 & 0 \\ t_1'/n_1' & 1 \end{pmatrix} \tag{16}$$

To get an equation which combines a refraction followed by a translation, we take the one-column matrix on the left side of **Eq. (9)** and substitute it into **Eq. (15)** to obtain

$$\begin{pmatrix} n_2\alpha_2 \\ x_2 \end{pmatrix} = T_{21}R_1 \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix} \tag{17}$$

Note that the 2×2 matrices appear in right to left order and that the multiplication of two square matrices is an extension of the rule given above. This equation gives the location and slope of the ray which strikes surface 2 after a refraction and a translation. To continue the ray trace at surface 2, we introduce a second refraction matrix R_2 by expressing k_2 in terms of the two indices at this surface and the radius. Then **Eq. (17)** is extended to give the relation:

$$\begin{pmatrix} n_2'\alpha_2' \\ x_2' \end{pmatrix} = R_2T_{21}R_1 \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix} \tag{18}$$

This process can obviously be applied to any number of components. The product of the three 2×2 matrices in the above equation is known as the system matrix S_{21} and it completely specifies the effect of the lens on the incident ray passing through it. It is also written as

$$S_{21} = R_2T_{21}R_1 = \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \tag{19}$$

where the four quantities a, b, c, and d are known as the Gaussian constants. They are used extensively in specifying the behavior of a lens or an optical system. We now state the sign and notation conventions which are necessary for the application of paraxial matrix methods:

1. Light normally travels from left to right.
2. The optical systems we deal with are symmetrical about the z-axis. The intersections of the refracting or reflecting surfaces with this axis are the vertices and are designated in the order encountered as V_1, V_2, etc.
3. Positive directions along the axes are measured from the origin in the usual Cartesian fashion, so that horizontal distances (that is, along the z-axis) are positive if measured from left to right. Angles are positive when measured up from the z-axis.
4. Quantities associated with the incident ray are unprimed; those for the refracted ray are primed.
5. A subscript denotes the associated surface.
6. If the center of curvature of a surface is to its right, the radius is positive, and vice versa.
7. There is a special rule for mirrors to be explained below.

Using the Gaussian Constants

Fig. 5 shows the double convex lens of Fig. 4 with the assumed locations of the cardinal points indicated. These positions, which we shall now determine accurately, are at distances designated as l_F, l'_F, l_H, and l'_H and are measured from the associated vertex. The object position can be called t, a positive quantity which is measured to the right from object to the first vertex, or it can be called t_1 if measured in the opposite direction. For the first choice, the matrix specifying the translation from object to lens will have the quantity t/n_1 in its lower left-hand corner. However, it is both logical and convenient to use the first vertex as the reference point. Hence, we replace t/n_1 in the translation matrix with the quantity $-t_1/n_1$, and remember to specify t_1 as a negative number when calculating the image position. The equation connecting object and image is obtained by starting with this matrix, multiplying it by the system matrix of Eq. (19), and finally by a translation matrix corresponding to the translation t'_2 from lens to image to obtain:

$$\begin{pmatrix} n'_2\alpha'_2 \\ x' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t'_2/n'_2 & 1 \end{pmatrix} \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \times \begin{pmatrix} 1 & 0 \\ -t_1/n_1 & 1 \end{pmatrix} \begin{pmatrix} n_1\alpha_1 \\ x \end{pmatrix} \tag{20}$$

where α or α_1 is the angle that the ray from the top of the object makes with the z-axis. This equation takes the initial value of the inclination α and of the position x of the ray leaving the object, and determines the final values, α' and x' at the image. The product of the three 2×2 matrices on the right-hand side can be consolidated into a single matrix called the object-image matrix $S_{P'P}$. Then Eq. (20) can be written as

$$\begin{pmatrix} n'_2\alpha'_2 \\ x' \end{pmatrix} = S_{P'P} \begin{pmatrix} n_1\alpha_1 \\ x \end{pmatrix} \tag{21}$$

and this matrix has the complicated form:

$$S_{P'P} = \begin{pmatrix} b + \dfrac{at_1}{n_1} & -a \\ \dfrac{bt'_2}{n'_2} + \dfrac{at_1t'_2}{n_1n'_2} - d - \dfrac{ct_1}{n_1} & c - \dfrac{at'_2}{n'_2} \end{pmatrix} \tag{22}$$

If we put this matrix into the previous equation, we obtain an unsatisfactory result: the value of x' will depend on the angle α_1 made by the incident ray, as can be seen by multiplying the two matrices on the right side. The ratio of x' to x is called the magnification m; that is

$$m = \frac{x'}{x} \tag{23}$$

A perfect image can be formed only if the magnification is determined solely by the object distance and the constants of the lens. To eliminate this difficulty, the lower left-hand element in the matrix, Eq. (22) is required to be equal to zero, and the magnification is then:

$$m = \frac{x'}{x} = c - \frac{at'_2}{n'_2} \tag{24}$$

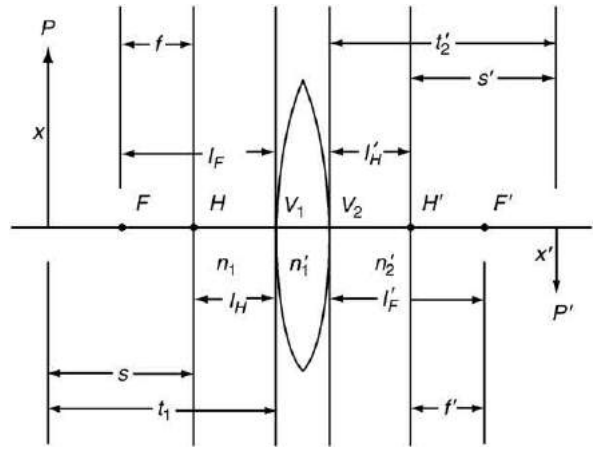


Fig. 5 Definitions of cardinal plane locations.

The determinant of this matrix is unity, since it is the product of three matrices whose individual determinants are unity. It follows that

$$\frac{1}{m} = b + \frac{at_1}{n_1} \tag{25}$$

so that

$$\begin{pmatrix} n_2'\alpha_2' \\ x' \end{pmatrix} = \begin{pmatrix} 1/m & -a \\ 0 & m \end{pmatrix} \begin{pmatrix} n_1\alpha \\ x \end{pmatrix} \tag{26}$$

Using the fact that the lower left-hand element of the matrix **Eq. (22)** is zero shows that

$$\frac{t_2'}{n_2'} = \frac{d + ct_1/n_1}{b + at_1/n_1}, \qquad \frac{t_1}{n_1} = \frac{d - bt_2'/n_2'}{-c + at_2'/n_2'} \tag{27}$$

This equation connects the object distance t_1 with the image distance t'_2, both quantities being measured from the appropriate vertex. This relation is known as the generalized Gauss lens equation. It applies to all lens systems, no matter how complicated.

We can determine the location of the unit planes by letting $m = 1$. This t'_2 in **Eq. (27)** is the location l'_H of the image space unit plane, and it becomes

$$l_H' = \frac{n_2'(c-1)}{a} \tag{28}$$

This relation expresses the location of the unit plane on the image side as the distance from V_2 to H'. Similarly, the location of H with respect of V_1 is

$$l_H = \frac{n_1(1-b)}{a} \tag{29}$$

To locate the focal planes, consider a set of parallel rays coming from infinity and producing a point image at F'. The terms containing t_1 in **Eq. (27)** are much larger than d or b, and this relation becomes

$$l_F' = \frac{n_2'c}{a} \tag{30}$$

Letting t'_2 become infinite in the right-hand half of **Eq. (27)**, the location of F is given by the equation

$$l_F = \frac{-n_1 b}{a} \tag{31}$$

The focal lengths f and f' shown in **Fig. 5** were defined earlier as the distances between their respective focal and unit planes. Hence

$$f' = l_F' - l_H' = \frac{n_2'}{a} \tag{32}$$

and

$$f = l_F - l_H = \frac{-n_1}{a} \tag{33}$$

If we let $n_1 = n'_2 = 1$ for air, these become

$$-f = f' = \frac{1}{a} \tag{34}$$

The Gaussian constant a is thus the reciprocal of the focal length in air. Even when the lens is not in air, it is still true that $f' = -f$ if the lens is in a single medium. It is also true regardless of whether or not the lens is symmetric.

It is customary in optics to work with a consistent set of units, such as centimeters or millimeters, and the sign conventions used for x and y are the standard Cartesian rules. As mentioned above, curved surfaces which open to the right have a positive radius and vice versa. As an example, let a double convex lens have radii of 2.0 and 1.0, respectively, an index of refraction of 1.5, and a thickness of 0.5. These quantities are then denoted as

$$\begin{aligned} &r_1 = +2.0, \qquad r_2 = -1.0, \qquad t_1' = t_2 = +0.5 \\ &n_1 = 1.0, \qquad n_1' = 1.5 = n_2, \qquad n_2' = 1.0 \end{aligned} \tag{35}$$

By **Eq. (10)**:

$$k_1 = \frac{n_1' - n_1}{r_1} = \frac{1.5 - 1.0}{2.0} = 0.25 \tag{36}$$

and

$$k_2 = \frac{1.0 - 1.5}{-1.0} = 0.50 \tag{37}$$

The system matrix S_{21} is

$$\begin{aligned} S_{21} &= R_2 T_{21} R_1 \\ &= \begin{pmatrix} 1 & -0.50 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0.5/1.5 & 1 \end{pmatrix} \begin{pmatrix} 1 & -0.25 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0.83 & -0.71 \\ 0.33 & 0.92 \end{pmatrix} \end{aligned} \tag{38}$$

and we note that the determinant is

$$(0.83)(0.92) - (0.71)(-0.33) = 1.00 \tag{39}$$

This property of the system matrix provides a very useful check on the accuracy of matrix multiplications. The locations of the four cardinal points can be determined from the formulas given above and their locations are shown in the scale drawing of **Fig. 6**. Since we know how to locate the cardinal points of a lens, we are in a position to ray trace in a precise manner, using the method of **Fig. 3** and adding a third ray. After locating F, H, H', and F', we no longer need the lens; the cardinal planes with the proper separation are fully equivalent to the lens they replace. This is very much like what electrical engineers do when they replace a complicated circuit with a black box; this equivalent circuit behaves just like the original. Ray tracing becomes more complicated when we deal with diverging lenses (**Fig. 7**), but the procedure gives above still applies. Parallel rays spread apart and do not come to a focus on the right-hand side of the lens. However, if the refracted rays are extended backwards, these extensions will meet as indicated. This behavior can be verified quantitatively by using the formulas developed above. It will be found that F and F' have exchanged places, but H and H' retain their original positions inside the lens. If an object is placed between the focal point F and the first vertex of a convex lens (**Fig. 8**), then our usual procedure produces two rays which do not intersect in image space. The ray from the object which is parallel to the axis is refracted downward so that it does not intersect the other ray in image space. However, the extensions to the left of these two rays meet to form an image which is erect, magnified, and virtual. This is the normal action of a magnifying lens; the image can be seen but cannot be projected onto a screen, unlike the real, inverted image of **Fig. 3**. This ray tracing diagram confirms what would be seen when a double convex lens is used to magnify an object lying close to the lens. For the double concave lens of **Fig. 9**, a parallel ray leaves an object point P and is refracted upward at the unit plane H' in such a way that its extension, rather than the ray itself, passes through F'. The ray headed for F is refracted at H before it can reach the object-space focal plane and becomes parallel to the axis. The third ray, going from P to H, emerges at H' parallel to its original direction and its extension to the left passes through the image point P' already determined by the intersection of the other two rays. The resulting virtual image is upright and reduced. All three rays in this diagram behave exactly like the corresponding rays in **Fig. 3**; the only change is the use of the extensions to locate the image. The eye receives the diverging rays from P and believes that they are coming from P'.

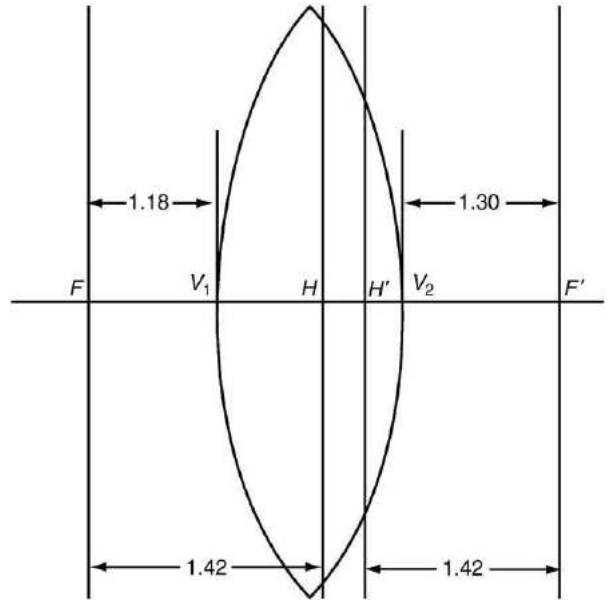


Fig. 6 Positions of cardinal planes for an asymmetric lens.

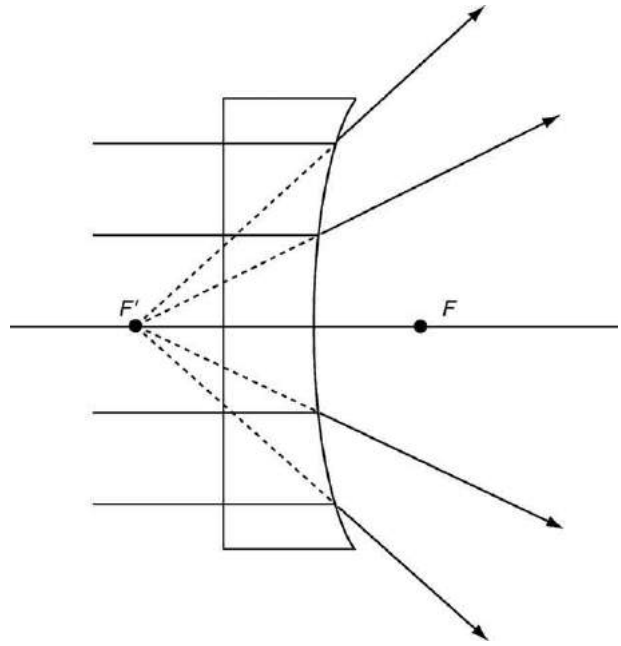


Fig. 7 Ray tracing for a planar-concave lens.

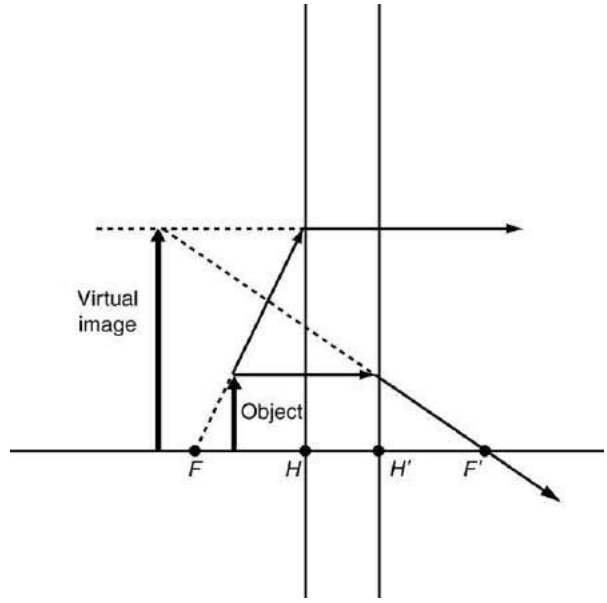


Fig. 8 Formation of a virtual image by a convex lens.

Nodal Points and Planes

Let a ray leave a point on the z-axis at an angle α_1, and have an angle α'_2 when it reaches image space. Since $x=0$ at the starting point, **Eq. (26)** shows that

$$n'_2\alpha'_2 = n_1\alpha_1/m \tag{40}$$

The ratio of the final angle to the initial angle in this equation is the angular magnification μ, which is then

$$\mu = n_1/mn'_2 \tag{41}$$

The locations of object and image for unit angular magnification are called the nodal points, labeled as N and N'. The above

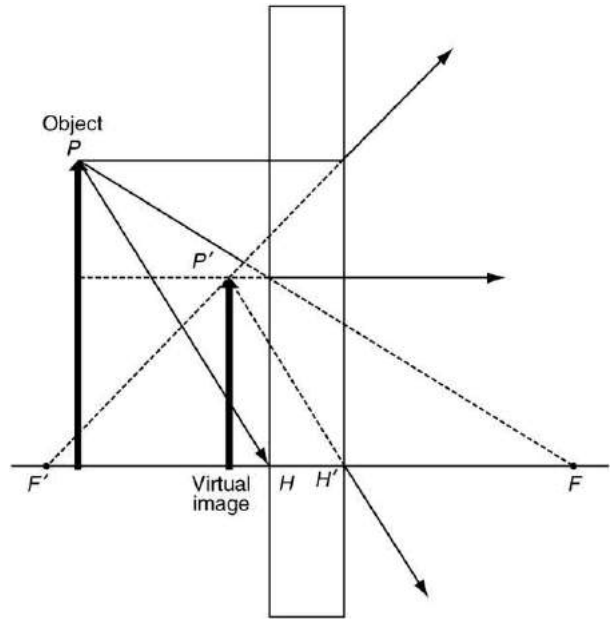


Fig. 9 Formation of a virtual image by a concave lens.

definition shows that the linear and angular magnification are reciprocals of one another when object and image are in air or the same medium. The procedure that located the unit points will show that the nodal points have positions given by

$$\frac{l_N}{n_1} = \frac{(n_2'/n_1) - b}{a} \tag{42}$$

and

$$\frac{l_N'}{n_2'} = \frac{c - (n_1/n_2')}{a} \tag{43}$$

When the two indices are identical, these relations are identical to those for the points H and H'. That is why the rays from P to H and H' to P' in Fig. 3 are parallel.

Compound Lenses

A great advantage of the paraxial matrix method is the ease of dealing with systems having a large number of lenses. To start simply, consider the cemented doublet of Fig. 10: a pair of lenses for which surface 2 of the first element and surface 1 of the second element match perfectly. The parameters of this doublet are given in the manner shown below. It is understood that the first entry under r is r_1, the second entry is r_2, and so on. Note that this doublet has only three surfaces; if the two parts were separated by an air space, producing an air-spaced doublet, then there would be four surfaces to specify. The values of the indices n' and the spacings t' are placed between the values of r. The constants of the doublet are then

$$\begin{array}{ccc} r & n' & t' \\ 1.0 & & \\ & 1.500 & 0.5 \\ -2.0 & & \\ & 1.632 & 0.4 \\ \infty & & \end{array} \tag{44}$$

Note that surface 3, which is flat, has a radius of infinity. Numbering the vertices in the usual manner, the system matrix is

$$S_{31} = R_3 T_{32} R_2 T_{21} R_1 \tag{45}$$

where

$$k_2 = (n_2' - n_2)/r_2 = (1.632 - 1.500)/-2.0 \tag{46}$$

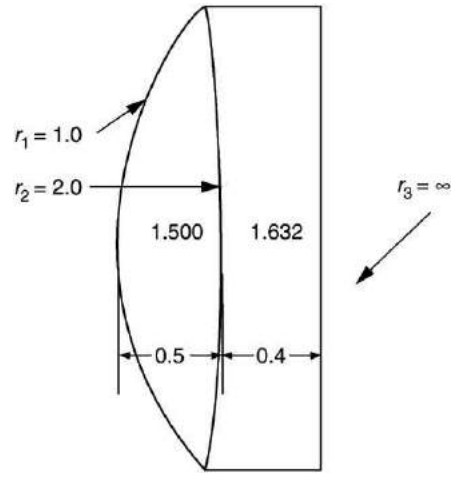


Fig. 10 Specification for a cemented doublet.

Approximations for Thin Lenses

An advantage of paraxial matrix optics is the ease of obtaining useful relations. For example, combining the several expressions for the Gaussian constant a, we obtain:

$$a = -\frac{n_1}{f} = \frac{n_2'}{f'} = k_1 + k_2 - \frac{k_1 k_2 t_1'}{n_1'} \tag{47}$$

and using the definitions of k_1 and k_2 with the lens in air leads to

$$\frac{1}{f'} = (n_1' - 1)\left[\frac{1}{r_1} - \frac{1}{r_2} + \frac{(n_1' - 1)t_1'}{n_1' r_1 r_2}\right] \tag{48}$$

which is the lensmakers' equation. It tells how to find the focal length of a lens from a knowledge of its material and geometry. This derivation is much more direct than what you will usually find. This equation becomes simpler when we are dealing with thin lenses – those for which the third term in brackets can be neglected by assuming that the lens thickness is approximately zero. Then

$$\frac{1}{f'} = (n_1' - 1)\left[\frac{1}{r_1} - \frac{1}{r_2}\right] \tag{49}$$

which is the form often seen in texts. We then realize that the original form is valid for lenses of any thickness, but the thin lens form can be used if the spacing is much less than the two radii. It is also seen that when the thickness is approximately zero, then

$$b = c = 1 \tag{50}$$

and the unit planes have locations

$$l_H = l_H' = 0 \tag{51}$$

Thus, they are now coincident at the center of the lens. A ray from any point on the object would pass undeviated through the geometrical center of the lens, as is often shown in the elementary books. These new values of the Gaussian constants give a system matrix of the form:

$$S_{21} = \begin{pmatrix} 1 & -1/f' \\ 0 & 1 \end{pmatrix} \tag{52}$$

This matrix contains a single constant, the focal length of the lens, so that the preliminary design of an optical system is merely a matter of specifying the focal length of the lenses involved and using the resulting matrices to determine the object-image relationship. We also note that this matrix looks like a refraction matrix; the non-zero element is in the upper right-hand corner. This implies that the refraction–translation–refraction procedure that actually occurs can be replaced, for a thin lens, by a single refraction occurring at the coinciding unit planes. Ophthalmologists take advantage of the thin lens approximation by specifying focal lengths in units called diopters. The reciprocal of the focal length in meters determines its refraction power in diopters. For example, a lens with $f' = 10$ cm will be a 10 diopter lens. If two lenses are placed in contact, the combined power is the sum of the individual powers, for multiplying matrices of the above form gives the upper right-hand element as $(-1/f'_1 - 1/f'_2)$.

The Paraxial System for Design or Analysis

The material given in this article can serve as the basis for an organized way of taking the specifications of an optical system and using them to gain a full understanding of this system. We shall demonstrate it with a practical example which has some interesting features. Chemical engineers have long known that the index of refraction of a liquid can be determined by observing the empty bore of a thick glass tube with a telescope having a calibrated eyepiece and then measuring the magnification of the bore when the liquid flows through it. Consider a liquid with an index of 1.333 and a tube with bore radius of 3 cm, outer radius of 4 cm, and glass of index equal to 1.500. The first step is to specify the parameters of the optical system. These are the radii $r_1 = -3$, $r_2 = -4$ and the thickness $t'_1 = 1$, regarding the tube as a concentric lens with an object to its left; the indices $n_1 = 4/3$ (fractions are convenient), $n'_1 = 3/2$, $n'_2 = 1$. Calculating the elements of the two refraction matrices and the translation matrix, the system matrix is

$$S_{21} = \begin{pmatrix} 1 & -1/8 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 2/3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1/18 \\ 0 & 1 \end{pmatrix} \tag{53}$$

Multiplying, the Gaussian constants are

$$a = 2/27, \quad b = 11/12, \quad c = 28/27, \quad d = -2/3 \tag{54}$$

An important check is to verify that $bc - ad = 1$, as is the case here. The expressions given previously for the positions of the cardinal points then lead to the values

$$\begin{aligned} l_H &= 3/2, \qquad l'_H = 1/2, \qquad l_F = -16\,1/2, \\ l'_F &= 14, \qquad l_N = -3, \qquad l'_N = -4 \end{aligned} \tag{55}$$

These points are shown in Fig. 11. Although the same scale has been used for both horizontal and vertical distances, this is usually not necessary; the vertical scale, which merely serves to verify the magnification, can be arbitrary. The entire liquid column is the object in question, but it is simpler to use a vertical radius as the object. Then the image, generated as described in all the previous examples, coincides with the object and it is enlarged, erect, and virtual. We confirm the image location shown by using the generalized Gaussian lens equation, giving

$$\begin{aligned} t' &= [(-2/3) + (28/27)(-3)/(4/3)]/[11/12) \\ &\quad + (2/27)(-3)/(4/3)] \\ &= -4 \end{aligned} \tag{56}$$

and this agrees with the sketch. The denominator of this expression was shown to be the reciprocal of the magnification $1/m$ with a value of 3/4, and m itself can be calculated from t' as $c - at' = 4/3$; this is another important check on the accuracy of the calculation; the sketch obeys this conclusion, and the purpose of this analysis is confirmed.

Reflectors

To show how reflecting surfaces are handled with matrices, consider a plane mirror located at the origin and normal to the z-axis (Fig. 12). The ray which strikes it at an angle α_1 leaves at an equal angle, resulting in specular reflection. Since $k = 0$ for a flat surface, the refraction matrix reduces to the unit matrix and by Eq. (9):

$$n'_1 \alpha'_1 = n_1 \alpha_1 \tag{57}$$

This contradicts Fig. 12, since α_1 and α'_1 are opposite in sign.

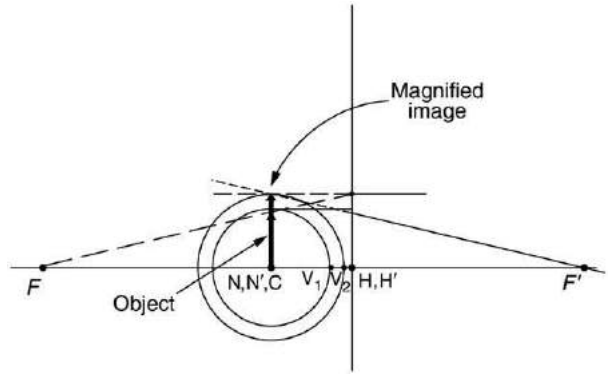


Fig. 11 Ray tracing diagram for a thick-walled glass tube.

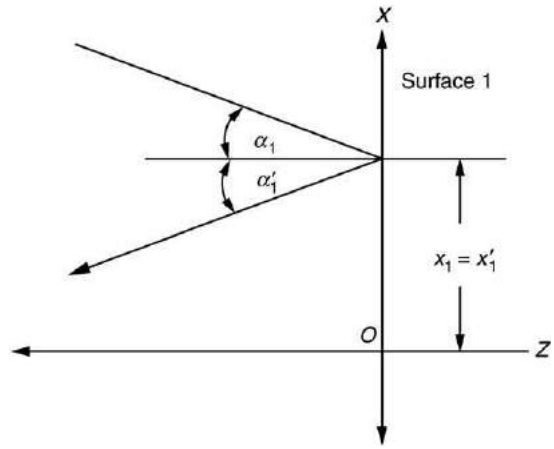


Fig. 12 Specular reflection.

If, however, we specify that

$$n_1' = -n_1 \tag{58}$$

then the difficulty is removed. Hence, reflections involve a reversal of the sign on the index of refraction, and two successive reflections restore the original sign. For a mirror, the refraction power is

$$k_1 = \frac{n_1' - n_1}{r_1} = \frac{-1-(1)}{r_1} = -\frac{2}{r_1} \tag{59}$$

The system matrix for a single refracting surface thus becomes

$$S_{11} = R_1 = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2/r_1 \\ 0 & 1 \end{pmatrix} \tag{60}$$

with

$$a = -2/r_1, \quad b = c = 1, \quad d = 0 \tag{61}$$

The connection between object and image can be expressed as

$$\begin{pmatrix} n_1'\alpha_1' \\ x_1' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t'/n_1' & 1 \end{pmatrix} \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -t/n_1 & 1 \end{pmatrix} \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix} \tag{62}$$

and using the same procedure that was applied to **Eqs. (28)–(31)**, the locations of the six cardinal points are

$$\begin{aligned} l_F &= -n_1 b/a, \quad l_F' = n_1' c/a \\ l_H &= n_1(1-b)/a, \quad l_H' = n_1'(c-1)/a \\ \frac{l_N}{n_1} &= \frac{(n_1'/n_1) - b}{a}, \quad \frac{l_N'}{n_1'} = \frac{c - (n_1/n_1')}{a} \end{aligned} \tag{63}$$

so that for the spherical mirror:

$$l_F = \frac{(-1)(1)}{-2/r_1} = \frac{r_1}{2} = l_F' \tag{64}$$

$$l_H = l_H' = 0 \tag{65}$$

and

$$l_N = \frac{-1-1}{-2/r_1} = r_1 \tag{66}$$

but

$$\frac{l_N'}{-1} = \frac{1-(-1)}{-2/r_1}, \quad l_N' = r_1 \tag{67}$$

Since r_1 is negative, these equations show that the unit points lie on the vertex, the foci coincide halfway between the center of curvature and the vertex, and the nodal points are at this center. The perfect focusing is a consequence of the paraxial approximation. Ray tracing for this mirror uses the procedure developed for lenses, which applies to optical systems of any complexity. **Fig. 13** shows an object to the left of the center of curvature. The ray from P parallel to the axis goes to H' and then through F', while the ray through F goes to H and then becomes parallel to the axis. Their intersection at P' produces a real, inverted image. The

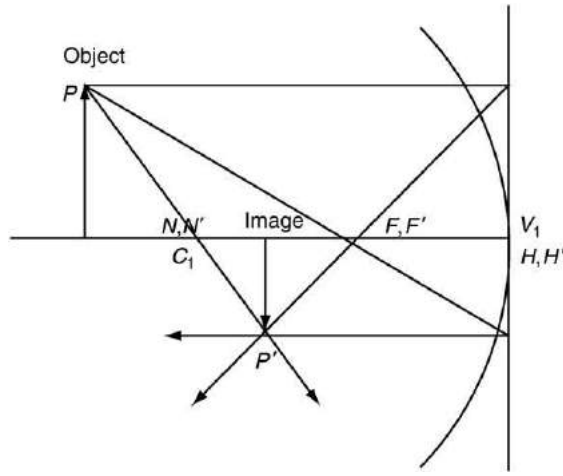


Fig. 13 Ray tracing for a concave mirror.

third set of rays represents something different: it requires a knowledge of the nodal point locations. The ray from P to N should be parallel to the ray from N' to P', by definition; in this example, they meet this requirement by being colinear.

As an example of the power of matrix optics to simplify the design or analysis of an optical system involving lenses and mirrors, consider an object that is 15 units to the left of a converging lens, with focal length $f' = 10$. A concave mirror, of radius $r_1 = 16$, is 20 units to the right of the lens. The refraction power of the mirror is $k_1 = (-1-1)/-16 = 1/8$ and using the thin lens form of the system matrix, the combined matrix is

$$\begin{pmatrix} 1 & -1/8 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 20 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1/10 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -3/2 & 1/40 \\ 20 & -1 \end{pmatrix} \tag{68}$$

This gives the constants a, b, c, and d, from which we find that the image is $-4\ 4/9$ units to the left of the mirror, its magnification is $-8/9$, and it is inverted. Compare the simplicity of this procedure with the usual way of doing this calculation, which involves first finding the image in the lens and then using that image as a virtual object.

Conclusion

It has been shown that the use of paraxial matrices provides a simple approach to an understanding of the kind of optical systems that provide perfect images. Further details and derivations of expressions given here will be found in the Further Reading below.

Further Reading

Brouwer, W., 1964. Matrix Methods in Optical Instrument Design. New York: WA Benjamin.
Nussbaum, A., 1998. Optical System Design. Upper Saddle River, NJ: Prentice-Hall.

Aberrations

A Nussbaum, University of Minnesota, Minneapolis, MN, USA

Introduction

The images may be magnified or reduced and they may be erect or inverted, but they are otherwise faithful replicas of the object. This behavior is a consequence of ray tracing based on an approximate or paraxial form of Snell's law which we write as

$$n\theta = n'\theta' \tag{1}$$

where n and n' are the indices of refraction on either side of an interface and θ and θ' are the angles of incidence and refraction. If, however, the following exact form of Snell's law:

$$n\sin\theta = n'\sin\theta' \tag{2}$$

governs the ray behavior, as is the case when the rays have large inclinations with respect to the symmetry axis, then it is necessary to devise a more involved ray tracing procedure. Fig. 1, which shows the first surface of a lens, will be used to explain how this procedure works. A ray leaves the object point P and strikes the lens surface at the point P_1. Rays which lie completely in the plane ZOX are said to be meridional; this is the plane which passes through the symmetry axis, just as meridians on the Earth's surface are determined by planes through the geographic axis. By regarding the various distances in the figure as vectors, it is possible to start with the initial coordinates of the ray and with the direction-cosines which specify its slope and find the coordinates of the point P' where the ray strikes the lens surface. A derivation of the equation governing this process will be found in the text of A Nussbaum (see Further Reading). This derivation involves only elementary algebra and vector analysis, and will be found simpler than the usual textbooks treatment of ray tracing procedures. The equation is

$$T_1 = \frac{F}{-E + \sqrt{E^2 - c_1 F}} \tag{3}$$

where

$$E = Bc_1/2 = c_1[(z - v_1)N + \mathrm{xL}] - N \tag{4}$$

and

$$F = Cc_1 = c_1\left[(z - v_1)^2 + x^2\right] - 2(z - v_1) \tag{5}$$

Knowing the coordinates z and x of the starting point P of the ray, as well as its starting cosines N and L, plus the curvature c_1 or radius $r_1 = 1/c_1$ of the lens surface and the location v_1 of the surface's vertex, we can readily calculate the length T_1 of the ray. Then its components on the two axes give the coordinates of P'. Next, we need to find the slope of the ray after it is refracted at the lens surface, and this can be calculated from Snell's law combined with the behavior of the incident and refracted rays regarded as vectors. Again we invoke the reference cited above for a derivation of the equations giving the value of the direction-cosines after refraction; these expressions are

$$n_1'N_1' = K_1/c_1 + n_1N_1 - K_1z_1 \tag{6}$$

and

$$n_1L_1' = n_1L_1 - K_1x_1 \tag{7}$$

where K_1 is known as the refracting power and has the definition

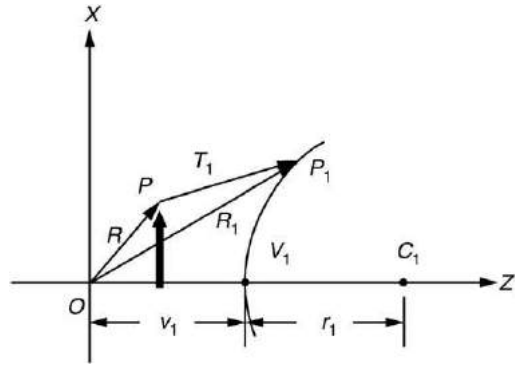


Fig. 1 Ray tracing in the meridional plane.

 doi:10.1016/B978-0-12-809283-5.00688-1

$$K_1 = c_1(n_1' \cos\theta' - n_1 \cos\theta) \tag{8}$$

The ray can now be traced to the next surface knowing the lens thickness, and then to the image plane. These two sets of equations, when incorporated into a computer program of about two dozen lines, will trace meridional rays to any desired accuracy through an optical system containing an arbitrary number of lenses and spherical reflectors. As an example, consider a symmetric double convex lens with radii of 50 units at surface 1, -50 units at surface 2, a thickness of 15 units, and an index of 1.50. Rays parallel to the axis – one at 0.2 units above the axis, a second one at 2 units above the axis, and a third one at 20 units above the axis – are started well to the left and traced to the paraxial focal plane, which is at 47.368 units from the second surface. The ray that is 0.2 units above the axis will cross the focal plane at the axis; this is a paraxial ray, passing through the paraxial focal point F'. The ray which is 2 units above the axis crosses the focal plane slightly below the focal point, and is therefore not quite paraxial, while the ray which starts at 20 units above the axis falls well short of the focal point, intersecting the axis before it reaches the focal plane. This behavior indicates that the lens possesses a defect known as spherical aberration and these calculations imply that spherical aberration should be defined as the failure of nonparaxial rays to conform to the behavior of paraxial rays. This definition does not appear in optics texts, and the few definitions that are given appear to be incorrect. It is meaningless to speak of a focal point as defined by a pair of nonparaxial rays; this infinite set of ray-pairs determines a corresponding number of intersection points, all of which fall short of the true focal point F' This situation is nicely illustrated by the next figure, which shows a large number of parallel rays, the central ones being paraxial and the others behaving differently. The rays which are very close to the axis will meet as expected at the focal point F'. The rays a little farther out – the intermediate rays – will cross the axis just a little to the left of F', and those near the edge of the lens – the *marginal* rays – will fall very short, as indicated. Rotating this diagram about the axis, we realize that spherical aberration produces a very blurry circular image at the paraxial plane. The three-dimensional envelope of the group of rays produced by this rotation is known as the caustic surface and the narrowest cross-section of this surface, slightly to the left of the paraxial focal plane, is called the circle of least confusion. If this location is chosen as the image plane, then the amount of spherical aberration can be reduced somewhat. Other methods of improving the image will be considered below. **Fig. 2** was obtained by adding a graphics step to the computer program described above. **Fig. 3** shows how to define spherical aberration

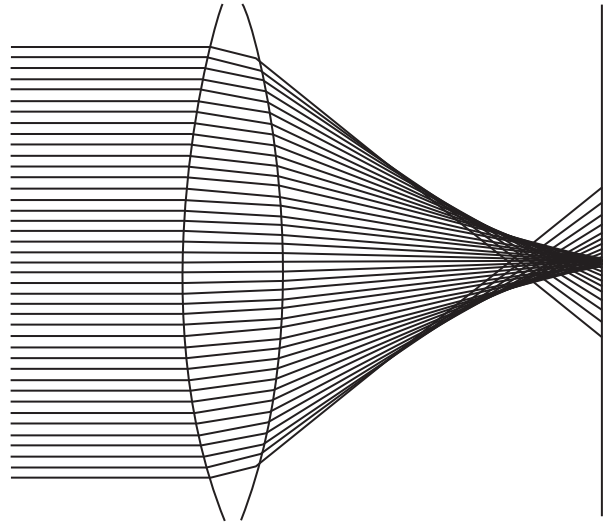


Fig. 2 Spherical aberration for parallel meridional rays.

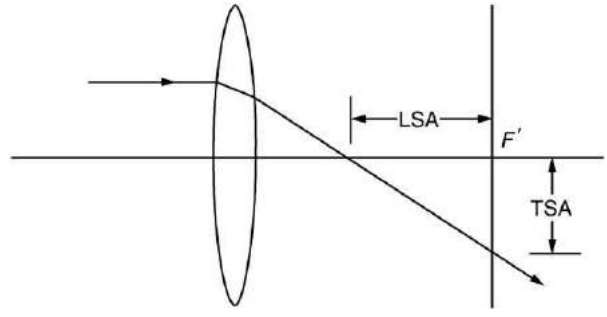


Fig. 3 Definitions of transverse and longitudinal spherical aberration.

quantitatively. Two kinds of aberration are specified in this figure. The place where the ray crosses the axis, lying to the left of F', is at a distance, measured from the paraxial focus, called the longitudinal spherical aberration (LSA), while the distance below the axis at the focal plane is the transverse spherical aberration (TSA). To reduce this aberration, note in **Fig. 2** that the amount of refraction at the two surfaces of the lens for marginal rays is not the same. Equalizing the refraction at the two surfaces will improve the situation. Suppose we alter the lens while keeping the index, thickness, and focal length constant; only the radii will change. This can be done if the new radii satisfy the paraxial equation giving the focal length in terms of the lens constants. The process of varying the shape of lens, while holding the other parameters constant, is called *bending the lens*. This can be easily accomplished by a computer program, obtaining what are known as optimal shape lenses, which are commercially available. To study their effect on spherical aberration, define the shape factor σ of a lens as

$$\sigma = \frac{r_2 + r_1}{r_2 - r_1} \tag{9}$$

from which $\sigma = 0$ for a symmetric lens ($r_1 = -\ r_2$). **Fig. 4(a)** shows the monitor screen for a symmetric lens, and **Fig. 4(b)** illustrates the result of lens bending. This lens has had the curvature of the first surface raised and the second surface has been flattened to keep the focal length constant. Let us now consider what lens designers can do about spherical aberration. Brouwer in his book (see Further Reading) gives the specifications for a cemented doublet consisting of a double convex lens followed by a lens with two concave surfaces. The longitudinal spherical aberration of the front lens alone varies as shown in **Fig. 5**. The way that the designer arranged for the marginal rays to meet at the paraxial focus, rather than falling short, was to recognize that a diverging lens following the front element would refract the rays away from – rather than closer to – the axis. When this second component is added, we would expect the behavior shown in **Fig. 6**. Note that the value of LSA for the doublet has been enormously reduced, as indicated by the horizontal scale change. Spherical aberration can be reduced in more elaborate systems as part of the design process. Telescopes, in particular, make use of corrector plates which are placed in front of the large mirrors.

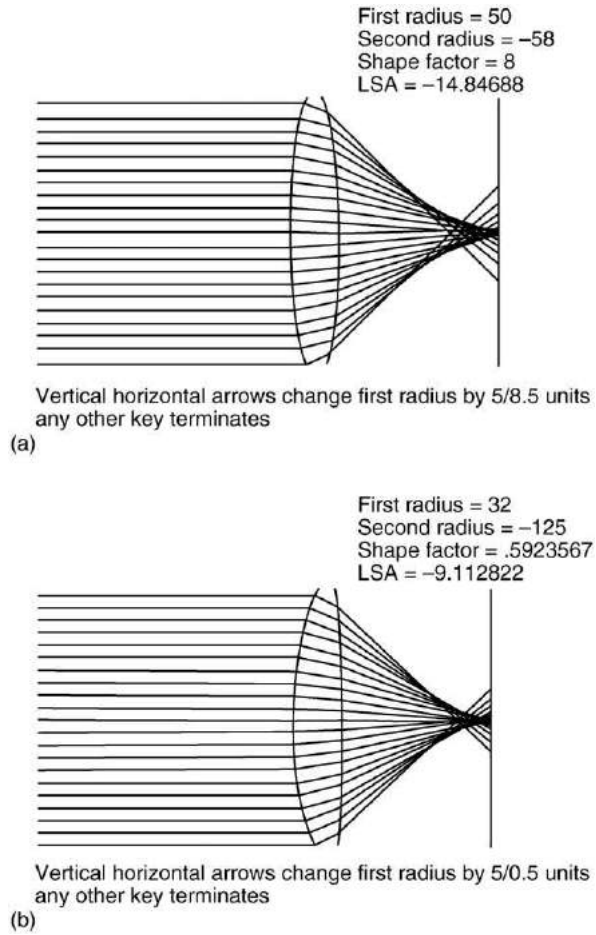


Fig. 4 Lens bending as performed numerically.

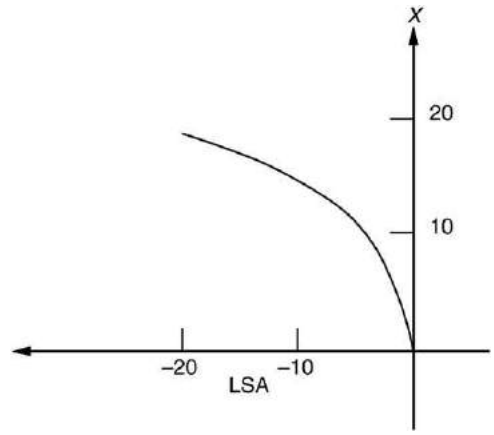


Fig. 5 Spherical aberration for a single lens.

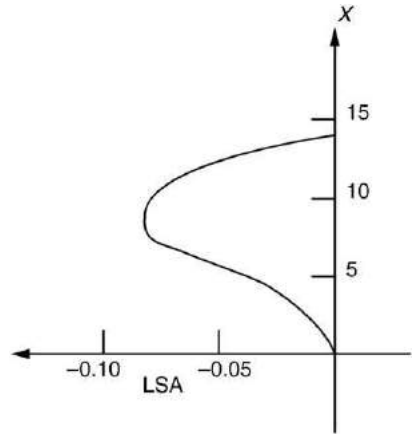


Fig. 6 Reduction of spherical aberration by a doublet.

Coma

We have seen how to trace nonparaxial, meridional rays and found that spherical aberration appears because the intermediate and marginal rays do not behave like the central rays. We now wish to deal with rays that do not lie in a meridional plane; these are nonmeridional or skew rays. Calculating the behavior of such a ray would appear to be a complicated problem, but it turns out to be a simple extension of what has already been accomplished. Let a skew ray start at an arbitrary point in space with coordinates x, y, and z and let it have direction cosines L, M, and N with respect to OX, OY, and OZ. The points P and P_1 in **Fig. 1** are now out of the plane ZOX and when the vectors in this figure are expressed as the sum of three, rather than two, components, the equations given above for meridional rays acquire terms in y and M which have the same form as those for x and L. Now trace a set of skew rays chosen to lie on a cylinder (**Fig. 7**) which is centered about the z-axis of a lens, so that all the rays are meridional. These rays are fairly far from the axis, so that they are also nonparaxial, but they all meet at a common image point, forming a cone whose apex is this image point. Then give this cylinder a downward displacement while holding fixed the intersection point of each ray with the dotted circle on the front of the lens. The tilting of the cylinder changes all its rays – except the top and the bottom rays – from meridional to skew. In addition, the cone on the image side will then tilt upwards; it should not change in any other way if the skew rays continue to meet at a well-defined apex. But this is not what happens in a simple lens, as will now be explained.

Let the top and bottom meridional rays meet at a point P in image space (**Fig. 8**) and use this point to determine the location of an image plane. The skew ray just below the uppermost ray in **Fig. 7** will pass through this plane fairly close to the intersection point. Let us guess (and confirm later) that it will pass through a point slightly to the right of P and a little below it, as we observe this plane from the lens position. This is point 1 in **Fig. 8**. The next skew ray on the front of the lens will strike the image plane still farther below and to the right, producing point 2 in the figure. As we continue to trace these rays down the front half of the circle, we realize that a ray almost at the bottom of **Fig. 7** will have to be very close to the lower meridional ray and that its image point is just below but to the left of P; this is the point labeled 'next-to-last' in **Fig. 8**. All of these points join to form a closed curve. The

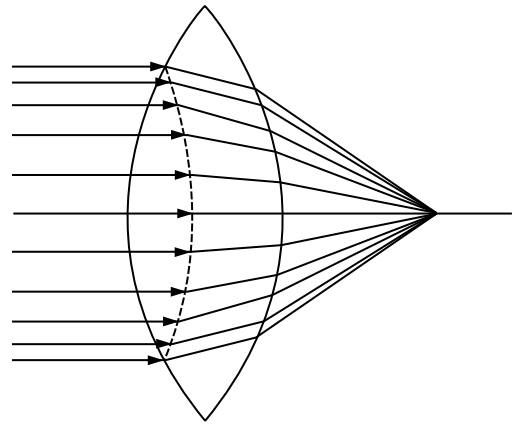


Fig. 7 Configuration used to illustrate coma.

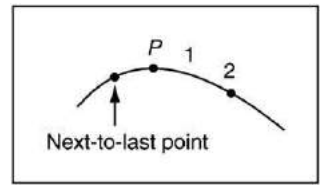


Fig. 8 Starting a coma plot.

back halves of the dotted circle and of the ray cylinder are not shown in the figure, but since we are again starting at P in Fig. 8, the pattern will repeat itself. That is, we get a second closed curve which – by symmetry – should be identical to the first curve. This is a most unusual effect; double-valued phenomena are quite rare. To confirm this prediction, we use the skew ray equations, and trace the pattern formed by a set of concentric cylinders of different sizes. The appearance of the coma pattern generated by the computer is shown in Fig. 9. Each half-circle on the front side of the lens produces a closed figure (an approximation to a circle) as an image. If coma were the only aberration (when it is called ideal coma), the plots would be the perfect, coinciding circles of Fig. 10. The comet-like shape explains where the name comes from. We can picture the generation of these circles in the way shown in this figure. The ray through the geometrical center of the lens produces a point in the paraxial focal plane. The top and bottom rays (the meridional rays) from the circles of increasing size produce points which define image planes that are successively closer to the lens; this is a consequence of spherical aberration. Moving the top ray parallel to itself and along the half-circle generates the image plane pattern. The lines tangential to the coma circles make an angle of 60° with each other, as can be seen by measurement on the computer output. Most optics books look at coma in a way different from that given here. Coma has been defined by WJ Smith (see Further Reading) as the variation of magnification with aperture, analogous to the definition of spherical aberration as the variation of focal length with aperture. And as was previously mentioned, focal length and magnification are purely paraxial concepts, not applicable to the description of aberrations. The definition that comes from the arguments given above is that coma is the failure of skew rays to match the behavior of meridional rays, just as spherical aberration is the failure of meridional rays to match the behavior of paraxial rays. Most texts show coma in its ideal form.

Astigmatism

We have shown that spherical aberration is the failure of meridional rays to obey the paraxial approximation and coma is the failure of skew rays to match the behavior of meridional rays. To see the connection between these two aberrations, consider the object point P of Fig. 11. This very complicated diagram can actually be very helpful in understanding the third of the five geometrical aberrations. Let P be the source of a meridional fan: this is the group of rays with the top ray labeled PA and the bottom ray is PB. If the lens is completely corrected for spherical aberration, this fan will have a sharp image point P'_T lying directly below the z-axis. Now let P also be the source of a fan bounded by the rays PC and PD. This fan is at right-angles to the other one; we can think of these rays as having the maximum skewness possible. If the lens has no coma, this fan also will produce a sharp image point P'_s which, for a converging lens, will be farther from the lens than the tangential focal point. In other words, even though the lens has been fully corrected for both aberrations, the two corrections will not necessarily produce a common image. In the figure, the meridional fan is called a tangential fan and the skew fan is called a sagittal fan; these are the terms commonly used by lens designers. The failure of the sagittal and tangential rays to produce a single image in a lens corrected for both spherical

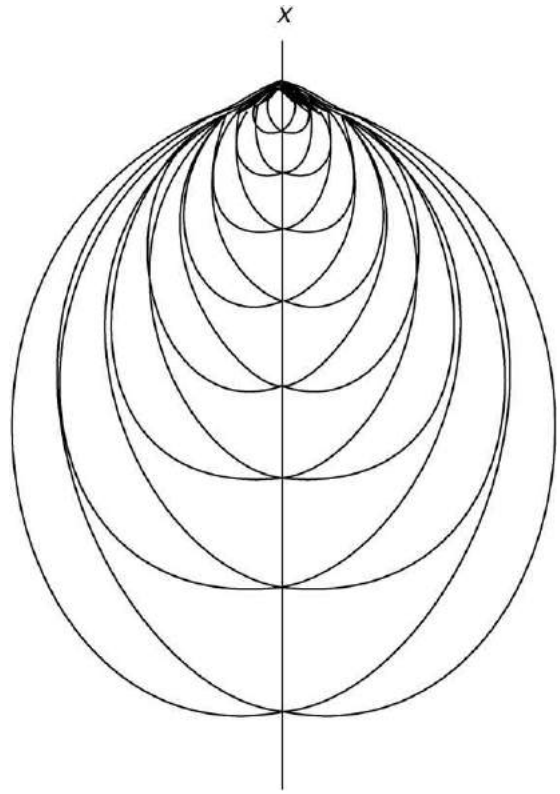


Fig. 9 Coma as calculated numerically.

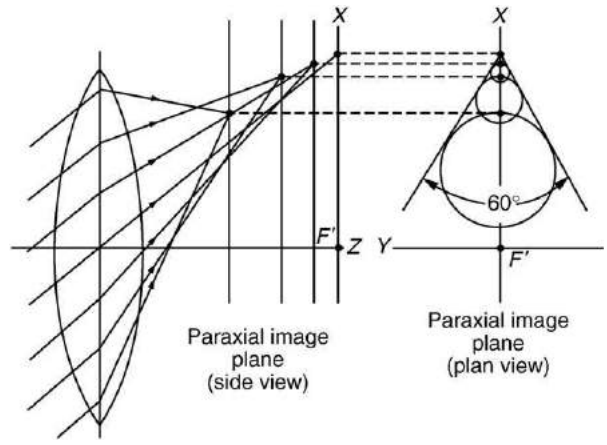


Fig. 10 Customary way of illustrating coma.

aberration and coma is known as astigmatism. Astigmatism, coma, and spherical aberration are the point aberrations that an optical system can have. We have already noted that spherical aberration is the only one that can be associated with a point on the z-axis; the others require an off-axis object. Astigmatism is very common in the human eye; to see how it shows up, look at the sagittal image point P'_s of **Fig. 11**. This point is a distance z_s, from the x,y-plane. A plane through this point will have a line image on it, the sagittal line, due to the tangential fan, which has already come to a focus and is now diverging. The converse effect, producing a tangential line, occurs at the other image plane. If we locate an image plane halfway between the two, then both fans contribute to the image and in the ideal case it will be a circle. As the image plane is moved forwards or backwards, these images become ellipses and eventually reduce to a sagittal or a tangential line, as shown in **Fig. 12**. Because there are two different image planes, an object with spokes (**Fig. 13**) will have an image for which the vertical lines are sharp and the others get gradually poorer as they become more horizontal, or vice versa, depending on which image plane is chosen. Eye examinations detect astigmatism if the spokes appear to go from black to gray. This example of the effect of astigmatism was often based on a radial pattern of arrows, which explains the origin of the word sagittal, derived from the Latin 'sagitta' for arrow.

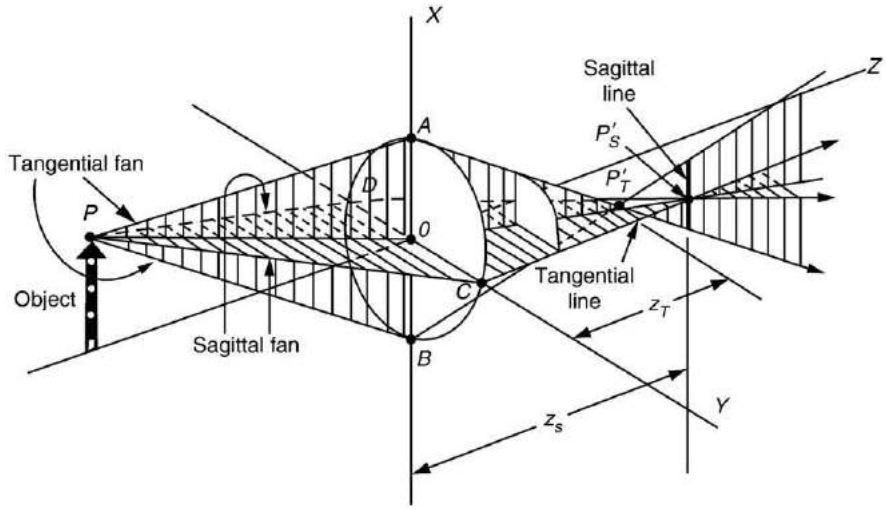


Fig. 11 Tangential and sagittal fans displaying astigmatism.

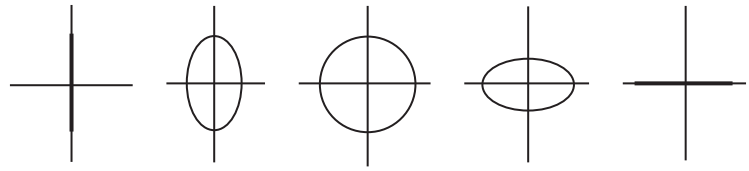


Fig. 12 Astigmatic image patterns.

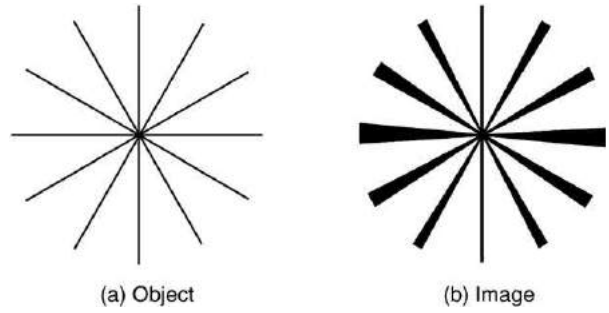


Fig. 13 Image degraded by astigmatism.

Curvature of Field and Distortion

Having covered the three point aberrations, there are two lens defects associated with extended objects. If we move the object point *P* in Fig. 11 closer to or farther from the *z*-axis, we would expect the positions of the tangential and sagittal focal planes to shift, for it is only when the paraxial approximation holds that these image points are independent of *x*. Hence, we obtain the two curves of Fig. 14, which shows what astigmatism does to the image of a two-dimensional object. If the astigmatism could be eliminated, the effect would be to make these curved image planes coincide, but we have no guarantee that the common image will be flat, or paraxial. The resulting defect is called Petzval curvature or curvature of field. For a single lens, the Petzval surface can be flattened by a stop in the proper place, and this is usually done in inexpensive cameras. Petzval curvature is associated with the *z*-axis. If we take the object in Fig. 11 and move it along the *y*-axis, then all rays leaving it are skew and this introduces distortion, the aberration associated with the coordinates normal to the symmetry axis. Distortion is what causes vertical lines to bulge outward (barrel distortion) or inward (pincushion distortion) at the image plane, as shown in Fig. 15. A pinhole camera will have no distortion, since there are no skew rays, and a single thin lens with a small aperture will have very little. Placing a stop near the lens to reduce astigmatism and curvature of field introduces distortion because, as shown in Fig. 16(a), the rays for object points far from the axis are limited to off-center sections of the lens. The situation in this figure corresponds to barrel distortion; placing the stop on the

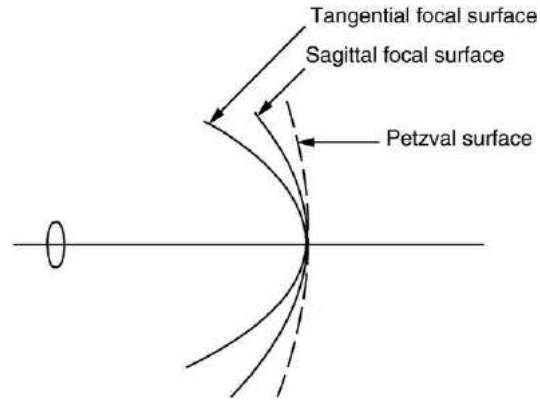


Fig. 14 Petzval of field curvature.

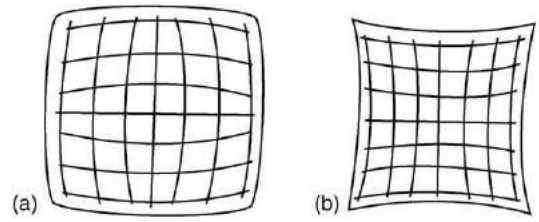


Fig. 15 Barrel and pincushion distortion.

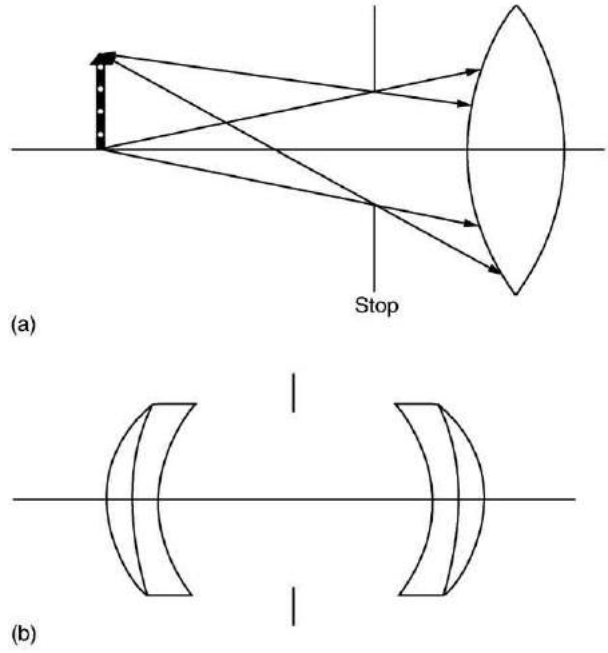


Fig. 16 (a) Distortion-creating stop. (b) Symmetric-component lens that reduces distortion.

other side of the lens produces pincushioning. Distortion can be therefore reduced by a design which consists of two identical groupings (Fig. 16(b)), with the iris diaphragm placed between them; the distortion of the front half cancels that of the back half. Fig. 17 shows the output for a calculation using approximate formulas rather than exact ray tracing. Note that the designer has arranged for astigmatism to vanish at the lens margin. This example indicates the nature of astigmatism in a system for which the coma and the spherical aberration are low but not necessarily zero; it is the failure of the tangential and sagittal image planes to coincide.

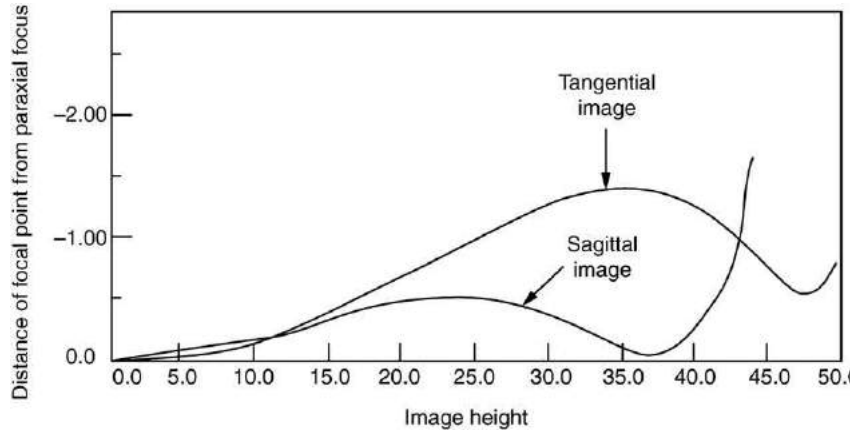


Fig. 17 Reduction of astigmatism through design.

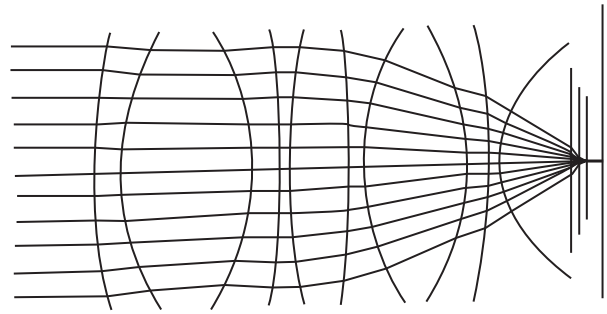


Fig. 18 High-power high-resolution microscope objective.

Nonspherical Surfaces

So far, we have considered optical systems that use only spherical or plane reflecting and refracting surfaces. However, many astronomical telescopes use parabolic and hyperbolic surfaces. In recent years, nonspherical or aspheric glass lenses have become more common because they can be mass produced by computer-controlled machinery and plastic lenses can be molded with very strict tolerances. Ray tracing procedures given above can easily be extended to the simplest kind of aspheric surfaces: the conic sections. Iterative procedures may be applied to surfaces more complicated than the conic sections. A well-known example is the curved mirror; a spherical reflecting surface will have all the aberrations mentioned here, but a parabolic mirror will bring parallel rays to a perfect focus at the paraxial focal point, thus eliminating the spherical aberration. Headlamp reflectors in automobiles are a well-known example. A plano-convex lens with this defect corrected will be one whose first surface is flat and whose second surface is a hyperbola with its eccentricity equal to the index of refraction of the glass composing the lens. One form of astronomical telescope, which has a large primary concave mirror to collect the light and reflect it back to a smaller secondary convex mirror, is the Ritchey–Chrétien design of 1924. Both mirrors are hyperbolic in shape, completely eliminating spherical aberration. This is the design of the Hubble space telescope. The two-meter diameter primary mirror, formed by starting with a spherical mirror and converted to a hyperbola by computer-controlled polishing, was originally fabricated incorrectly and caused spherical aberration. From a knowledge of the computer calculations, it was possible to design, build, and place in orbit a large correcting optical element which removed all the aberrations from the emerging image.

Conclusion

The three-point aberrations and the two extended object aberrations have been defined and illustrated. Simple ways of reducing them have been mentioned, but the best way of dealing with aberrations is in the process of lens design. In addition, there is a completely separate defect known as chromatic aberration. All the above calculations are based on ray tracing for a single wavelength (or color) of visible light. However, the index of refraction of a transparent material varies with the wavelength of the light passing through it, and a design good for one value of the wavelength will give an out-of-focus image for any other wavelength. It has been known for many years that an optical system composed of two different kinds of glass will provide a

complete correction for two colors – usually red and blue – and other colors will be partially corrected. Such lenses are said to be achromatic. For highly demanding applications, such as faithful color reproductions of works of art, it is possible to design elaborate and expensive lenses which correct for yellow light as well. This kind of optical system is said to be apochromatic. With perfect imaging at the two ends and the middle of the visible spectrum, the color error elsewhere is virtually nonexistent. As an example of a design which is free of all aberrations – geometric and chromatic – **Fig. 18** shows the ray trace for a high-power aberration-free and apochromatic microscope objective. Note that the trace shows parallel rays brought to a focus inside the cover glass; in use, the light source would be a point on the slide and the light would converge at the eyepiece.

Further Reading

Brouwer, W., 1964. Matrix Methods in Optical Instrument Design. New York: WA Benjamin.
Laikin, M., 1990. Lens Design. New York: Marcel Dekker.
Nussbaum, A., 1998. Optical System Design. Upper Saddle River, NJ: Prentice-Hall.
Nussbaum, A., Phillips, R.A., 1976. Contemporary Optics for Scientists and Engineers. Englewoods Cliffs, NJ: Prentice Hall.
Smith, W.J., 1990. Modern Optical Engineering, 2nd edn. New York: McGraw-Hill Book Co.

Prisms

A Nussbaum, University of Minnesota, Minneapolis, MN, USA

Introduction

Prisms are solid structures made from a transparent material (usually glass). **Fig. 1** shows the simplest type of prism – one that has many uses. The vertical sides are two identical rectangles at right angles to each other and a larger third rectangle is at 45° to the other two. The top and bottom are 45°–45°–90° triangles. The most common applications of prisms are:

1. To bend light in a specified direction
2. To fold an optical system into a smaller space
3. To provide proper image orientation
4. To combine or split optical beams, using partially reflecting surfaces
5. To disperse light in optical instruments such as spectrographs.

Single Prisms as Reflectors

A prism like that of **Fig. 1** can act as a very effective reflector; it is more efficient than a mirror since there is no metal coating to absorb light. **Fig. 2** shows a light ray entering the prism at right angles to one of the smaller sides, reflected at the larger side, and emerging at right angles to the other small side. The path shown is determined by Snell's law. This law has the form:

$$n \sin \theta = n' \sin \theta' \tag{1}$$

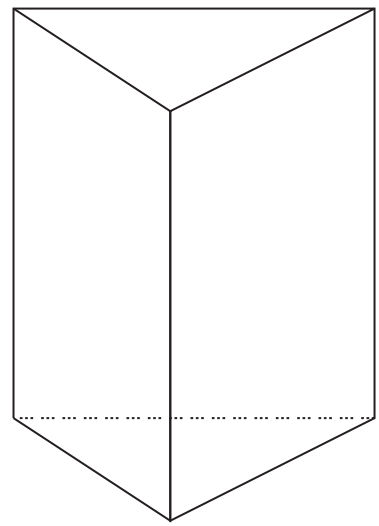


Fig. 1 A 45–45–90 degree prism.

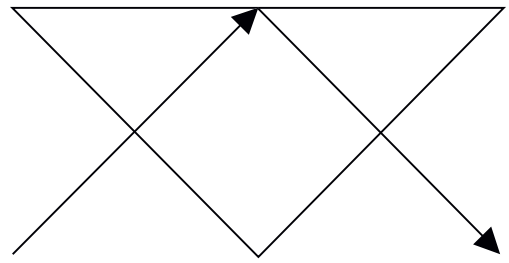


Fig. 2 Total internal reflection by a prism.

 doi:10.1016/B978-0-12-809283-5.00687-X

This indicates that a ray of light making an angle θ to the normal at the interface between media of indices of refraction n and n' will be refracted and the angle to the normal becomes θ'. Notice that the light ray strikes the surface where the reflection takes place at an angle of 45° to the normal; this angle is too large to permit it to emerge from the prism after refraction. To understand why, we recognize that when the ray is able to leave the glass, Snell's law requires that the emerging angle be larger than the incident angle. The largest possible incident angle – known as the critical angle – will correspond to an emerging angle of 90°; that is, this ray will be parallel to the interface. For an index of 1.50, as an example, the critical angle θ_c satisfies Snell's law in the form:

$$1.5 \sin \theta_c = 1.0 \sin 90^\circ \tag{2}$$

from which $\theta_c = 41.8^\circ$. The ray shown in the figure will therefore remain in the glass. Furthermore, since the indices appearing on both sides of Snell's law are now those of the glass, then the incident and refracted angles must be identical, and the prism reflects the ray in the same way that a mirror would. This phenomenon is known as total internal reflection.

This prism has an important application as an optical component of a fingerprint machine, whose action depends on modifying the total internal reflection which normally occurs. **Fig. 3** shows the prism and the reflecting behavior of **Fig. 2**, with a person's finger pressed against the long side. When the ray strikes a place on the surface corresponding to a valley in the fingerprint pattern, the index at this region is that of air, and the ray is reflected in the usual manner. But a ray striking a portion of the prism which lies immediately below a raised section of the skin will make an angle with the normal which is now smaller than the critical angle, since the index is much greater than that of air, so that the ray will cross the interface and be absorbed by the finger. The reflected beam, with the fingerprint pattern sharply reproduced, is sent to a detector and the resulting signal goes to a computer, where it is processed and used to drive a printer. The fingerprint cards generated by this kind of equipment are sharper than those done by the traditional smeared-ink method. It is possible to observe the phenomenon just described by looking at your fingers while holding a glass of water – the fingerprints will show up in an enhanced manner. In addition to the better quality, the manufacturer of one version of this equipment has incorporated a feature in their computer program which can detect attempts by the person being fingerprinted to degrade the image.

A more elaborate single prism – one with five sides – is the pentaprism of **Fig. 4**, used extensively in single lens reflex cameras. This prism does not take advantage of the total internal reflection process, as described above. Below the bottom face is a mirror at

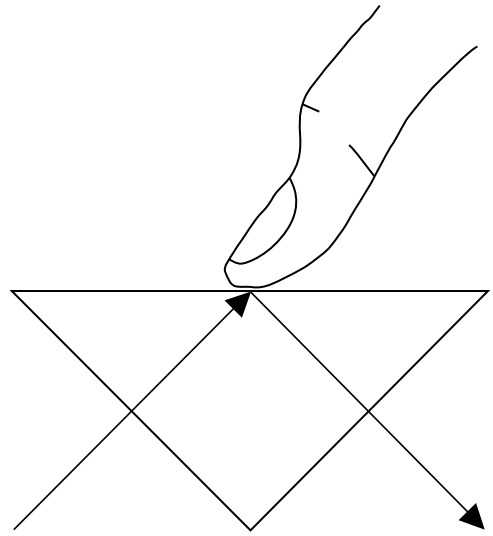


Fig. 3 Use of a prism to produce a fingerprint.

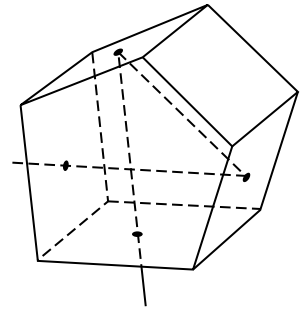


Fig. 4 A pentaprism.

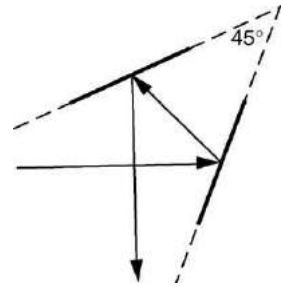


Fig. 5 Action of coated faces of a pentaprism.

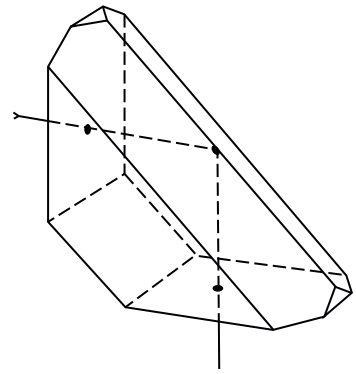


Fig. 6 Amici prism.

a 45° angle. Light from the camera's lens is reflected upward and emerges from the left-hand face where the scene being photographed is observed by the eye. Since there are two reflections, the reversal caused by a single mirror is eliminated. The camera's mirror is hinged and when the exposure button is pressed, it flies up and the incoming light can reach the film. The two reflecting faces, which lie at an angle of 45° to one another (**Fig. 5**), are coated. To see why this is necessary, note that the two angles in the isosceles triangle formed by the light ray and the face extensions must equal $(180^\circ - 45^\circ)/2$ or 67.5°. This ray must then lie at an angle of $90^\circ - 67.5^\circ$ to the normal, or 22.5°, which is much less than the critical angle needed for reflection.

There are at least a dozen other forms of prism which have been designed and used. For example, the Amici prism (**Fig. 6**) is a truncated right-angle prism with a roof section added to the hypotenuse face. It splits the image down the middle and thereby interchanges the right and left portions. These prisms must be very carefully made to avoid producing a double image. One use is in telescopes to correct the reversal of an image.

Double Prism Reflectors

Another extensive application of prisms is their use in pairs as components of telescopes and binoculars. The most widely used is a pair of Porro prisms: two right angle prisms arranged as shown in **Fig. 7**. The double change in direction has two effects; it eliminates image reversal and it shortens the spacing between objective and eyepiece lenses, making the binoculars lighter and more compact. It is possible to calculate how light rays behave in Porro prisms with a ray tracing procedure. This method involves the use of 4×4 matrices and its development is rather complicated. A simple understanding of the way light passes through this optical component can be based on **Fig. 8**, which shows the two prisms separated for clarity. An arbitrary ray (the heavy line) coming from the object will be reflected at a 45° angle at each of the four faces that it strikes. Using two line segments at right angles as an object, the ray from the top of the vertical arrow and the ray from the end of the horizontal arrow will behave as indicated. The image is then identical to the object but rotated by 180°. Since the optics of the binoculars – the combined effect of the objective and the eyepiece – result in an inverted image, the Porro prism corrects this problem.

Prisms as Instruments

Prisms are the crucial optical element in many laboratory instruments. To show how they can be used to measure the index of refraction of transparent materials – solids or liquids – consider the prism of **Fig. 9**, with apex angle A. Light comes to the prism at

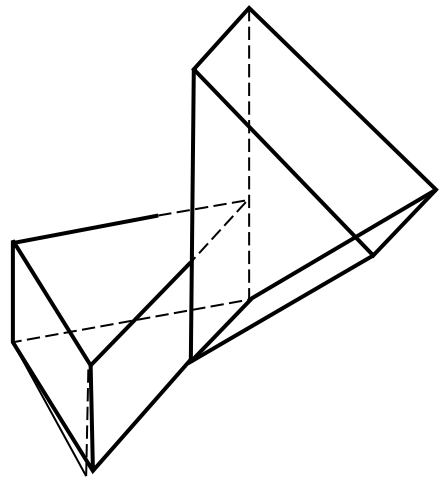


Fig. 7 A pair of Porro prisms.

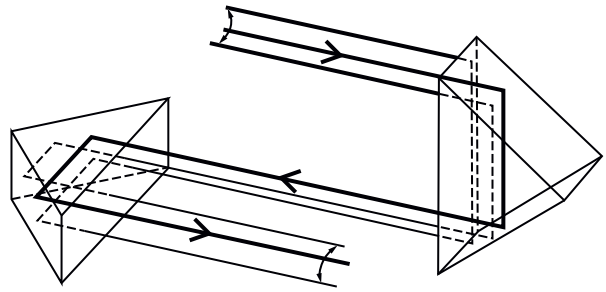


Fig. 8 Action of Porro prisms.

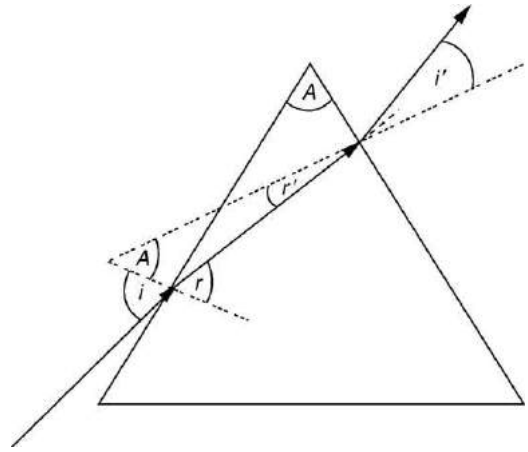


Fig. 9 Calculation of prism deviation.

an incident angle i, enters the prism at the angle of refraction r, crosses the prism and is refracted a second time when it emerges, the associated angles being designated as r' and r', When the ray leaves the prism, it is traveling in a direction which represents an angular deviation δ from its original orientation. This deviation is the sum of the individual changes in direction at each face, or

$$\delta = (i - r) + (r' - i') \tag{3}$$

For the triangle formed by the projected normals, which meet at the angle A:

$$r = r' + A \tag{4}$$

Hence

$$\delta = i + i' - A \tag{5}$$

The minimum value of the deviation is an important property of the prism; we find it by differentiation to obtain:

$$\mathrm{d}\delta = \mathrm{d}i + \mathrm{d}i' = 0 \tag{6}$$

Differentiation of the form of Snell's law valid at each interface, namely:

$$\sin i = n \sin r, \quad \sin i' = n \sin r' \tag{7}$$

gives

$$\mathrm{d}i = n \cos r \, \mathrm{d}r / \cos i \tag{8}$$

$$\begin{aligned} \mathrm{d}i' &= n \cos r' \, \mathrm{d}r' / \cos i' \\ &= n \cos r' \, \mathrm{d}r / \cos i' \end{aligned} \tag{9}$$

where $\mathrm{d}r'$ can be replaced by $\mathrm{d}r$, as obtained from the equation connecting r, r', and A. It now follows that:

$$\begin{aligned} \mathrm{d}\delta = n[&\{\cos r / \sqrt{(1 - n^2 \sin^2 r)}\} \\ &- \{\cos(A - r) / \sqrt{(1 - n^2 \sin^2 (A - r)}\}]\mathrm{d}r = 0 \end{aligned} \tag{10}$$

from which

$$\begin{aligned} &\cos^2 r\{1 - n^2 \sin^2 (A - r)\} \\ &\quad = \cos^2 (A - r)\{1 - n^2 \sin^2 r\} \end{aligned} \tag{11}$$

or:

$$\cos^2 r = \cos^2 (A - r) \tag{12}$$

and finally:

$$r = A/2, \; r' = A/2, \; i' = i \tag{13}$$

The differentiation thus leads to the conclusion that the ray enters and leaves the prism in a symmetrical manner. The amount of deviation is now:

$$\delta = 2i - A \tag{14}$$

or:

$$i = (\delta + A)/2 \tag{15}$$

and:

$$n = \sin i / \sin r = \sin\{(\delta + A)/2\} / \sin(A/2) \tag{16}$$

Using this expression to calculate n as a function of δ, it is found that δ is a minimum, so that a measurement of the minimum deviation produced by a prism gives the value of its index of refraction in a direct and accurate way. This method can also be used for liquids; a triangular glass cell with walls of reasonable thickness is made very accurately and used to measure δ. The refraction due to the walls does not affect the measurement, since the refraction occurring at the entrance wall is equal but opposite to that at the exit wall.

The description of prism behavior given so far assumes that we are dealing with monochromatic light; that is light of a single wavelength or single color in the visible spectrum. However, the index n of a prism varies with wavelength and white light passing through a prism will undergo dispersion – each component of a beam will emerge at a different angle and can be observed. The prisms that are part of expensive glass chandeliers display such an effect. This property of a prism is used in spectrometers – instruments which can measure the wavelength of light. **Fig. 10** shows a schematic diagram of this kind of instrument of high quality. Note that the light is brought into and taken out of the spectrometer with doublets. Various kinds of spectrometers have been put into use, with the ability to resolve closely spaced wavelengths being a crucial factor in the design.

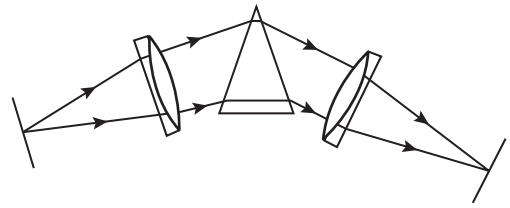


Fig. 10 A prism spectrometer.

Further Reading

Brouwer, W., 1964. Matrix Methods in Optical Instrument Design. New York: WA Benjamin.
Longhurst, R.S., 1967. Geometrical and Physical Optics, 2nd edn. New York: John Wiley & Sons.
MIL-HDBK-141, 1962. Military Standardization Handbook. Washington, DC: US Government Printing Office.

Acousto-Optics

M Gottlieb and D Suhre, Carnegie Mellon University, Pittsburgh, PA, USA

Nomenclature

b	dispersion constant (–)
c	elastic stiffness (dynes cm^{-2})
D,d	optical beam diameters (cm)
e	elastic strain (–)
f	acoustic frequency (hertz)
F	optical focal length (cm)
F_ϕ,F_θ	angular form factors (–)
k	optical momentum wavevector (cm^{-1})
K	acoustic momentum wavevector (cm^{-1})
M_2	acousto-optic figure of merit (sec^3 gram^{-1})
n	refractive index
p	photo-elastic constant (–)
P	pressure (dynes cm^{-2})
P_a	acoustic power (watts cm^{-2})
V	acoustic velocity (cm sec^{-1})
β	compressibility (cm^2 dyne^{-1})
	diffraction efficiency (–)
ε	dielectric constant (–)
ω	optical frequency (hertz)
Ω	solid angle (steradians)
ρ	density (grams cm^{-3})
θ	Bragg angles (radians)
τ	acoustic wave travel time (sec)

Acousto-optics deals with the interaction of light waves with sound waves and has given rise to a large number of devices related to various laser systems for display, information handling, optical signal processing, and numerous other applications requiring the spatial or temporal modulation of coherent light. The basic principles and phenomena controlling these interactions were largely understood by the mid-1930s, but had little practical significance because very high acoustic power levels were required to attain good optical efficiency and there were few good sources of well-collimated monochromatic light. During the period from 1960 to the present, several key technologies were developing rapidly, at the same time that many applications of the laser were being suggested which require high-speed, high-resolution scanning methods. These new technologies gave rise to high-efficiency, wideband acoustic transducers capable of operation to several gigahertz, high-power wideband solid-state amplifiers to drive such transducers, and the development of a number of new, synthetic acousto-optic crystals with low-drive-power requirements and low acoustic losses at high frequencies. This combination of properties makes acousto-optics feasible for many systems, and for several is the method of choice to satisfy demanding requirements. This chapter will summarize the basic features of acousto-optic interactions and the operating principles of the most common acousto-optic devices.

The Photo-Elastic Effect

The change induced in the refractive index of a transparent material by the pressure change produced by an acoustic disturbance is the underlying mechanism of all acousto-optic interactions. An acoustic wave produces periodic regions of compression and rarefaction in the material, which modulates the density. The Lorentz–Lorenz relation relates the refractive index to the density, for the simplest case of an ideal gas

$$\frac{(n^2-1)}{(n^2+2)} \propto \rho \tag{1}$$

where n is the refractive index and ρ is the density. This relation is followed to a good approximation for most simple solid materials as well. The elastooptic coefficient is

$$\frac{\rho\, \mathrm{d}n}{\mathrm{d}\rho} = \frac{(n^2-1)/(n^2+2)}{6n} = p \tag{2}$$

The fundamental quantity given by Eq. (2), also known as the photo-elastic constant p, is related to the pressure applied by

$$p = \frac{1}{\beta}\left(\frac{\mathrm{d}n}{\mathrm{d}p}\right) \tag{3}$$

where P is the applied pressure and β is the compressibility of the material. The photo-elastic constant of an ideal material with refractive index of 1.5 is 0.59. The photo-elastic constants of a wide variety of materials lie in the range from about 0.1 to 0.6.

The relation in Eq. (3) follows from the usual definition of the photo-elastic constant:

$$\Delta(1/\varepsilon) = \Delta(1/n^2) = pe \tag{4}$$

where ε is the dielectric constant, $\varepsilon = n^2$, and e is the acoustic strain amplitude. The induced change in refractive index, Δn, is

$$\Delta n = -n^3 pe \tag{5}$$

 doi:10.1016/B978-0-12-809283-5.00766-7

Strain amplitudes typically lie in the range of 10^{-8} to 10^{-5}, with Δn in the range of about 10^{-8} to 10^{-5} (for $n = 1.5$). Devices based upon such a small change in refractive index are capable of generating large effects because these devices are configured in a way that can produce large phase changes at optical wavelengths.

The more complete relation defining the photoelastic interaction is more complex than the simple scalar **Eq. (5)** in which the photoelastic constant is independent of the directional properties of the material. In fact, even for an isotropic material such as glass, longitudinal acoustic waves and transverse (shear) acoustic waves cause the photoelastic interaction to assume different parameters. A tensor relation between the dielectric properties, the elastic strain, and the photo-elastic coefficient describes the interaction, particularly for anisotropic materials. The tensor equation is

$$\Delta(1/n^2) = \sum_{kl} p_{ijkj} e_{kl} = \Delta(1/\varepsilon)_{ij} \tag{6}$$

where $(1/\varepsilon)_{ij}$ is a component of the optical index ellipsoid, e_{kl} are the Cartesian strain components, and p_{ijkl} are the components of the photo-elastic tensor. The crystal symmetry of the photoelastic material places limits on the possible configurations of interaction geometry.

Diffraction by Acoustic Waves

For typical acousto-optic devices the acoustic wave acts as a diffraction grating, made up of periodic changes in optical phase, moving at sound velocity. These features determine the properties of acousto-optic diffraction. Using a quantum-based model, the light and sound may be thought of as particles, photons and phonons, undergoing collisions in which energy and momentum are conserved. Either of these descriptions may be used to obtain all the important diffraction effects, but some are more easily understood on the basis of one or the other. A description of both is given here. Consider **Fig. 1** in which the light wave, of frequency ω and wavelength λ, is incident into the material with an acoustic wave of frequency f and wavelength Λ. If the refractive index of the medium is $n + \Delta n$ in the presence of the acoustic wave, the phase of the optical wave will be changed by an amount

$$\Delta\phi = 2\pi(L/\lambda)\Delta n \tag{7}$$

where L is the length of the interaction region. Some typical values of $\Delta\phi$ can be obtained by assuming $L = 2.5$ cm and $\lambda = 0.5$ μm, with Δn reaching a peak value of 10^{-5}. This yields a phase change of π rad, which is, of course, quite large. It is large because L/λ, the number of optical wavelengths, is 50 000, so that a very small Δn can still produce a sizable $\Delta\phi$. If the electric field incident on the delay line is represented by

$$E = E_0 e^{i\omega t} \tag{8}$$

then the field of the phase-modulated emerging light will be

$$E = E_0 e^{(i\omega t + \Delta\phi)} = e^{i\omega t} e^{2\pi i(L/\lambda)\sin(ft)} \tag{9}$$

where f is the acoustic frequency.

A detailed derivation of the resulting temporal and spatial distribution of the light field is mathematically complex, but, analogy with radio-wave modulation can be used to arrive at the resultant fields. The spectrum of a phase-modulated carrier of frequency ω consists of components separated by multiples of the modulation frequency f, as shown in **Fig. 2**. Sidebands are produced about the carrier frequency, such that the frequency of the mth sideband is $\omega + mf$, where m may be positive or negative. The amplitude of each of the sidebands is proportional to the Bessel function of order equal to the sideband number, and whose argument is the modulation index $\Delta\phi$. The odd-numbered negative orders are 180° out of phase with the even-numbered ones, a feature that may be useful for certain signal processing applications. The light emerging from the delay line is composed of a

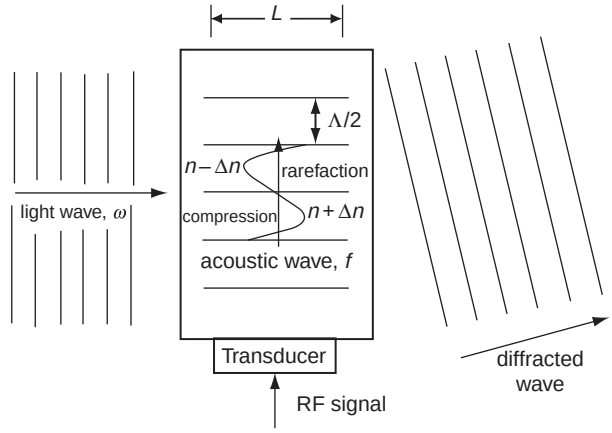


Fig. 1 Interaction of light waves with acoustic waves.

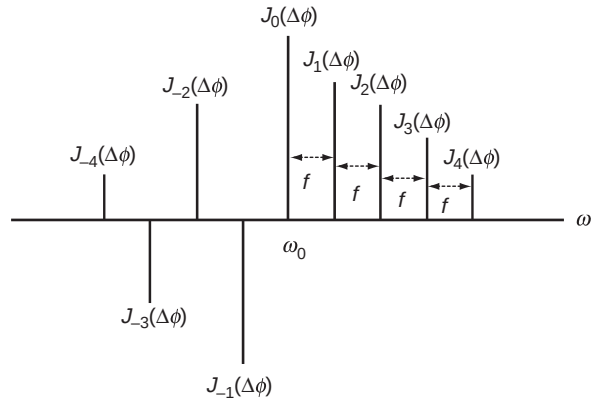


Fig. 2 Spectrum of a phase-modulated wave of carrier frequency ω_0, and modulation index $\Delta\phi$.

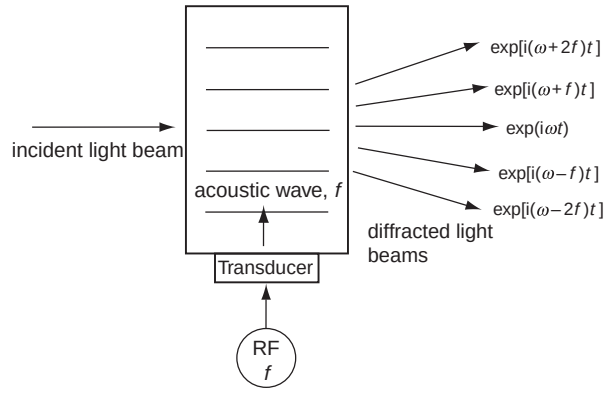


Fig. 3 Diffraction of light in the Raman–Nath limit.

number of light waves whose frequencies have been shifted by mf from the frequency ω of the incident light. The relative amplitudes will be determined by the peak change in the refractive index. The two extremes of the diffraction process are simplified in the 'thin grating' and the 'thick grating' models. The ratio of the interaction length, L, to the acoustic wavelength, Λ, will determine the character of the diffraction process. The plane wave approximation in which the acoustic energy propagates as a plane wave is valid when this ratio is very large. However, when this ratio is small, the acoustic propagation can be described in terms of a sum of plane waves, the angular spectrum of which increases as the ratio decreases. In this phenomenological model the partial wave which is propagating at an angle λ/Λ to the forward direction may diffract the light second time into an angle $2\theta = 2\Lambda/\Lambda$, and the frequency of this light will once again be up-shifted, for a total frequency shift of $2f$. If the spectrum of acoustic waves contains sufficient power of still higher orders, then this rediffraction process can be repeated, so that light will be multiply diffracted m times into higher-order angles, $m\theta = m\lambda/\Lambda$ each with a frequency shift mf. A similar argument holds for the negative orders, so that a complete set of diffracted light beams will appear as shown in Fig. 3, where the deflection angle corresponding to the mth order is given by $\sin\theta_m = m\lambda/\Lambda$ and the frequency of the light deflected into the mth order is $\omega_m = \omega + mf$. The intensity of the carrier wave, or zero order, will be zero when the modulation index $\Delta\phi$ is equal to 2.4. The first order will be a maximum for $\Delta\phi = 1.8$, corresponding to the maximum of the first-order Bessel function B_1, and decreasing for larger modulation indices. These phenomena were described by Debye and Sears and so are often referred to as Debye–Sears diffraction. An extensive theoretical analysis of the effect was given by Raman and Nath and is alternatively referred to as Raman–Nath diffraction. As the interaction length is increased the Raman–Nath diffraction gradually weakens. The weakening begins around an interaction length $L \sim \Lambda^2/4\lambda$. This value of L is expressed in the Q parameter

$$Q = 4L\lambda/\Lambda^2 \tag{10}$$

which is known as the Raman–Nath parameter. A different regime of diffraction takes effect for values of the interaction length $Q \gg 1$.

The phenomena in this regime can be more easily understood in terms of the quantum description of the light and sound waves as collisions between photons and phonons. In this model, the dynamics of the collisions of the light and sound are governed by the laws of conservation of energy and momentum. The momentum vectors magnitudes of the light and sound, k and

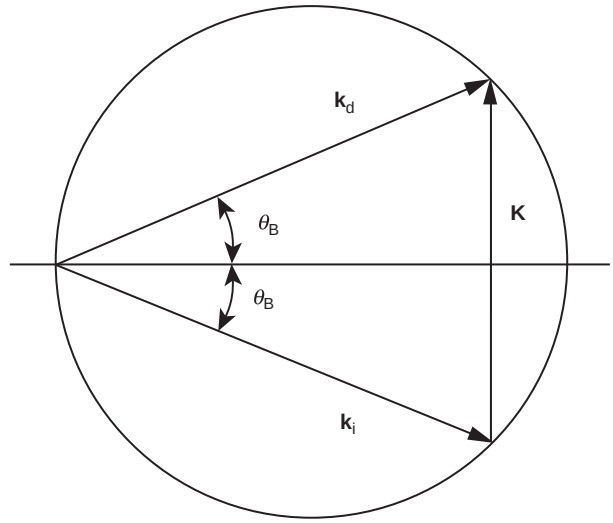


Fig. 4 Momentum diagram for diffraction of light.

K, are given by the well-known expressions

$$|\mathbf{k}| = 2\pi n/\lambda \text{ and } |\mathbf{K}| = 2\pi/\Lambda \tag{11}$$

where the acoustic wavelength, Λ, is related to the acoustic velocity, V, by $\Lambda = V/f$.

Conservation of momentum is expressed by the vector relation

$$\mathrm{k_i} + \mathrm{K} = \mathrm{k_d} \tag{12}$$

the diagram for which is shown in Fig. 3, where $\mathrm{k_i}$, and $\mathrm{k_d}$ represent the momentum of the incident photon and the diffracted photon, respectively. In this notation

$$\mathrm{k_i} = 2\pi n_i/\lambda, \tag{13a}$$

and

$$\mathrm{k_d} = 2\pi n_d/\lambda \tag{13b}$$

If the material is anisotropic (birefringent), n_i and n_d may be different.

Conservation of energy requires that

$$h\omega_d = h\omega_i + hf \tag{14}$$

in which h is Planck's constant. Since ω_i lies in the optical frequency range, $\sim 10^{13}$ Hz, and f lies in the RF or microwave range, 10^6–10^9 Hz, then $\omega_d \sim \omega_i$. This results, for isotropic materials, in the magnitudes of $\mathbf{k}_i$ and $\mathbf{k}_d$ being equal and resulting in the isosceles momentum triangle of Fig. 4. The angles of incidence and diffraction (with respect to the planar acoustic wavefronts), called the Bragg angles, are equal for this case, and are

$$\theta_B = (1/2)\lambda/\Lambda \tag{15}$$

A schematic diagram for this process is shown in Fig. 5. The interaction will be large only if the light is well-collimated, and incident at this angle. Unlike the Debye–Sears regime, there is only a single diffracted beam. The energy of the phonon may either be added to that of the incident photon, increasing its frequency to $\omega_d = \omega_i + f$, or the reverse, resulting in $\omega_d = \omega_i - f$. The sense of the momentum vectors determines which of these occurs. The Debye–Sears effect and Bragg diffraction are not different phenomena, but are the limits of the same mechanism. The Raman–Nath parameter Q determines which is the appropriate limit for a given set of values λ, Λ, and L.

Anisotropic Bragg Diffraction

In optically anisotropic materials, acousto-optic Bragg diffraction can occur between an ordinary and extraordinary optical beam, and vice versa. This is generally referred to as birefringent diffraction, and it offers additional design capabilities compared to the isotropic case, such as enhancing the angular aperture, and extending the aperture–bandwidth product.

The theory of diffraction of light so far described has assumed that the optical medium is isotropic. A number of important acousto-optic devices make use of the properties of birefringent materials. It is different from diffraction in isotropic media in that the magnitude of the momentum vector of the light will in general be different for different light polarization directions. Thus, the vector diagram representing conservation of momentum will no longer be the simple isosceles triangle of Fig. 4. The momentum vector for ordinary polarized light will, in general, be different from the momentum vector for light that is extraordinary polarized.

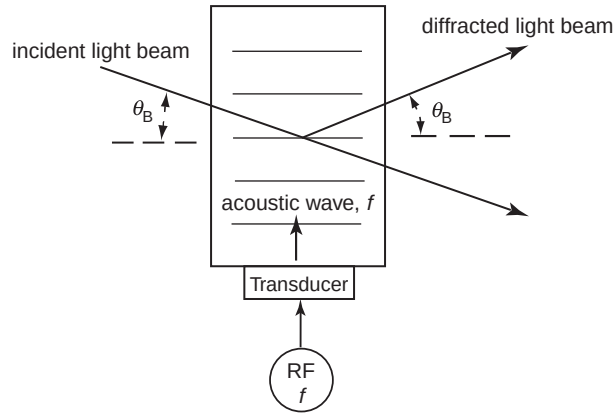


Fig. 5 Acousto-optic diffraction cell.

To understand the effect of anisotropy on diffraction, it is necessary to mention another phenomenon which occurs when light interacts with shear acoustic waves, i.e., waves in which the displacement of matter is perpendicular to the direction of propagation of the acoustic wave. A shear acoustic wave may cause the direction of polarization of the diffracted light to be rotated by 90°. The underlying reason for this is that the shear disturbance induces a birefringence which causes the plane of polarization to be rotated. The acousto-optic tunable filter (AOTF) is a particularly interesting birefringent device. It was first developed for the case of collinear optical and acoustic beams, and was used for wavelength selection in a dye laser.

Diffraction Efficiency

The acoustic power required to diffract light will be determined by the geometric parameters as well as by the properties of the medium. A simplified calculation leads to results that are useful. Referring to the spectrum of a phase-modulated wave shown in Fig. 2, we can see that the ratio of the intensity in the first order to that in the zero order is

$$I_1/I_0 = [B_1(\Delta\phi)/B_0(\Delta\phi)]^2 \tag{16}$$

By analogy this same result comes about for acousto-optically diffracted light. The acoustic power flow is given by

$$P_a = \frac{1}{2}cVe^2 \tag{17}$$

where c is the elastic stiffness constant. The elastic stiffness constant is related to the bulk modulus β, the density ρ, and acoustic velocity v through the expression

$$c = \frac{1}{\beta} = \rho V^2 \tag{18}$$

and the acoustic power density is

$$P_a = \frac{1}{2}\rho V^3 e^2 \tag{19}$$

The optical phase modulation of the medium resulting from the change in refractive index, Δn, is

$$\Delta\phi = 2\pi(L/\lambda)\Delta n \tag{20}$$

Using Eq. (5) for Δn, the phase modulation is related to the acoustic power density by

$$\Delta\phi = -\pi(L/\lambda)n^3 p\left(2P_a/\rho V^3\right)^{1/2} \tag{21}$$

The diffraction efficiency can now be calculated by using this value for the optical phase change in Eq. (16). The first- and second-order Bessel functions can be approximated for small modulation index by

$$B_0(\Delta\phi) \sim \cos(\Delta\phi) \sim (\Delta\phi)^2, \text{ and}$$
$$B_1(\Delta\phi) \sim \sin(\Delta\phi) \sim \Delta\phi \tag{22}$$

The small signal approximation to the diffraction efficiency is then

$$I_1/I_0 \sim (\Delta\phi)^2 = \left(\pi^2 L^2/2\lambda^2\right)\left(n^6 p^2/\rho \mathrm{V}^3\right) P_\mathrm{a} \tag{23}$$

The expression $n^6p^2/\rho V^3$ is known as the figure of merit of the material, and is designated as M_2; it is comprised entirely of intrinsic material properties.

Acousto-optic Materials

Acousto-optic device technology has matured to the point that performance is chiefly limited by material parameters, particularly the figure of merit and acoustic attenuation. Nature has arranged that materials with high figures of merit usually have high attenuation and vice versa. The widely used acousto-optic materials are fused quartz, tellurium dioxide, and lithium niobate. Development work on new infrared materials has been reported recently. A list of commonly used acousto-optic materials is given in Table 1. For materials with a low figure of merit, a higher drive power can be used to obtain the required efficiency.

Experience has indicated that the upper limit for very small devices (active area ~ 0.1 mm^2) is a drive power density of 100–500 mW/mm^2, provided there is proper heat sinking to transfer the heat energy. For larger devices, sizes greater than ~ 1 cm^2, the limit is closer to a few W/cm^2. At the high drive power levels, the acoustic attenuation may cause significant optical distortion.

AO Devices

Resolution, bandwidth, and speed are the important characteristics of acousto-optic scanners, shared by all types of scanning devices. Depending upon the application, only one, or all, of these characteristics may have to be optimized. Consider the acousto-optic scanner in Fig. 6 with a collimated incident beam of width D, diffracted to an angle θ_B at its RF bandcenter, and whose bandwidth is Δf. If the diffracted beam is focused onto a plane by a lens, or lens combination, at the scanner, the diffraction spread of the optical beam will be

$$\Delta x = F\Delta\phi \sim F - \lambda/D \tag{24}$$

Table 1 Selected acousto-optic materials

Material	*Transmission range (μm) mission*	*Acoustic mode & propagation direction*	*v (cm/s × 10^5)*	*Acoustic attenuation (dB/cm GHz^2)*[c]	n	*M_2 (s^3/g × 10^{-18})*
Visible–near-infrared (VIS-NIR)						
$LiNbO_3$	0.04–4.5	L[100][a]	6.57	0.15	2.20	7.0
		S[001][b]	3.59	2.6	2.29	2.92
TiO_2,	0.45–6	L[001]	10.3	0.55	2.58	1.52
α-SiO_2;	0.12–4.5	L[001]	6.32	2.1	1.54	1.48
		L[100]	5.72	3.0	1.55	2.38
TeO_2.	0.35–5	L[001]	4.20	15	2.26	34.5
		S[110]	0.616	90	2.26	793
Far IR						
Ge	2–20	L[111]	5.50	30	4.00	840
$T1_3AsS_4$ (TAS)	0.6–12	L[001]	2.5	29	2.63	510
GaAs	1–11	L[110]	5.15	30	3.37	104
ZnTe	0.55–20	L[110]	3.37	130	2.77	18
GaP	0.6–10	L[110]	6.32	60	3.31	30
$T1_3PSe_4$	0.85–9	L[100]	2.0	150	2.9	2069
Te	5–20	L[100]	2.2	60	4.8	4400
CdS	0.5–11	L[100]	4.17	90	2.44	12
GaP	0.6–10	L[110]	6.32	60	3.31	30
ZnS	0.4–12	L[001]	5.82	27	2.35	3.4
		S[001]	2.63	130	2.35	8.4

[a]Longitudinal acoustic waves propagating along the [100] crystallographic direction.
[b]Shear acoustic waves propagating along the [001] crystallographic direction.
[c]Attenuation constant at 1 GHz; the frequency dependence of the attenuation for most crystals is quadratic.

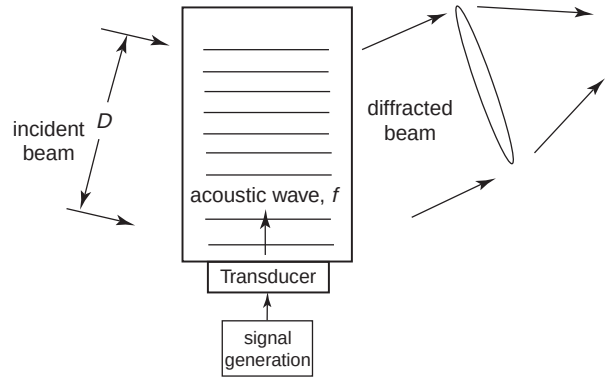


Fig. 6 Acousto-optic scanner.

where F is the focal length of the lens. The light intensity will be distributed in the focal plane with a sinc^2 distribution by aperture diffraction. The number of resolvable spots, N, will be the angular scan range divided by the angular diffraction spread,

$$N = \frac{\Delta\theta}{\Delta\phi} \tag{25}$$

where $\Delta\theta$ is the range of the angular scan. Differentiating the Bragg angle formula yields

$$\Delta\theta = (\lambda/V \cos\theta_0)\Delta f \tag{26}$$

and

$$N = \Delta f(D/V \cos\theta_0) = \Delta f \tau \tag{27}$$

where θ_0 is the Bragg angle at band center, and τ is the time that it takes the acoustic wave to cross the optical aperture. The resulting expression is the time–bandwidth product of the acousto-optic scanner, a concept applied to a variety of electronic devices as a measure of information handling capacity. The time–bandwidth product of an acousto-optic Bragg cell is equivalent to the number of bits of information which may be instantaneously processed by the system. In order to maximize the number of resolution elements, it is desirable to have as large a bandwidth as possible (i.e., large frequency range) and also as large an aperture delay time as possible. There are two factors limiting the bandwidth of an acousto-optic device: the bandwidth of the transducer structure (discussed later), and acoustic absorption in the delay medium. The acoustic absorption increases with increasing frequency; for high-purity single crystals the increase generally goes with the square of the frequency. For glassy materials, on the other hand, the attenuation will increase more slowly with frequency, often approaching a linear function. The maximum frequency is generally taken as that for which the attenuation of the acoustic wave across the optical aperture is equal to 3 dB. A reasonable approximation of the maximum attainable bandwidth is

$$\Delta f = 0.7 f_{\max} \tag{28}$$

Birefringent Scanners

The birefringent scanner is a significant use for anisotropic Bragg diffraction. There are a number of advantages with the birefringent scanner over the anisotropic scanner, such as a larger angular scan range along with lower frequencies. There are also disadvantages, such as lower speed due to generally lower shear wave velocities, and the particular application will dictate whether an isotropic or birefringent scanner is better. Scanners represent a fairly important aspect of acousto-optic devices due to the widespread commercial use of laser beam deflectors for displays and laser printers. For applications where the required speed is beyond that of mechanical scanners, the acousto-optic scanner is an ideal candidate. However, unlike mirrors, acousto-optic deflectors are wavelength sensitive, and can only be used with single-wavelength laser beams.

The birefringent scanner can be described with the wavevector diagram, as shown in Fig. 7 for a positive uniaxial crystal where the extraordinary index of refraction is larger than the ordinary. The acoustic wavevector is tangent to the diffracted surface, which produces the largest scan angle for a given acoustic bandwidth. This is also the degenerate case, where only a single diffracted beam results, whereas two beams at two different acoustic frequencies will generally result from arbitrary input and acoustic propagation directions. The azimuthal acceptance angles (angles normal to the plane of incidence) of the acoustic and optical wavevectors can be different, although propagation and polarization directions must be selected that will produce high efficiency.

For the positive uniaxial case, the acoustic and diffracted wavevectors will be perpendicular at the design point, which allows the center frequency to be calculated from geometry as

$$f_0 = (V/\lambda)\left(n_i^2 - n_0^2\right)^{1/2} \tag{29}$$

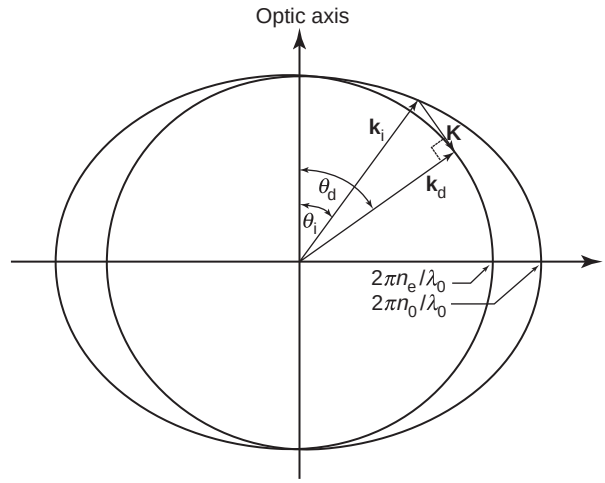


Fig. 7 Wavevector diagram for birefringent scanner using a positive uniaxial crystal.

where n_i is the index of refraction at the incidence angle. Due to the typically low velocity of shear waves, this frequency can be much lower than an isotropic scanner, and the incidence angle can also be chosen to adjust the frequency. It is also possible to use the optical activity of certain materials such as TeO_2, along with circularly polarized light to further reduce the acoustic frequency. It is important to maintain as low a frequency as possible due to acoustic attenuation, which is especially high with the soft materials typically used for birefringent applications.

The bandwidth over which the scanner can efficiently operate is fairly large due to having the acoustic and diffracted wavevectors perpendicular, and is approximately $\Delta f = 2f_0$, assuming an octave of bandwidth for the isotropic scanner. This will produce a larger scan angle than an isotropic scanner, and more spots of resolution. The number of resolution spots is determined through diffraction by the aperture size and diffraction, along with the scan angle. Since a larger aperture requires a longer time for the acoustic waves to propagate across the aperture, the response time τ to access the scanner will increase, and the product of τ and f is related to the number of spots by **Eq. (27)**. The advantage of the birefringent scanner is that it can operate at a lower frequency f_0 with better performance. However, since other factors such as acoustic attenuation are important in designing a scanner, for some applications it might be better to operate an isotropic scanner at a higher frequency.

AOTFs

With anisotropic Bragg diffraction, the magnitude of the diffracted wavevector $\mathbf{k}_d$ will differ from the incident wavevector $\mathbf{k}_i$, which cannot occur for the isotropic case. This is illustrated in **Fig. 8** for an AOTF utilizing a negative uniaxial crystal where an extraordinary input wave propagating at an incident angle θ_i relative to the crystal axis is diffracted into an ordinary output wave at an angle θ_d. This occurs through an acoustic wave propagating at an angle θ_a with wavevector $\mathbf{K}_a$, where all the wavevectors lie in a plane through the optic axis of the crystal. This diagram is identical to the index ellipsoid for the crystal, scaled by $2\pi/\lambda$, where λ is the vacuum wavelength. The acoustic wavevector is shown as adding to the incident wavevector, which increases the frequency of the optical wave by the acoustic frequency. It can also be represented in the reverse direction, which would reduce the optical frequency by the same amount.

An AOTF spectrally filters the optical input, and maintains its spectral purity over a large angular aperture. These conditions are maximized when the power flow or ray directions of the input and output beams are collinear. This produces a parallel-tangents condition, where lines drawn tangent to the wavevector surfaces and connecting the input and diffracted wavevectors are parallel. For beams at 90° to the optical axis, it is referred to as the collinear case, whereas all other parallel-tangents conditions are noncollinear. The diffracted beam is rotated by 90° during anisotropic Bragg interaction, so that crossed polarizers can be used to separate the input and diffracted beams. For the collinear case polarization separation must be used, whereas angular separation can also be used for the noncollinear case.

The AOTF design equations can be derived from the geometrical conditions imposed by the wavevector matching condition

$$\mathbf{k}_d = \mathbf{k}_i + \mathbf{K} \tag{30}$$

along with decomposing the acoustic wave into its Fourier components which result from the finite interaction length produced by the transducer. These components allow for phase matching over a range of input angles, and they also allow for a spectral spread of the interaction. The full width half maximum (FWHM) resolution is given by

$$\Delta\lambda = \frac{1.8\lambda^2}{bL \sin 2\theta_i} \tag{31}$$

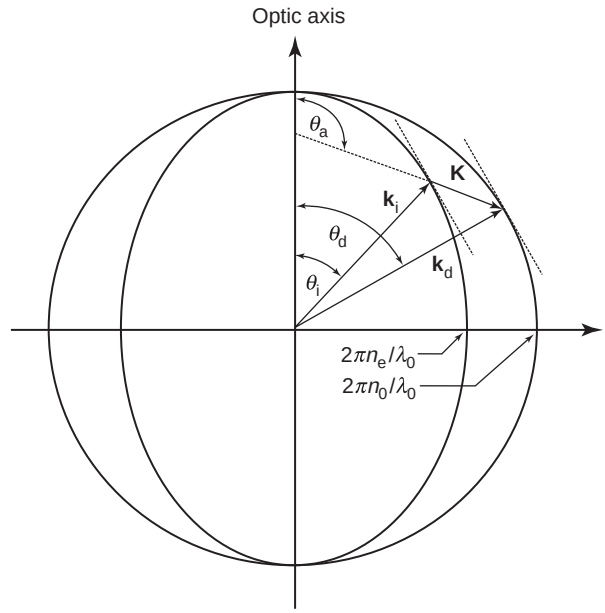


Fig. 8 Wavevector diagram for an AOTF using a negative uniaxial crystal.

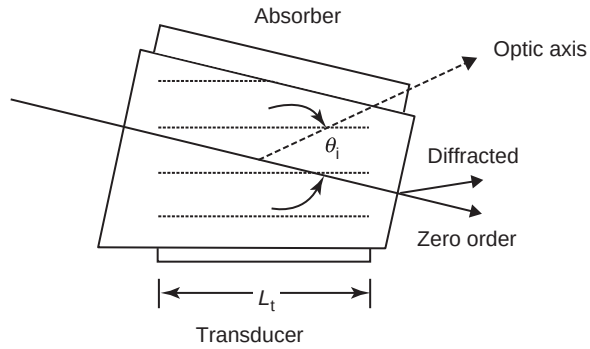


Fig. 9 Noncollinear AOTF orientation of optical and acoustic beams.

where λ is the vacuum wavelength, b is the dispersion term (essentially $2\pi\Delta n$, where the birefringence Δn is the difference between the ordinary index n_0 and extraordinary index n_e of refraction), L is the interaction length of the input beam defined geometrically by the acoustic beam, and θ_i refers to the angle of the central or cardinal ray of the input beam, and the AOTF is designed around this angle. It is therefore a sensitive function of the incidence design angle, with the highest resolution occurring for the collinear case. The resolution is also narrower for larger birefringence.

The geometry of a noncollinear AOTF is shown schematically in **Fig. 9**. The acoustic waves propagate in the correct direction and are generated by a transducer of length L_t, which is related geometrically to the interaction length. These waves are generally absorbed at the other side of the AOTF to prevent interfering reflections, and the sides may also be wedged to eliminate reflections into the interaction region. The optical beam enters the AOTF at the correct angle to the crystal optic axis, and either diffracts or passes through as the zero order. The input beam will have an angular spread, producing an acceptance angle that is a function of the interaction length, wavelength, and crystal parameters, and the external solid angular aperture is roughly given by

$$\Delta\Omega = \frac{\pi n^2 \lambda}{\Delta n L} \tag{32}$$

where n refers to the diffracted beam index of refraction, which can be either n_0 or n_e depending upon the design. The resolution and solid angle are therefore related through the transducer length, and the product of the resolving power times the solid angle is given by

$$(\lambda/\Delta\lambda)(\Delta\Omega) = 2\pi n^2 \sin^2\theta_i / |F_\theta F_\phi|^{1/2} \tag{33}$$

where

$$F_\theta = 2\cos^2\theta_i - \sin^2\theta_i \tag{34a}$$

$$F_\varphi = 2\cos^2\theta_i + \sin^2\theta_i \tag{34b}$$

are form factors for the polar and azimuthal components of the optical field. Since all the wavectors lie in a plane containing the optic axis, the azimuthal components of the incident, diffracted, and acoustic wavevectors must always be identical under the parallel-tangents condition.

The product $(\lambda/\Delta\lambda)(\Delta\Omega)$ forms a figure of merit for spectrometers, and for the collinear case, the product is identical to a Fabry–Perot etalon having an index of n, and at other angles it is more complicated due to the angular dependence. The angular field is also symmetrical for the collinear AOTF, but nowhere else other than along the optic axis, where the resolving power becomes zero.

The great advantage of the AOTF designed under the parallel-tangents condition is the large angular aperture compared to the general case. Under this condition, the dependence of the angular aperture on resolving power is second order rather than first order, and the angular aperture can be tens of degrees for a typical resolution, which is useful for imaging applications. At the special angle of 54.7°, the second-order dependence is also zero, and the aperture can be extremely large relative to the resolving power, although this requires a high-resolution device with a corresponding narrow field, since the condition would soon be violated at larger field angles.

The efficiency of the AOTF under phase matching is given by

$$\eta = \sin^2\left[\pi^2 M_2 P_a L^2/\left(2\lambda^2\right)\right]^{1/2} \tag{35}$$

which depends on θ_i and various material properties. Since the interaction is anisotropic, M_2 must be taken as a tensor quantity, in which both the polarization and propagation directions of the acoustic and optical waves must be accounted for. For acoustic waves, the polarization direction is the particle motion, which is perpendicular to the propagation direction for shear waves. In general, M_2 is much larger for a specific configuration of a particular material, and AOTFs are designed around this condition. The range of useful design angles is also usually limited since M_2 is generally a sensitive function of θ_i. The angular dependence on M_2 is important in designing an efficient device since it can be near zero at specific design angles for some materials, such as with TeO_2 for the collinear case.

The AOTF must be designed to optimize performance. The resolution is generally given as a system requirement, and the design angle along with the transducer length must be adjusted to optimize the throughput, or total optical power through the AOTF. This requires maximizing the efficiency, angular aperture, and aperture dimension. The maximum crystal size that can be grown ultimately limits these parameters, both in the interaction length and in the aperture size.

AOTFs have been used for a wide range of spectral filtering applications for spectroscopy and laser applications. Both collinear and noncollinear devices have electronically tuned a variety of lasers, including dye, semiconductor, and Ti:sapphire lasers. The AOTF is placed within the laser cavity, which requires high transmission on the laser line, with enough resolution to separate nearby laser transitions. A related application is the multiplexing of optical communications systems, where the AOTF is used to separate the various laser diode wavelengths. Perhaps the most widely used application is spectroscopy, and single-point detection systems have been used for a variety of biological and chemical applications in the visible and infrared. These techniques have been extended to spectral imaging, again in the visible and the infrared. Due to a variety of AOTF aberrations the image quality will be somewhat degraded. Since the AOTF is electronically tunable, it can be rapidly modulated both in amplitude and wavelength, which allows for modulation spectroscopy to detect small signals in a large background. By applying multiple frequencies, the AOTF has also been used as a rejection filter, in which all wavelengths other than that selected are allowed to pass.

Modulators

Acousto-optic modulators are used to vary and control laser beam intensity. A Bragg configuration gives a single first-order output beam, where intensity is directly linked to the power of RF control signal. The rise time of the modulator is the time required for the acoustic wave to traverse the laser beam. For minimum rise time the laser beam will be focused down, forming a beam waist as it passes through the modulator.

A Bragg configuration gives a single first-order output beam, the intensity of which is controlled by the RF signal power, and diffaction angle controlled by the drive frequency. Two deflectors can be used in series and at right angles to each other to give full two-dimensional scanning.

One requirement in the design of acousto-optic modulators is that the diffracted beam and undiffracted beam must be well separated. For an adequate extinction ratio, the Bragg angle should be at least as large as the divergence of the optical beam. This condition puts a minimum value on the center frequency. Equating the Bragg angle

$$\theta_B = \frac{\lambda f_0}{2nV} \tag{36}$$

and the diffraction angle of the Gaussian beam

$$\delta\theta_0 = \frac{4\lambda}{\pi(nd)} \tag{37}$$

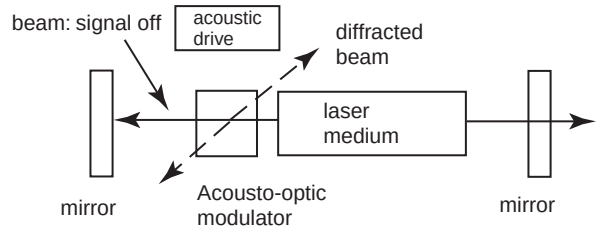


Fig. 10 An acousto-optic *Q*-switch in a laser cavity.

where d is the beam diameter at the minimum, it follows that the lower limit of the acoustic frequency is given by

$$f_0 = \frac{8}{\pi\tau} \tag{38}$$

We may combine these expressions to determine a limit of modulator bandwidth of acousto-optic modulators, with the result

$$\Delta f \sim \frac{1}{4} f_m \tag{39}$$

i.e., the modulation bandwidth is approximately equal to 25% of the midband acoustic frequency, f_m. In view of the present status of transducer technology, the modulation bandwidth of acousto-optic modulators is limited to several hundred MHz.

A frequency shifter uses the shift inherent in the acousto-optic interaction to up- or down-shift a laser's frequency. Two kinds of shifters can be distinguished: the fixed frequency shifter, and the variable frequency shifter. The frequency shift is equal to the signal frequency applied to the transducer. Various device configurations can be used to shift the laser beam, such as multiple travels inside the shifter to double or triple the frequency shift, or a combination of two frequency shifters in series. Acousto-optic frequency shifters can be used, for example, in optical heterodyning and interferometry, or in laser Doppler velocimetry for particle velocity analysis.

Q-Switches

A *Q*-switch is a special modulator which introduces high repetition rate losses inside a laser cavity (typically 1 to 100 KHz). They are designed for minimum insertion loss and to be able to withstand very high laser powers. In normal use an RF signal is applied to diffract a portion of the laser cavity flux out of the cavity. This increases the cavity losses and prevents oscillation. When the RF signal is switched off, the cavity losses decrease rapidly and an intense laser pulse evolves. A diagram of a typical acousto-optic *Q*-switched laser cavity is shown in **Fig. 10**. *Q*-switches are the preferred method of intracavity modulation where high speed, high stability, and modest cost are important design factors.

Further Reading

Chang, I.C., 1976. Acousto-optic devices and applications. IEEE Transactions: Sonics and Ultrasonics SU-23, 2–22.
Chang, I.C., 1981. Acousto-optic tunable filter. Optical Engineering 20, 824–829.
Dixon, R.W., 1967. Photo-elastic properties of selected materials and their relevance for applications to acoustic light modulators and scanners. Journal of Applied Physics 38, 5149–5153.
Gordon, E.I., 1966. A review of acousto-optical deflection and modulation devices. Proceedings of the IEEE 54, 1391–1401.
Gottlieb, M., Ireland, C.L.M., Ley, J.M., 1983. Electro-optic and Acousto-optic Scanning and Deflection. New York: Marcel Dekker.
Goutzoulis, A.P., Pape, D.R., 1994. Design and Fabrication of Acousto-optic Devices. New York: Marcel Dekker.
Uchida, N., Niizeki, N., 1973. Acousto-optic deflection materials and techniques. Proceedings of the IEEE 61, 1073–1092.
Xu, J., Stroud, R., 1992. Acousto-optic Devices: Principles, Design, and Applications. New York: Wiley.

Electro-Optics

LR Dalton, University of Washington, Seattle, WA, USA

The electro-optic (EO) effect can be described as any one of a number of phenomena that occur when an electromagnetic wave in the optical spectrum (e.g., characterized by wavelengths in the range 200 to 2000 nm) interacts with an electric field, or with matter under the influence of an electric field. Two of the most important electro-optic phenomena are the Kerr effect (discovered by John Kerr in 1875) and the Pockels effect (discovered by Friedrich Pockels in 1893), in which birefringence is induced or modified in a liquid (the Kerr effect) or a solid (the Pockels effect). Birefringence is the difference in refractive indices for light of orthogonal line polarizations, one of which is parallel to the induced optical axis. The Kerr effect involves creation of birefringence in a liquid that is otherwise not birefringent. The degree of birefringence is quadratically proportional to the applied electric field strength. Hence, the Kerr effect is frequently referred to as the quadratic electro-optic effect while the Pockels effect involves a linear dependence on the applied electric field and is referred to as the linear electro-optic effect. In considering practical applications of the electro-optic effect, a commonly encountered term is that of 'electro-optic modulator', which can be described as a device wherein a signal control element is used to modulate a beam of light. The control element is typically an electric field with frequency components in the zero (DC) to hundreds of gigahertz ($1\text{GHz} = 10^9\text{Hz}$) range (or even tens of terahertz ($1\text{THz} = 10^{12}\text{Hz}$). The modulation may be imposed on the phase, frequency, amplitude, or direction of the modulated optical beam.

Electro-optic effects are one class of second-order nonlinear optical phenomena. Other important second-order nonlinear optical phenomena include sum (e.g., second-harmonic) and difference frequency generation. Such phenomena derive from the second-order term in the power series expansion of macroscopic (material) polarization, P, in terms of applied electric fields:

$$P_i = \chi^{(1)}_{ij} E_j + \chi^{(2)}_{ijk} E_j E_k + \chi^{(3)}_{ijkl} E_j E_k E_l + \ldots \tag{1}$$

where $\chi^{(1)}_{ij}$, $\chi^{(2)}_{ijk}$, and $\chi^{(3)}_{ijkl}$ are the linear, second-, and third-order optical susceptibilities, respectively. Second-harmonic generation (SHG) or frequency doubling (2ω) effects, where a beam at twice the frequency of the input beam is generated, can be seen by substituting a sinusoidal field, $E_\omega \cos(\omega t - kz)$; for E_j and E_k. After using a well-known trigonometric identity, the equation for macroscopic polarization becomes (through second-order and dropping the subscript i):

$$P = \chi^{(1)} E_\omega \cos(\omega t - kz) + (1/2)\chi^{(2)} E_\omega^2 [1 + \cos(2\omega t - 2kz)] + \ldots \tag{2}$$

The electro-optic effect can be appreciated by considering interaction of the medium with an optical field $E_\omega \cos(\omega t - kz)$ and a low-frequency field E_0. Note that although we have used the symbol zero to denote the frequency of the low-frequency electrical field, actual frequencies can extend to hundreds of gigahertz or even tens of terahertz. Again substituting (and now keeping terms to third order), the equation for polarization can be written as:

$$P = \chi^{(1)} E_\omega \cos(\omega t - kz) + 2\chi^{(2)} E_0 E_\omega \cos(\omega t - kz) + 3\chi^{(3)} E_0^2 E_\omega \cos(\omega t - kz) + (3/4)\chi^{(3)} E_\omega^3 \cos(\omega t - kz) = \chi_{\text{eff}} E_\omega \cos(\omega t - \text{kz}) \tag{3}$$

The equation for the nonlinear index of refraction, n, can be expressed as

$$\begin{aligned} n^2 &= 1 + 4\pi\chi_{\text{eff}} \\ &= 1 + 4\pi\left[\chi^{(1)} + 2\chi^{(2)} E_0 + 3\chi^{(3)} E_0^2 + (3/4)\chi^{(3)} E_\omega^2\right] \end{aligned} \tag{4}$$

Denoting the linear index of refraction as n_0, the above equation can be rewritten as

$$n^2 - n_0^2 = 8\pi\chi^{(2)} E_0 + 12\pi\chi^{(3)} E_0^2 + 3\pi\chi^{(3)} E_\omega^2 \tag{5}$$

which in turn leads to

$$n = n_0 + (4\pi/n_0)\chi^{(2)} E_0 + (6\pi/n_0)\chi^{(3)} E_0^2 + (3\pi/2n_0)\chi^{(3)} E_\omega^2 \tag{6}$$

The definition of light intensity in cgs units is $E_\omega^2 = (8\pi/cn_0)I(\omega)$, which when substituted into the above equation gives

$$n = n_0 + (4\pi/n_0)\chi^{(2)} E_0 + (6\pi/n_0)\chi^{(3)} E_0^2 + \left(12\pi/cn_0^2\right)\chi^{(3)} I(\omega) \tag{7}$$

The index of refraction can now be written as

$$n(\omega) = n_0(\omega) + n_1(\omega) E_0 + n_2(0) E_0^2 + n_3(\omega) I(\omega) \tag{8}$$

where the terms $n_1(\omega)$, $n_2(0)$, $n_3(\omega)$ correspond to the linear electro-optic effect, the quadratic electro-optic effect, and the optical Kerr effect, respectively. The above equation can also be expressed as

$$n = n_0 - (1/2)n^3 r E_0 - (1/2)n^3 s E_0^2 - \ldots \tag{9}$$

where r and s are the linear and quadratic electro-optic coefficients, respectively.

Unfortunately, a great deal of confusion exists concerning the use of the terms electro-optic effect and electro-optic modulator. First of all, the terms 'electro-optic' and 'opto-electronic' are frequently confused and used interchangeably. Opto-electronic devices function as electrical-to-optical or optical-to-electrical signal transducers. Examples of commonly used opto-electronic devices include light-emitting diodes (LEDs), photodiodes (PDs), injection laser diodes (ILDs), and integrated optical circuit (IOC)

elements. An electro-optic device can also function as an electrical-to-optical signal transducer (e.g., a Mach–Zehnder interferometer or a birefringent modulator employing polarizers at the input and the output); however, the mode of operation is entirely different for electro-optic and opto-electronic devices. For example, with LEDs or modulated lasers, the applied electric field produces modulation of transmission of light by altering the excited state population of the light emitting (lasing) state of the material. With electro-optic materials, no actual excited state population is generated. Electro-optic devices typically operate in regions of the optical spectrum removed from resonant transitions, i.e., regions of relatively high transparency. The applied electric field acts only to perturb the charge distribution of the material (the spatial positions of the electrons and nuclei). The altered charge distribution interacts with the optical beam transmitting the material with the result that the speed of light in the material (i.e., the index of refraction or birefringence) is altered. The response times of opto-electronic and electro-optic phenomena (and hence bandwidths of devices exploiting these phenomena) are quite different. In the former case, response will be limited by the lifetime of the relevant (emitting) excited state while in the latter case, response will be defined by the relaxation time of the material (e.g., reorientation time of a liquid or lattice relaxation time of a solid) following removal of the perturbing electric field.

An even greater confusion can arise due to the jargon used in particular technologies, such as telecommunications. A telecommunications engineer frequently refers to 'electro-optic switching' when describing opto-electronic transduction of an optical signal to an electrical signal, followed by re-routing in the electrical domain, and finally opto-electronic transduction of the electrical signal back to the optical domain. Electro-optic switching (in the sense used in this article) involves quite a different operation and would be described by that same telecommunications engineer as 'all-optical switching'–a term reserved by physicists for the optical Kerr effect (control of one light beam by another light beam).

A second point of confusion involves distinguishing between the terms 'electro-optic' and 'electro-absorptive' modulation. Again, the terms are frequently used interchangeably. However, they involve quite different physical mechanisms. An electro-absorptive modulator is, like a modulated laser or LED, a 'resonant' device. In a device such as a gallium arsenide (GaAs) or indium phosphide (InP) quantum dot electro-absorptive modulator, the position and width of an optical absorption (resonant transition) are defined by the physical dimensions of the quantum dot. Application of an electric field shifts the spectral wavelength of the quantum dot absorption. This phenomenon can be used to dramatically change optical transmission through (or equivalently, absorption by) the material by choosing the wavelength of the propagating beam of light to be near the resonant absorption. With the electric field off, the material is reasonably transparent but becomes strongly absorbing when the electric field is applied, shifting the resonant absorption to the wavelength of the propagating beam. Obviously, the voltage required to achieve a desired change in optical transmission will depend on control of quantum dot dimensions in fabrication, which has been and continues to be a topic of research focus for electro-absorptive materials. Likewise, control of insertion losses of such devices is a concern as the wavelength used for the transmissive state of the device must be sufficiently close to the optical resonance to achieve the desired attenuation with application of a modest control voltage. Another issue to be faced with electro-absorptive devices is that of 'chirp' (optical signal distortion arising from the fact that both absorption and index of refraction – the imaginary and real parts of the optical susceptibility – change with application of the electric control field). On the positive side, electro-absorptive modulators are 'polarization-insensitive' modulators and are thus conveniently used with multimode (as well as single mode) optical transmission.

Further confusion exists because widely different types of materials can be used in electro-optic and electro-absorptive devices. As these materials are frequently competing for the same applications (signal transduction, optical switching, etc.), the advantages and disadvantages of different materials are often compared without maintaining the context that quite different phenomena are involved. Even restricting our discussion to electro-optic materials, we note that materials can range from organic liquid crystalline materials (nematic, smetic, ferroelectric, etc.) to organic electro-optic materials (both crystalline and 'macromolecular') to inorganic crystalline materials (lithium niobate, $LiNbO_3$; barium titanate, $BaTiO_3$; barium borate, BBO; potassium dihydrogen phosphate, KDP; lithium tantalate, $LiTaO_3$; zinc telluride, ZnTe, etc.). The response times of these materials to transient application of an applied electric field will be quite different, reflecting the different types of lattice motion involved. With liquid crystalline materials, particularly nematic materials, considerable molecular reorientation is involved and due to the mass that must be moved, response times are quite slow although the index change will be relatively large. With use of liquid crystalline materials, device bandwidths will typically be limited to tens of megahertz (MHz) or less. In contrast, conjugated π-electron organic materials typically exhibit response times of tens to hundreds of femtoseconds, which translates to potential device bandwidths of tens of terahertz (actual device bandwidths may be less due to factors other than material response). For conjugated π-electron organics, the response time is the phase relaxation time of the π-electron system. Because only a slight change in bond length alternation of the conjugated π-electron system occurs with application of an applied electric field and because strong electron–phonon coupling and resonance stabilization of the π-electron system act to reduce the barrier to lattice relaxation, the response times of π-electron organic materials are among the fastest observed in nature. With crystalline inorganic materials, the ionic constituents move to new locations with application of an electric field with the exact movement determined by the field strength, the charge on the ions, and the restoring force. Unequal restoring forces along the three mutually orthogonal (perpendicular) axes of the crystal lead to anisotropy in the optical properties of the material. The applied electric field will induce a change in the anisotropy (the principal refractive indices). The symmetry of the electro-optic tensor will be closely related to the symmetry of the piezoelectric tensor. The linear electro-optic effect requires that the crystal exhibit noncentrosymmetric (acentric or ferroelectric) symmetry. A centrosymmetric crystal possesses a center of symmetry defined by identical particles existing in the lattice at vectors $\mathbf{r}$ and $-\mathbf{r}$, where $\mathbf{r}$ is a position vector measured from an arbitrary origin. A centrosymmetric crystal, like an isotropic liquid or gas, can exhibit a quadratic electro-optic effect.

For the sake of completeness, it can also be noted that index of refraction changes can be induced by acoustic waves (acousto-optic modulation) and by heating (thermo-optic modulation). Also, elasto-optic and photo-elastic effects can produce index of refraction changes.

To keep this article to a manageable length, discussion is limited to solid-state electro-optic materials. Design principles being used to produce new organic electro-optic materials will be illustrated. Since inorganic materials exist as crystals, limited chemical modification of such materials is possible, although processing techniques for fabricating thin films of such crystalline materials have been developed in a number of cases. For organic macromolecular materials, development of new materials is an on-going activity. Several representative devices being fabricated, using electro-optic materials, will be discussed. Four of the most commonly encountered types of EO devices include: (i) stripline devices; (ii) prism, cascaded prism, and superprism devices; (iii) resonated devices, such as ring microresonator and photonic bandgap devices, and (iv) waveplates. Of course, devices such as stripline devices can be configured in a variety of ways to produce polarization-sensitive and polarization-insensitive electrical-to-optical signal transducers, optical switches either using simple 1×2 or 2×2 switch architectures or multimode interference (MMI) switches, optical gyroscopes, photonically controlled phased array radar systems, spectrum analyzers, optical digital signal processors, analog-to-digital (A/D) and digital-to-analog (D/A) signal converters, electromagnetic (radiofrequency to millimeter wave) signal generators, voltage sensors, etc. Prism-type devices are typically used for spatial light modulation or laser beam deflection. Ring microresonator devices afford a rich array of applications ranging from active wavelength division multiplexing (WDM) to active optical interconnect reconfiguration, to voltage-controlled wavelength tuning of laser outputs.

Basic Relationships

The effect of the applied field is to deform the index of refraction ellipsoid, which can be represented as

$$\left(\frac{1}{n^2}\right)_1 X^2 + \left(\frac{1}{n^2}\right)_2 Y^2 + \left(\frac{1}{n^2}\right)_3 Z^2 + 2\left(\frac{1}{n^2}\right)_4 YZ + 2\left(\frac{1}{n^2}\right)_5 XZ + 2\left(\frac{1}{n^2}\right)_6 XY = 1 \tag{10}$$

leading to the corrections

$$\Delta\left(\frac{1}{n^2}\right)_i = \sum_{j=1}^{3} r_{ij} E_{0j} \tag{11}$$

where the electro-optic tensor components are related to the tensor components of the second-order nonlinear optical susceptibility by $r_{ij} = -8\pi / n_{0i}^2 n_{0j}^2 \chi_{ji}^{(2)}$. The full tensorial equation for change of the 'indicatrix' (index of refraction ellipsoid) can be expressed as

$$\begin{pmatrix} \Delta\left(\frac{1}{n^2}\right)_1 \\ \Delta\left(\frac{1}{n^2}\right)_2 \\ \Delta\left(\frac{1}{n^2}\right)_3 \\ \Delta\left(\frac{1}{n^2}\right)_4 \\ \Delta\left(\frac{1}{n^2}\right)_5 \\ \Delta\left(\frac{1}{n^2}\right)_6 \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \\ r_{41} & r_{42} & r_{43} \\ r_{51} & r_{52} & r_{53} \\ r_{61} & r_{62} & r_{63} \end{pmatrix} \begin{pmatrix} E_{0X} \\ E_{0Y} \\ E_{0Z} \end{pmatrix} \tag{12}$$

The electro-optic tensor reduces considerably for specific materials reflecting the symmetry of the particular material. For inorganic crystalline materials, the electro-optic tensor can be reasonably complex (containing many nonzero elements). It is also helpful to note that for crystalline inorganic materials, the electric-field-induced change in crystal shape leads to strain and the orientation of the indicatrix is altered, leading to an additional contribution to the change in index coefficients:

$$\Delta\left(1/n^2\right)_i = \sum_k p_{ik} S_k + \sum_j r_{ij} E_j \tag{13}$$

where S_k is a component of the strain and p_{ik} is the elasto-optic tensor. At high frequencies, the inertia of the crystal prevents macroscopic straining so that the first term of the above equation vanishes. At low frequencies, elasto-optic effects cannot be ignored. Because the deformation leading to strain is generally caused by the inverse piezoelectric effect, it is linearly related to the applied electric field (the same dependence as the linear electro-optic effect). This can lead to a frequency dependence of the 'effective' electro-optic coefficient for crystalline materials.

A brief comment on the photo-elastic effect (the change in index coefficients produced directly by applied stress) is warranted. This effect has the form:

$$\Delta\left(1/n^2\right)_i = \sum_l \pi_{il}\sigma_1 \tag{14}$$

where σ_l are the components of the stress and π_{il} are the piezo-optical coefficients. Note that this effect is independent of the applied electric field.

For axially symmetric 'charge-transfer type' organic chromophores prepared by deposition or electric field poling, only two unique tensor elements, r_{33} and r_{13}, survive:

$$\tilde{r} = \begin{pmatrix} 0 & 0 & r_{13} \\ 0 & 0 & r_{13} \\ 0 & 0 & r_{33} \\ 0 & r_{13} & 0 \\ r_{13} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{15}$$

The electro-optic effect for organic macromolecular materials derives from molecules with conjugated π-electrons confined within molecules (chromophores). The electro-optic tensor elements can be related to molecular first hyperpolarizability, β, and the chromophore number density, N (molecules/cm^3), by

$$\begin{aligned} r_{33} &= -2Nf(0)\beta\langle\cos^3\theta\rangle/n_0e^4 \\ r_{13} &= -Nf(0)\beta\langle\sin^2\theta\cos\theta\rangle/n_{0o}^2n_{0e}^2 \end{aligned} \tag{16}$$

The expressions in brackets are 'order parameters' describing the degree of noncentrosymmetric order. $\langle\cos^3\theta\rangle = 1$ corresponds to all 'dipolar' (charge transfer type) molecules pointing in the same direction (perfect noncentrosymmetric or ferroelectric order) while $\langle\cos^3\theta\rangle = 0$ corresponds to complete disorder and no electro-optic effect. The reason that three angles appear in the order parameter is that an angle is required to represent the principal symmetry axis for the chromophore, the principal symmetry axis of the optical field, and the principal symmetry axis of the electric control field. Thus, averaging (denoted by brackets) must be carried out over each of these three angles. The tensor element r_{33} corresponds to the electric control field applied along the principal axis of the ordered chromophores and the principal axis of the optical field vector, transverse magnetic (TM) polarized light. The tensor element r_{13} corresponds to the control electric field applied along the principal axis of the optical field, transverse electric (TE) polarized light, and orthogonal to the principal chromophore axis. The notation r_{33} is contracted notation for r_{333}. The factor $f(0)$ takes into account the dielectric nature of the medium into which the chromophores (molecules with large first hyperpolarizability) are embedded.

For organic electro-optic materials, elasto-optic effects do not appear to make significant contributions. Moreover, electro-optic activity can be systematically improved by improving β and by optimizing the product of chromophore number density and order parameter. Due to the presence of intermolecular electrostatic interactions, W, among chromophores, order parameters and number density are not independent, e.g., for materials that are prepared by electric field poling the order parameter relevant to the principal component of the electro-optic tensor can be expressed approximately as

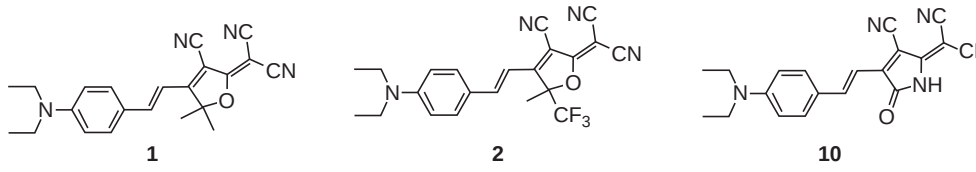
$$\langle\cos^3\theta\rangle = L_3(\mu f(0)E_p/kT)[1 - L_1(W/kT)^2] \tag{17}$$

where L_3, and L_1 are Langevin functions, μ is the chromophore dipole moment, E_p is the strength of the electric poling field, k is the Boltzmann constant, and T is the Kelvin poling temperature. W is a quadratic function of chromophore number density, N. The above equation indicates that a maximum will be observed in the graph of r_{33} versus N.

Materials

With inorganic and organic crystalline materials, very little can be done to optimize electro-optic activity other than discovering new EO crystalline materials or employing isotopic (e.g., deuterium for hydrogen) or ion substitution with existing materials.

As with organic liquid crystalline materials, the electro-optic activity of macromolecular organic second-order nonlinear materials can be systematically improved by molecular modification. As shown in **Fig. 1**, quantum mechanical calculations can provide useful guidance as to which structural modifications will lead to improvements in molecular (first) hyperpolarizability and ultimately electro-optic activity. The molecular hyperpolarizability (long wavelength limit) can be increased by a factor of two by simple variation of the acceptor structure. The calculated values of wavelengths for the interband charge transfer transitions are 390 nm, 403 nm, and 430 nm. Since dipole moments do not change significantly with structure, intermolecular electrostatic interactions should be similar for these three chromophores and the improvement in β should translate to an improvement in electro-optic activity. This theoretical prediction has been experimentally verified. This is just one example of use of quantum mechanics to guide the improvement of electro-optic activity for organic materials. This figure also illustrates the typical structure of an organic electro-optic chromophore, which consists of an electron donor region (an amine donor in the example shown), a



Dipole moment:	11.8	11.9	11.7	Debye
Hyperpolarizability	23.8	34.4	43.0	$*10^{-30}$ esu

Fig. 1 Density functional calculations (DFT) of molecular first hyperpolarizability (β) and dipole moment (μ) for three chromophore structures.

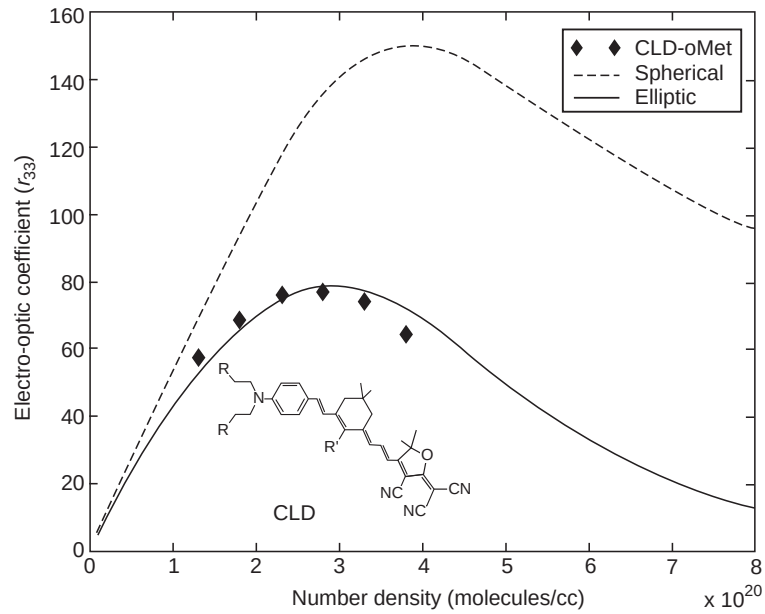


Fig. 2 Two theoretical (statistical mechanical) simulations of the variation of electro-optic activity with chromophore number density are shown and compared with experimentally observed results for the CLD chromophore (structure shown) dissolved in an amorphous polycarbonate host polymer.

connective conjugated π-electron bridge, and an electron deficient electron acceptor region. The theoretically predicted (by density functional theory, DFT, calculations) variation of molecular first hyperpolarizability with molecular structure has been verified for the structures shown. However, the critical point is that electro-optic activity of organic materials (currently exhibiting a maximum value of 182 pm/V at 1.3 microns wavelength) is likely to be improved for some time to come. This value is to be compared to a value of 32 pm/V for lithium niobate. Another factor influencing the value of macroscopic electro-optic activity realized for electrically poled organic materials is the effect of intermolecular electrostatic interactions on optimization of $N\langle\cos^3\theta\rangle$. Statistical mechanical calculations can provide guidance as to how this is achieved (see **Fig. 2** where the role of chromophore shape is illustrated). Approximating the chromophore shape by a rigid prolate ellipsoid with embedded dipole moment yields nearly quantitative reproduction of the experimental data. If the π-electron structure (dipole moment) is left unchanged but the shape of the chromophore is altered to a spherical shape, a significant improvement in electro-optic activity is predicted. The shape effect arises because there are two components to the electronic dipole–dipole intermolecular electrostatic potential function describing the many body interaction of chromophores. One component favors centrosymmetric ordering while the second component favors noncentrosymmetric ordering. Their relative weighting in defining macroscopic order depends on chromophore shape. Theoretical calculations have also shown that embedding chromophores in nanoscopic macromolecular objects (such as dendrimers and dendronized polymers) can dramatically enhance electro-optic activity. This type of engineering can also be used to control auxiliary properties such as optical loss and stability. The exceptional processability of organic electro-optic materials has permitted conformal and flexible devices to be fabricated and has permitted electro-optic circuitry to be integrated vertically on top of semiconductor VLSI (very large-scale integration) electronics. Mach–Zehnder modulators, with 3 dB operational bandwidths of 200 GHz, have been fabricated from organic electro-optic materials as have Mach–Zehnder devices with operational drive voltages of less than one volt. The major advantage that crystalline inorganic materials possess, relative to macromolecular organic EO materials, is superior thermal and photostability. A comparison of lithium niobate and the best organic EO materials, is given

Table 1 Comparison of electro-optic performance for inorganic and organic EO materials at 1.3 microns wavelength

Parameter	*$LiNbO_3$*	*NLO polymer*
Effective EO coefficient, r_{eff}, pm/V	32	182
Index of refraction (n)	2.2	1.6–1.7
Dielectric permittivity (ε)	30	3
Material Figure of Merit, $n^3(r_{eff})/\varepsilon$	10	270
Bandwidth × length product, GHz cm	7	>100
$V_\pi L$, V cm	5	1.5
Optical loss, dB/cm	0.2	0.2–1.0
Processing temperature (°C)	1000	<200
Multiple layers possible	No	Yes

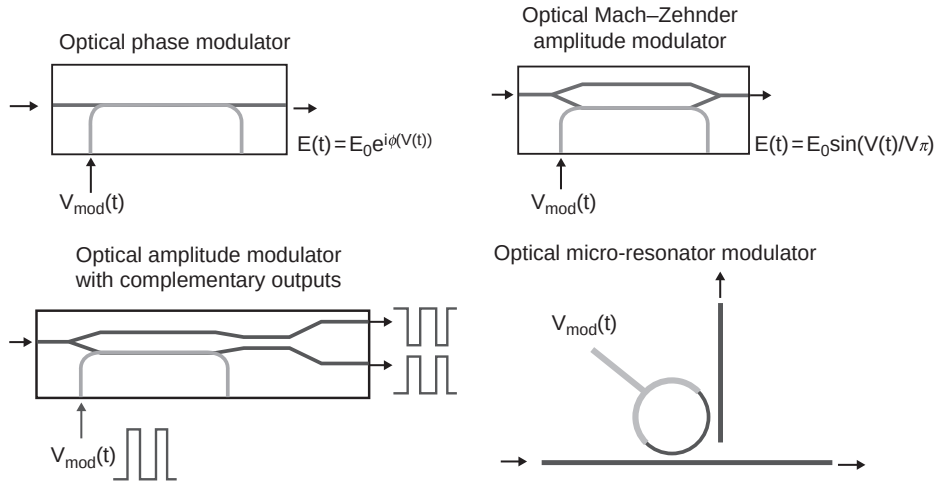


Fig. 3 Schematic representation of simple electro-optic device structures are shown: Upper left (birefringent modulator), upper right (Mach–Zehnder interferometer), lower left (1 × 2 optical switch), lower right (ring microresonator).

in **Table 1**. Cost is a significant factor if electro-optic materials are to compete with modulated lasers and other opto-electronic devices for applications in communications, computing, transportation, and defense. Currently, manufacturing costs for both crystalline and macromolecular materials are very high (several thousand dollars per modulator). Recently, it has been demonstrated that organic macromolecular materials can be used with soft lithography techniques to mass-produce electro-optic circuitry. This could conceivably dramatically reduce manufacturing costs. Also, the fact that organic macromolecular materials can be directly integrated with silicon photonics and VLSI semiconductor electronics, may influence future commercial utilization of electro-optic materials, which is currently dominated by lithium niobate.

Devices

Representative simple device structures are shown in **Fig. 3**. For a simple stripline phase (birefringence) modulator (**Fig. 3**, upper left), the phase retardation produced by application of an electric field is given by

$$\Delta\phi; = 2\pi\Delta nL/\lambda = \pi n^3 rEL/\lambda = \pi n^3 rVL/\lambda d \tag{18}$$

where L is the electrode length (the distance over which the electrical and optical fields co-propagate), λ is the optical wavelength, $V = E/d$ is the electric field felt by the material, and d is the electrode spacing. Electro-optic device performance is frequently characterized by the voltage, V_π, required to produce a phase shift of π radians. For a TM polarized light propagating through a birefringent modulator:

$$V_\pi = \lambda d/(n_{0e}^3 r_{33} L\Gamma) \tag{19}$$

where Γ characterizes the overlap of the optical and electrical waves.

A simple Mach–Zehnder interferometer (electro-to-optical signal transducer) is shown in the upper right of **Fig. 3**. When such a device is constructed from electrically poled organic electro-optic chromophores and a voltage is applied to one arm, a retardation of light propagating in that arm will be effected. This, in turn, will result in a change in the constructive/destructive interference of

the beams at the output of the device. The consequence of applying an electrical signal to one arm of a Mach–Zehnder interferometer is that the applied voltage is transduced onto the optical output beam as an amplitude modulation of that beam. The voltage required to achieve full wave modulation for an optical beam of TM polarization is

$$V_\pi = \lambda d/(n_{0e}^3 r_{33} L\Gamma) \tag{20}$$

Amplitude modulation can also be achieved using a birefringent modulator by placing polarizers at the input and output of the device. If an input polarizer is used to launch TM and TE modes, the birefringent modulator functions to produce a rotation of light, which is turned into an amplitude modulation by inserting a polarizer at the output. For such a device, V_π depends upon the difference in r_{33} and r_{13}. Note that for poled organic chromophores, $r_{33} \approx 3r_{13}$, $(V_\pi)_{\mathrm{BRM}} \approx 1.5(V_\pi)_{\mathrm{MZM}}$. For the above devices, it is clear that $V\pi$ will depend on the length over which the electrical and optical fields co-propagate in phase. For low electrical field frequencies, L will be the electrode length. The electrode spacing d, that can be used, will depend on the index of refraction difference between the core (electro-optic material) and a cladding material used to prevent the optical field experiencing the metal electrodes (an event that would dramatically increase propagation loss). For poled organic macromolecular materials and standard commercial claddings, d is typically in the order of 10 microns. Electrode lengths usually range from 1 to 3 cm. Two factors limit bandwidth in stripline devices: (i) velocity mismatch between the electrical and optical waves (which relates to the difference between n_0^2 and ε, where ε is the dielectric permittivity of the material – see **Table 1**); and (ii) microwave and millimeter wave loss in the metal electrodes. For organic electro-optic materials, the latter defines operational 3 dB bandwidth of devices.

Optical routing switches (see **Fig. 3**, lower left) are more complicated to understand and typically require somewhat larger V_π voltages to effect a switching operation (e.g., $(V_\pi)_{\mathrm{simple2\times 2crossbar}} \approx 1.7(V_\pi)_{\mathrm{MZM}}$).

Two conditions must be satisfied for light to be coupled into a ring microresonator device (see **Fig. 3**, lower right): (i) the circumference of the ring must be a multiple of the wavelength of light, and (ii) the velocity of light in the ring must equal the velocity of light at the input. If a ring microresonator is fabricated from an electro-optic material, then an applied electrical voltage can be used to control coupling of light into and out of the ring microresonator. The same electrode structure, that is used to achieve voltage-controlled wavelength selective filtering, can also be used to transduce electrical information onto the optical beam while it is in the ring resonator. The quality (Q) factor of a ring microresonator defines both the wavelength selectivity and the bandwidth of the device. Ring microresonator devices also have the advantage of permitting a long interaction length to be achieved in a very compact device structure. The dimensions (radii) of ring microresonator structures are defined by bending loss and thus by the difference between core and cladding materials. Typical radii can range from hundreds of nanometers to hundreds of microns. With ring microresonators, a reduction in drive voltage requirements can be achieved by trading off bandwidth. Another price paid in the use of ring microresonator device structures is that of bending loss, which is not present in the other device structure shown. Ring microresonators can be used for active wavelength division multiplexing (WDM), voltage-controlled optical signal routing, and wavelength tuning of optical sources.

The interaction length of a prism optical beam deflector can be increased by cascading prisms together. For such a device structure, the angle of deflection becomes

$$\theta = n_0^3 r_{33}(V/d)(L/h) \tag{21}$$

where h is the height of an individual prism and L is the length of the prism array. For currently available electro-optic materials only small deflection angles (a few degrees or less) can be achieved with practical voltage levels. Superprism structures may permit large deflection angles to be achieved but such devices have yet to be demonstrated in a convincing manner.

Further Reading

Butcher, P.N., Cotter, D., 1990. The Elements of Nonlinear Optics. New York: Cambridge University Press.

Dagli, N., 1999. Wide-bandwidth lasers and modulators for RF photonics. IEEE Trans. Microwave Theor. Techniques 47, 1151–1171.

Dalton, L.R., 2003. Rational design of organic electro-optic materials. Journal of Physical Condensed Materials 15, R897–R934.

Dalton, L., Harper, A., Ren, A., *et al.*, 1999. Polymeric electro-optic modulators: from chromophore design to integration with semiconductor very large scale integration electronics and silica fiber optics. Ind. Eng. Chem. Res. 38, 8–33.

Hornak, L.A., 1992. Polymers for Lightwave and Integrated Optics. New York: Marcel Dekker.

Khazaei, H.R., Wang, W.J., Berolo, E., Ghannouchi, F., 1998. High speed GaAs/AlGaAs traveling wave electro-optic modulators. Proceedings of SPIE 3278, 94–100.

Lawetz, C., Cartledge, J.C., Rolland, C., Yu, J., 1997. Modulation characteristics of semiconductor mach-zehnder optical modulators. IEEE Journal of Lightwave Technology 15, 697–702.

Lee, K.S., 2002. Advances in Polymer Science, vol. 158. Heidelberg: Springer.

Lee, M., Katz, H.E., Erben, C., *et al.*, 2002. Broadband modulation of light using an electro-optic polymer. Science 298, 1401–1403.

Nalwa, H.S., Miyata, S., 1997. Nonlinear Optics of Organic Molecules and Polymers. Boca Raton, FL: CRC Press.

Noguchi, K., Mitomi, O., Miyazawa, H., 1998. Millimeter-wave Ti:$LiNbO_3$ optical modulators. IEEE Journal of Lightwave Technology 16, 615–619.

Shen, Y.R., 1984. The Principles of Nonlinear Optics. New York: John Wiley & Sons.

Shi, Y., Zhang, C., Zhang, H., *et al.*, 2000. Low (sub-1 volt) halfwave voltage polymeric electro-optic modulators achieved by controlling chromophore shape. Science 288, 119–122.

Wise, D.L., Wnek, G.E., Trantolo, D.J., Cooper, T.M., Gresser, J.D., 1998. Electrical and Optical Polymer Systems. Fundamentals, Methods and Applications. New York: Marcel Dekker.

Yariv, A., Yeh, P., 1984. Optical Waves in Crystals. New York: John Wiley & Sons.

Polarization Introduction

JM Bennett, Michelson Laboratory, China Lake, CA, USA

Introduction

Polarization is one property of light waves and will be defined after a brief overview of the properties of light. Light is part of the electromagnetic spectrum of waves that have both particle-like and wave-like properties. Light waves carry energy in the form of photons which act like particles; the photon energy increases in proportion to the frequency of the wave. The particle-like properties of light and other electromagnetic waves are described by quantum mechanics. Light also acts like transverse waves that travel in straight lines in air and vacuum and can be described by classical electromagnetic theory. Polarization deals with the wave-like properties of light, and is described mathematically by Maxwell's equations, the key relations in electromagnetic theory. Polarization effects occur when light interacts with matter. In order to understand polarization, there will be a brief introduction to the subject of electromagnetic waves.

For a more in-depth treatment of the material in this article, the reader is directed to the Further Reading list at the end of this article and in particular to the two chapters on 'Polarization' and 'Polarizers' written by the author in Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chapter 5 and vol. II, chapter 3. New York: McGraw-Hill Inc.

Electromagnetic waves have both electric and magnetic fields associated with them. These are vibrations in directions perpendicular to the direction the wave is traveling, i.e., the direction of propagation. $\vec{E}$ represents the vector of the electric field and $\vec{H}$ perpendicular to $\vec{E}$ represents the magnetic field vector. Both these vectors are complex and have real and imaginary parts. Polarization effects are always associated with the $\vec{E}$ vector. Specifically, the plane of polarization is defined as the plane in which the $\vec{E}$ vector is vibrating. Waves having different amplitudes, phases, or angular orientations (azimuths) of their electric or magnetic vectors can be combined by conventional vector addition methods. Also the $\vec{E}$ vector of a particular vibration can be resolved into two components in mutually perpendicular directions that are vibrating in phase.

If a light source such as the sun, a candle flame, or an electric light bulb is considered on a microscopic scale, each vibrating atom or molecule emits linearly polarized light (see the definition below). But the individual atoms or molecules do not act together, so their vibrations have no fixed phase relationships to each other and they cannot be added into a single linearly polarized beam. Thus, we call light from these sources unpolarized. In an unpolarized light beam, the $\vec{E}$ vector vibrates in all directions perpendicular to the direction of propagation. If a snapshot is taken at a particular instant of time, different parts of the beam will have $\vec{E}$ vectors vibrating with different amplitudes and phases at different angles to each other, but all in a plane perpendicular to the direction of propagation. In the most common convention used in optics, the wave travels in the $+z$ direction in a right hand coordinate system and the $\vec{E}$ vectors are all vibrating at various angles in the x–y plane. The angle of vibration is measured from the positive x axis in a counterclockwise direction when the observer is looking against the direction of propagation of the light beam.

For linearly polarized or plane polarized light, if a snapshot of a light beam is taken at a particular instant, the $\vec{E}$ vector will be vibrating at a certain angle in the x–y plane. As time (or position on the traveling wave) varies, the amplitude of the $\vec{E}$ vector will vary in a sinusoidal manner, but the vibration will remain at the same angle in the x–y plane. As an example, the real part of the electric vector of a linearly polarized beam that is vibrating in the $+x$ and $-x$ directions and traveling in the $+z$ direction is given by the relation:

$$\vec{E}_x(z,t) = \hat{i}E_0 \cos(kz - \omega t) \tag{1}$$

where t is the time, $\hat{i}$ is a unit vector in the $+x$ direction, E_0 is the amplitude of the vibration, $\vec{k}$ is the propagation vector in the direction the wave is traveling, i.e., the z direction (its magnitude is $2\pi/\lambda$), and $\omega = 2\pi f$ where f is the vibration frequency of the wave. In free space, $\omega = 2\pi c/\lambda$ where c is the velocity of light and λ is the wavelength of the vibration.

Light can also be circularly or elliptically polarized. Circularly polarized light is produced by adding the electric vectors of two waves, each having the same amplitude but which are 90° out of phase; the resulting vibration sweeps out a circle in the x–y plane. When the first wave is vibrating in the $+x$ and $-x$ directions (**Eq. (1)**), the second wave is vibrating in the $+y$ and $-y$ directions:

$$\vec{E}_y(z,t) = \hat{j}E_0 \sin(kz - \omega t) \tag{2}$$

and is added to the first wave. The resultant wave is the sum of **Eqs. (1)** and **(2)**:

$$\vec{E} = E_0[\hat{i}\cos(\mathrm{kz} - \omega t) + \hat{j}\sin(kz - \omega t)] \tag{3}$$

The direction of vibration of this wave will rotate in a circle as either the time increases or as the distance along the wave increases, but the amplitude of the vibration will not change. As time increases, the vibration will make a circle in the clockwise direction (to negative angles). This light is defined to be right circularly polarized. If the + sign between the two parts of **Eq. (3)** is changed to a minus sign, as time increases, the vibration will make a circle in the counterclockwise direction (to positive angles) and the light is now defined to be left circularly polarized. These definitions are for conventional (traditional) optics. However, the opposite

 doi:10.1016/B978-0-12-809283-5.00691-1

definitions with right (left) circularly polarized light defining circles in the counterclockwise (clockwise) direction are also in use. In modern physics, there is still another convention that defines right (left) circularly polarized light as having negative (positive) helicity.

Right and left elliptically polarized light beams have the same angle conventions as for circularly polarized beams but are produced by adding two electric vectors that have different amplitudes. In **Eqs. (1)** and **(2)**, the amplitude terms will be, for example, E_1 and E_2 instead of E_0 and **Eq. (3)** can no longer be used. There is now a major axis and a minor axis of the ellipse. If $E_1 > E_2$, the major axis will be along the x axis; for $E_1 < E_2$, the major axis will coincide with the y axis.

The preceding discussion has dealt entirely with the electric vector of the electromagnetic field. However, one cannot observe the $\vec{E}$ vector. The irradiance (energy per unit area per unit time), $\vec{E}\vec{E}^*$ is what can be observed visually and measured by electronic detectors. Thus, measurements of the polarization are irradiance measurements. Light transmitted by a polarizer is called the transmittance of the polarizer; similarly, light reflected from a polarizer is called the reflectance of the polarizer.

One of the most important parts of the subject of polarization is how to produce linearly polarized light starting with unpolarized light. This is done by using polarizers, as discussed in the section on polarizers below. Certain materials have special properties that make them able to polarize light. Depending on the application, different kinds of polarizers are preferred. Optics textbooks by Hecht and Guenther discuss the most important polarizers and two articles by Bennett describe many kinds of polarizers including special ones (see the Further Reading at the end of this article for full references). Only the basic principles will be described here.

Sunlight scattered by air molecules in the atmosphere (Rayleigh scattering) is also partially linearly polarized. The air molecules act like tiny dipoles and vibrate when they absorb sunlight. They emit radiation that is polarized in certain directions relative to the vibration direction. When the viewer is at a 90° angle with respect to the sun, the polarization of the skylight is a maximum. Rayleigh scattering is the subject of several books and will not be further discussed here.

Retarders are used to change linearly polarized light into circularly or elliptically polarized light and compensators, which are a form of retarders, can analyze an unknown type of polarized light and determine its composition. They are discussed below.

Polarimetry and ellipsometry are closely related techniques that are used to determine the optical properties of a material by shining a known type of polarized light on the material at non-normal incidence and analyzing the polarization properties of the light after it has been reflected from the material. These subjects are treated in other articles in this encyclopedia and in several references at the end of this article.

Changes can be produced in the optical properties of some materials by applying an electric field, a magnetic field, an acoustic field, or another form of variable pressure. The materials changed in these ways are said to be electro-optic, magneto-optic, or piezo-optic. The changes in the optical properties modify some parameter of a light wave passing through a material or reflecting from it. The amplitude, phase, frequency, polarization, or direction of the light wave can be modified. In this article, we are only concerned with modifications that produce changes in the polarization; these are discussed towards the end of this article.

Matrix methods have been developed to handle problems involving polarization and there is also a visual representation of the matrix algebra that is based on the Poincaré sphere. Both topics are covered at the end of this article.

Polarizers

Basic Relations for Polarizers

A linear polarizer is anything which, when placed in an incident unpolarized beam, produces a beam of light whose electric vector is vibrating primarily in one direction with only a small component vibrating in the direction perpendicular to it. The transmittance T of the linear polarizer is

$$T = \frac{1}{2}(T_1 + T_2) \tag{4}$$

where T_1, the principal transmittance of the polarizer, is $\gg T_2$, the transmittance of the polarizer at 90° to the principal transmittance. Thus, a perfect polarizer would transmit only 50% of an incident unpolarized beam.

If a linear polarizer is placed in a linearly polarized beam and is rotated about an axis parallel to the beam direction, the transmittance will vary between a maximum value T_1 and a minimum value T_2 according to the law:

$$T = (T_1 - T_2)\cos^2\theta + T_2 \tag{5}$$

where θ is the angle between the plane of vibration of the principal transmittance and the plane of vibration of the electric vector in the incident beam.

The extinction ratio ρP of a polarizer is defined as

$$\rho_P = \frac{T_2}{T_1} \tag{6}$$

and the degree of polarization of a polarization P is

$$P = \frac{T_1 - T_2}{T_1 + T_2} \tag{7}$$

When two identical polarizers are placed in an unpolarized beam, and the directions of their principal transmittances, T_A and T_B, are inclined at an angle θ to each other, the transmittance of the pair will be

$$T_\theta = \frac{1}{2}\left(T_1^2 + T_2^2\right)\cos^2\theta + T_1 T_2 \sin^2\theta \tag{8}$$

Thus, when the directions of the principal transmittances are aligned, $T_\parallel = \frac{1}{2}\left(T_1^2 + T_2^2\right)$, and when they are perpendicular, $T_\perp = T_1 T_2$.

Birefringent Materials (Calcite)

The majority of high-quality polarizers are made from calcite. This is a birefringent (doubly refracting) crystalline material that is uniaxial (i.e., there is one preferred direction in the crystal). A birefringent material acts differently for light going in different directions through the crystal. For example, if an unpolarized light beam passes through the crystal in a certain direction, it will be split into two spatially separated beams that are parallel but are linearly polarized at right angles to each other. A uniaxial crystal has an optic axis (i.e., a certain direction in the crystal). When light rays travel parallel to the optic axis, they travel at the same velocity and there is no difference between them. When light passes through the crystal in other directions, the ray whose vibration direction is perpendicular to the optic axis is governed by the ordinary laws of geometrical optics (the same as for isotropic materials) and is called the ordinary ray. The ray whose vibration direction is parallel to the optic axis does not follow the normal geometrical optics laws and is called the extraordinary ray. One ray travels faster than the other, so there is a phase retardation for one ray relative to the other. This is the principle of a retarder or retardation plate (see the next section).

Calcite is a negative uniaxial crystal which means that the refractive index for the ordinary ray is larger than the refractive index for the extraordinary ray. When the ordinary ray enters a block of calcite from air at non-normal incidence, it is bent more than the extraordinary ray.

Calcite can be easily cleaved along three distinct planes, making it possible to produce rhombs of the form shown in **Fig. 1**. The optic axis, going in the HI direction, makes equal angles with all three faces at point *H*. Any plane, such as DBFH, which contains the optic axis and is perpendicular to the two opposite faces of the rhomb ABCD and EFGH is called a principal section. If the plane of incidence of light on the rhomb coincides with a principal section, the light entering the crystal will be split into two components polarized at right angles to each other which travel in slightly different directions and leave the crystal as two beams slightly displaced but parallel to each other.

The large birefringence of calcite and its excellent transmission through the visible spectral region and into the ultraviolet and infrared regions has made it possible to make excellent high extinction ratio polarizers with different designs. Some of these are shown in **Fig. 2**. There are two main types: Glan types with rectangular shapes and Nicol types with rhombohedral shapes. The polarizers are made of two pieces of calcite cemented together. Glan types have their optic axes in the plane of the entrance face. In the Nicol types, the principal section is perpendicular to the entrance face, but the optic axis is neither parallel nor perpendicular to the face. The two halves of conventional polarizing prisms are cemented together with cement that has a refractive index intermediate between the ordinary and extraordinary refractive indices. This enables one ray (generally the extraordinary, or *e* ray) to be transmitted and the other to be reflected at the cut, so that only one ray exits from the prism in the direction of the

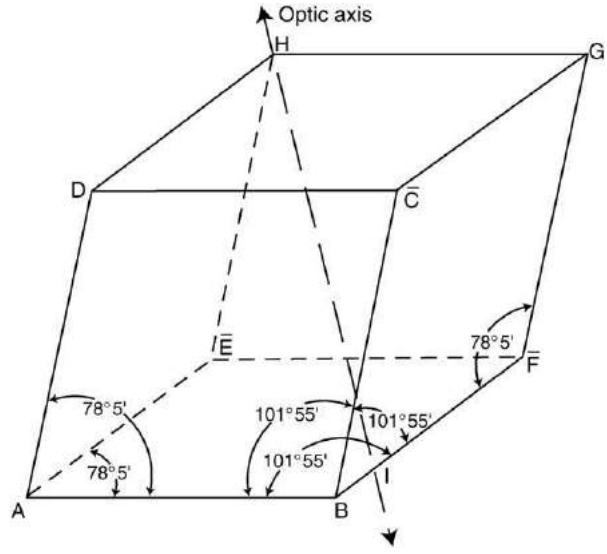


Fig. 1 Schematic representation of a rhombohedral calcite crystal showing the angles between faces. The optic axis passes through corner H and point I on side BF. Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

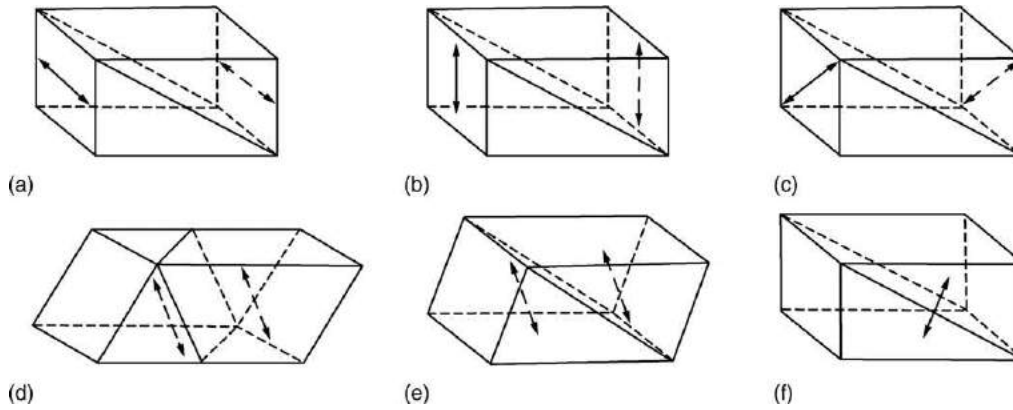


Fig. 2 Types of polarizing prisms. Glan types: (a) Glan-Thompson, (b) Lippich, and (c) Frank-Ritter; Nicol types: (d) conventional Nicol, (e) Nicol, Halle form, and (f) Hartnack-Prazmowsky. The optic axes are indicated by the double-pointed arrows. Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

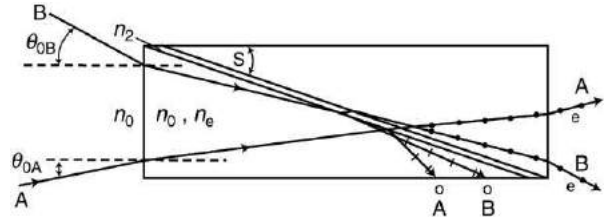


Fig. 3 Extreme rays (A and B) that can pass through a cemented Glan-Thompson prism. Both rays change to ordinary and extraordinary rays in the calcite. Both ordinary rays are reflected at the boundary between the two halves of the prism and the extraordinary rays exit in the directions indicated. Rays entering the prism between rays A and B would be transmitted by the prism. The field angle of the prism is twice the smaller of the two angles of incidence (ray A).

incoming ray. The Glan types are also used without cement with only an air space between the two halves. They can be used at shorter wavelengths in the ultraviolet where the cement absorbs. The extinction ratio can be very high for air-spaced prisms. For example, for a Glan–Foucault prism (an air-spaced Glan–Thompson prism, **Fig. 2**), the extinction ratio can be better than 1×10^{-5} and the prism is usable from about 0.214 μm in the ultraviolet to 2.3 μm in the infrared. However, air-spaced polarizers have very small field angles, so they are mainly used with laser sources where the beam is parallel.

The extreme paths of light through a cemented Glan–Thompson prism are shown in **Fig. 3**. This prism has a length that is three times the width of the entrance aperture (i.e., an L/A ratio of 3). The optic axis is perpendicular to the plane of incidence which is in the plane of the paper. In the first half of the polarizer, the paths of the ordinary and extraordinary rays, both of which follow the conventional law of refraction (Snell's Law, **Eq. (8)** above) nearly coincide. Ray A is incident on the entrance face of the polarizer at an angle such that the angle of incidence on the cut is the smallest angle for which the *o* ray is totally internally reflected. Ray B is incident at an angle such that the angle of refraction in the first half of the prism is essentially equal to the cut angle, S, so that the *e* ray just passes through the cut. The field angle of the polarizer is twice the smaller of angles θ_A and θ_B. Thus, all rays having angles of incidence between rays A and B will be polarized when exiting the prism. The paths of similar rays can be traced through the other prism designs.

In addition to conventional polarizing prisms, there are also polarizing beamsplitter prisms and Feussner prisms. In polarizing beamsplitter prisms, the two beams that are polarized at right angles to each other both emerge from the prism but are separated spatially. **Fig. 4** shows a diagram of several types of polarizing beamsplitter prisms and **Fig. 5** shows the paths of rays through side views of these prisms.

A Feussner prism is made of isotropic material but the film separating the two halves is birefringent. These prisms have the advantage that much less birefringent material is required but they have a more limited wavelength range when calcite or sodium nitrite (another birefringent material) is used because the transmitted ordinary ray has a shorter transmission range than the extraordinary ray which is transmitted by the conventional and air-spaced polarizing prisms.

Dichroic Absorbers

A dichroic material is one that absorbs light polarized in one direction more strongly than light polarized at right angles to that direction. Dichroic materials are different from birefringent materials because the latter usually have negligible absorption coefficients for both the ordinary and extraordinary rays. Stretched polyvinyl alcohol sheets treated with absorbing dyes or

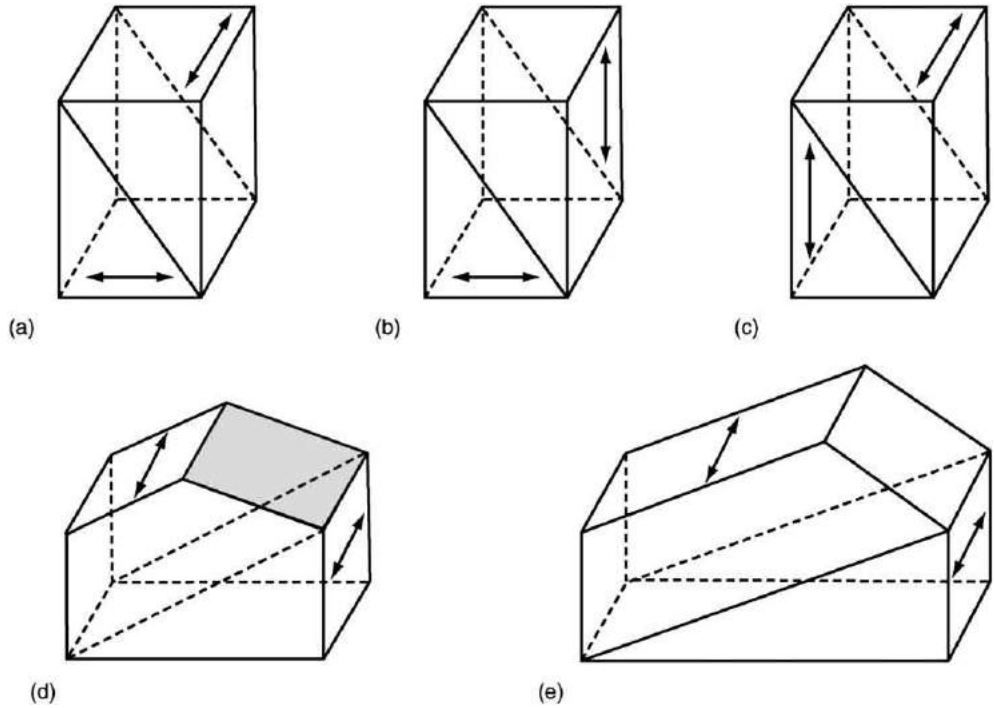


Fig. 4 Three-dimensional views of various types of polarizing beamsplitter prisms: (a) Rochon; (b) Sénarmont; (c) Wollaston; (d) Foster (shaded face is silvered); and (e) beamsplitting Glan-Thompson. Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

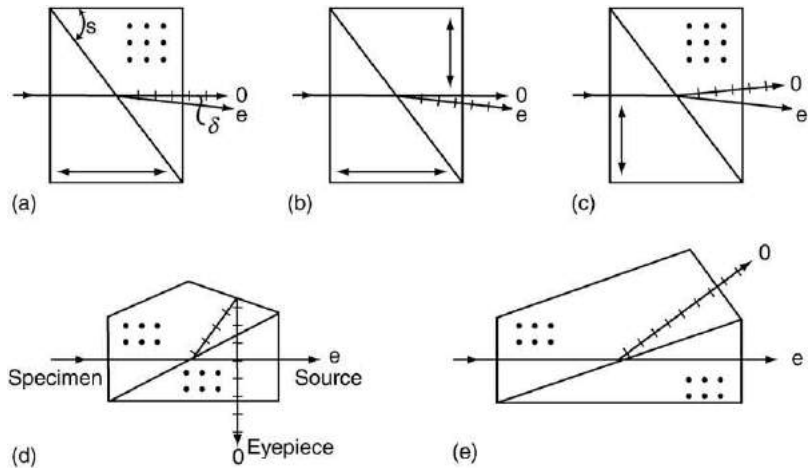


Fig. 5 Side views of the polarizing beamsplitter prisms in **Fig. 4**. The directions of the optic axes are indicated by the dots and the heavy double-pointed arrows. When the Foster prism is used as a microscope illuminator, the source, specimen, and eyepiece are in the positions indicated. Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

polymeric iodine are the most common type of dichroic absorbers and are primarily sold under the tradename Polaroid. These polarizers do not have as good an extinction ratio as the calcite prism polarizers (see above), but they are inexpensive, come in large sizes, are easily rotated, and produce negligible beam deviation. Also, they are thin, lightweight, and can be made in any desired shape. One of their main advantages is that they are insensitive to the degree of collimation of the beam and can be used in strongly convergent or divergent light. Depending on the density and type of absorbing dye used to make the polarizer, the transmission can be maximized for the visible or near infrared spectrum. The extinction ratio of the Polaroid HN-22 sheet compares favorably with that of the Glan–Thompson prisms throughout the visible spectral region, but the transmission of the Glan–Thompson prism is superior. As the dichroic polarizer transmission increases, the extinction ratio becomes worse.

Transmission and extinction ratio curves for various types of dichroic polarizers are shown in Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

A wire grid polarizer is another kind of dichroic absorber. It is made of a series of equally spaced conducting bars or wires that are either free standing or deposited onto a transparent substrate (backing plate). Energy that is polarized parallel to the length of the bars is absorbed out of the incoming wave by inducing oscillations in the electrons in the metal. Thus, only light polarized perpendicular to the bars will be transmitted. In order to have an appreciable degree of polarization, the wavelength must be at least twice the spacing between grids in the polarizer. Because of the difficulty of making wire grids with small enough spacings, these polarizers are limited to the mid- and long-wavelength infrared spectral region, beyond about 5 μm. The polarizer will have the best extinction ratio if the substrate has a low refractive index; if a high refractive index substrate such as silicon or germanium is used, it must be covered with an antireflection coating before the grid is deposited.

Most of the wire grid polarizers have been used with microwaves, so the theory is in the form for transmission lines. Similar transmission line theory has been applied to infrared polarizers in the form of bars and wires.

Reflection and Transmission

Light can be polarized by reflecting it from the flat surface of a material inclined at non-normal incidence to the light beam or by transmitting it through a transparent plate at non-normal incidence. These polarizers work because light has different reflectances when the electric vector is linearly polarized parallel and perpendicular to the plane of incidence. The plane of incidence contains both the incoming light beam (incident beam) and the reflected light beam. The angles for both the incident and reflected light beams are measured relative to an axis perpendicular to the surface (the surface normal). The reflection coefficients are given by the Fresnel equations and, for nonabsorbing materials, are:

$$R_s = r_s^2 = \frac{\vec{E}_s^{\,2}(\text{reflected})}{\vec{E}_s^{\,2}(\text{incident})} = \frac{\sin^2(\theta_0 - \theta_1)}{\sin^2(\theta_0 + \theta_1)} \tag{9}$$

$$R_p = r_p^2 = \frac{\vec{E}_p^{\,2}(\text{reflected})}{\vec{E}_p^{\,2}(\text{incident})} = \frac{\tan^2(\theta_0 - \theta_1)}{\tan^2(\theta_0 + \theta_1)} \tag{10}$$

where R_s and R_p are called the reflectances (previously called the intensity reflection coefficients), and r_s and r_p are the amplitude reflection coefficients. $\vec{E}_s$ ($\vec{E}_p$) is the incident or reflected electromagnetic wave vibrating perpendicular (parallel) to the plane of incidence and θ_0 and θ_1 are the angles of incidence and refraction, respectively. The angle of refraction is the angle of the light beam in the material, measured from the surface normal. For a nonabsorbing material, it can be obtained in terms of the refractive index n_1 of the material from Snell's Law:

$$\frac{\sin\theta_0}{\sin\theta_1} = \frac{n_1}{n_0} \tag{11}$$

where n_0 and n_1 are the refractive indexes of the incident medium (usually air with $n_0 = 1$) and the material. Sometimes the refractive index of the material is expressed as a ratio measured relative to the refractive index of the incident medium: $n = n_1/n_0$. The reflectances of absorbing materials are similar to **Eqs. (9)–(11)** but involve complex refractive indexes and angles of refraction.

Fig. 6 shows curves for R_s and R_p as a function of angle of incidence for four nonabsorbing materials that have different refractive indexes. In all cases the reflectance is higher for the *s* component than for the *p* component except at 0° and 90° angles of incidence where they are the same. At the so-called Brewster angle, θ_B, $R_p = 0$, $\tan\theta_B = n_1/n_0$, and incident unpolarized light is now completely linearly polarized perpendicular to the plane of incidence (s-polarized light). Note that high refractive index materials produce more intense polarized beams (i.e., have higher reflectances) than low refractive index materials.

Fig. 7 shows the reflectances and extinction ratios for materials having different refractive indexes. Each curve is for a single reflection. Since plates have two surfaces, there will be two reflections from an actual reflection polarizer so the reflectance and extinction ratios will increase. If the plates are thick, the reflection from the back surface of the plate will be displaced from the reflection from the front surface of the plate. A reflection polarizer is normally used close to the Brewster angle because a high degree of polarization (i.e., a very small extinction ratio) is desired. Because the extinction ratio changes rapidly for angles of incidence close to the Brewster angle, the incident beam must be well collimated to obtain a high degree of polarization. A main disadvantage of reflection polarizers is that the reflected beam is no longer parallel to the incident beam and two additional reflections are required to align the beam.

Light transmitted through a plate will only be partially linearly polarized at any angle of incidence including the Brewster angle because both s- and p-components are partially transmitted. Transmission polarizers are thus made of several plates to increase the degree of polarization. **Fig. 8** shows the transmittance and extinction ratio for a stack of four plane parallel plates assuming multiple incoherent reflections within each plate and no reflections between plates. At the Brewster angle the p-transmittance is theoretically 1 (assuming that there is no absorption within the material) but the extinction ratio greatly depends on the refractive index of the material. A stack of low refractive index plates ($n = 1.5$) has a poor extinction ratio, while a pile of high refractive index plates ($n = 4.0$) has a much better extinction ratio. The plates are often inclined at small angles to each other so the light multiply reflected between plates (which decreases the polarization) is removed from the transmitted beam. The sides of each plate are plane parallel to increase the polarization, but the plates are too thick for the amplitudes of the multiple internally reflected beams

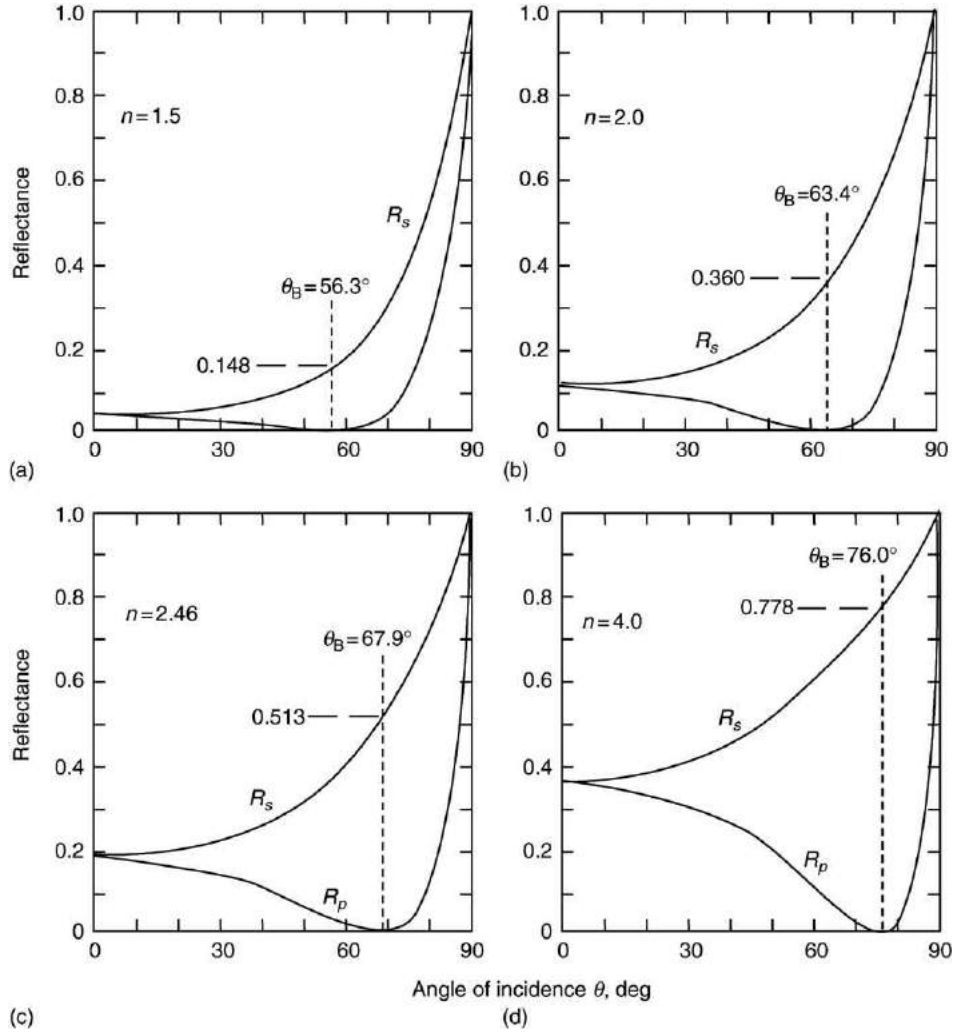


Fig. 6 Reflectance of light polarized parallel R_p and perpendicular R_s to the plane of incidence from materials of different refractive index n as a function of angle of incidence: (a) $n=1.5$ (alkali halides in the ultraviolet, glass (approximately) in the visible, and sheet plastics in the infrared), (b) $n=2.0$ (AgCl in the infrared), (c) $n=2.46$ (Se in the infrared), and (d) $n=4.0$ (Ge in the infrared). The Brewster angle θ_B (at which R_p goes to 0) and the magnitude of R_s at θ_B are also indicated. Reproduced with permission from Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.

to add or subtract. If the amplitudes of the beams could be added, there would be interference effects and the transmittance would vary with the thickness of each plate and with the wavelength. These are so-called interference polarizers and are mentioned below.

Transmission polarizers do not have the angle deviation problem of the reflection polarizers, but the transmitted beam may be slightly displaced parallel to the incident beam if the plates are thick. The main problem with transmission polarizers is that the light is not completely polarized even with several plates in a stack. If the plates are slightly absorbing, the transmission of the polarizer is reduced.

Miscellaneous Types

There are numerous other types of polarizers that mostly use thin films. Interference effects in thin films can increase the polarization efficiency at certain wavelengths depending on the thicknesses of the films and the design of the multilayer coating. Many of the applications involve non-normal incidence designs. For example, a single high refractive index film or a multilayer coating evaporated onto a low refractive index substrate increases the degree of polarization of reflection and transmission polarizers. Polarizing beamsplitters also use multilayer dielectric coatings. Free-standing films of different thicknesses were formerly used for infrared polarizers before wire grid polarizers became available.

Dichroic absorbers have been made from two-phase lamelar eutectics of thin needles of a conducting material embedded in a transparent matrix. Thin sheets of pyrolytic graphite material also act as dichroic absorbers. There are also numerous miniature polarizer designs for fiber optics and nanotechnology applications.

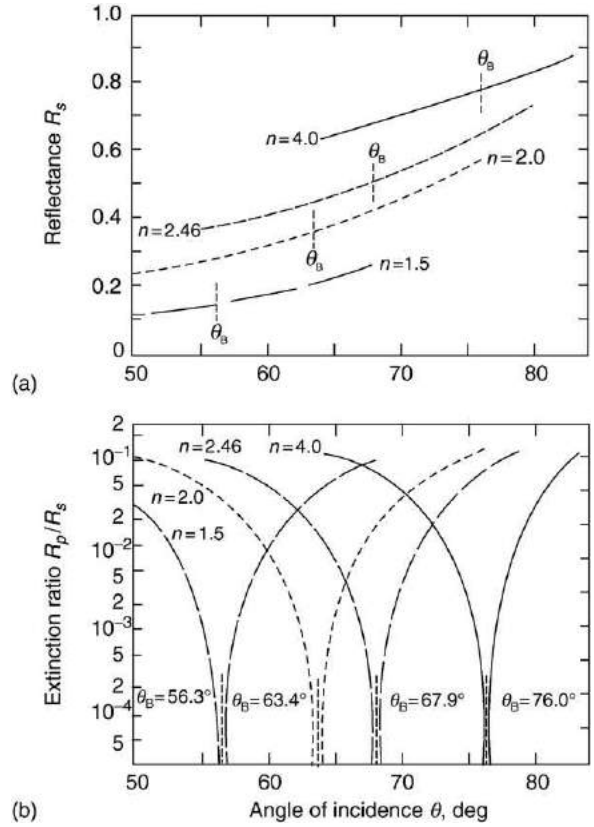


Fig. 7 (a) Reflectance R_s and (b) extinction ratio R_p/R_s for materials of different refractive index at angles near the Brewster angle θ_B. A single surface of the material is assumed. Reproduced with permission from Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.

Retarders or Retardation Plates

Retarders, or retardation plates, are devices that can change linearly polarized light into circularly or elliptically polarized light. They can also rotate the plane of polarization of linearly polarized light. A retardation plate is made from a uniaxial crystal that has an optic axis (as discussed above); the light travels in a direction perpendicular to the optic axis, as described below and shown in Fig. 9. The ordinary and extraordinary rays travel at different velocities through the crystal depending on their refractive indexes: $n=c/v$, where c is the velocity of light in a vacuum and v is the velocity of light in the material. The larger the refractive index, the slower is the velocity of light in the material. The ordinary ray with a refractive index n_o and the extraordinary ray with a refractive index n_e have mutually perpendicular vibration directions. If the ray has a vibration direction at another azimuth in the crystal, the refractive index is intermediate between n_o and n_e. For a positive uniaxial crystal $n_e > n_o$ so the extraordinary ray travels slower than the ordinary ray through the crystal. Thus, the designation 'fast axis' is often used for the ordinary ray and 'slow axis' is used for the extraordinary ray, as shown in Fig. 9. In a negative uniaxial crystal, $n_e < n_o$, i.e., the velocities along the two axes are reversed.

Because there is a velocity difference between the ordinary and extraordinary rays traveling through a uniaxial crystal, they get out of phase. Depending on their velocities and the optical thickness of the material (the refractive index times the physical thickness), the addition of their amplitudes results in a wave whose vibration direction is either: (a) perpendicular to the original vibration direction; (b) rotates in a circle (right or left circularly polarized light); or (c) rotates in an ellipse (right or left elliptically polarized light). The path difference N between two rays in the crystal, measured in terms of the wavelength of the light, is $N\lambda = \pm d(n_e - n_o)$, where d is the physical thickness of the material. The corresponding phase difference δ between the two rays is $\delta = 2\pi N = \pm(2\pi d/\lambda)(n_e - n_o)$. The quantity N is important because if beams of light vibrating along the fast and slow axes of a crystal get out of step by one quarter of the wavelength of the light, the device is called a quarter wave (or $\lambda/4$) retardation plate, or simply a quarter-wave plate. There are also half-wave ($\lambda/2$) plates that rotate the plane of polarization, and full-wave (λ) plates that rotate the plane of polarization through 180°. If light passes through a full-wave plate, theoretically it is indistinguishable from the original beam. However, materials are normally temperature sensitive so that as the temperature changes, the retardation is no longer exactly one wave, but may be slightly less than or greater than one wavelength. Other thicknesses of retardation plates produce elliptically polarized light.

Fig. 10 shows what happens to a beam of linearly polarized light when the axis of vibration is at 45° to the fast and slow axes of a positive uniaxial crystal. The fast axis is in the horizontal direction, the same as in Fig. 9.

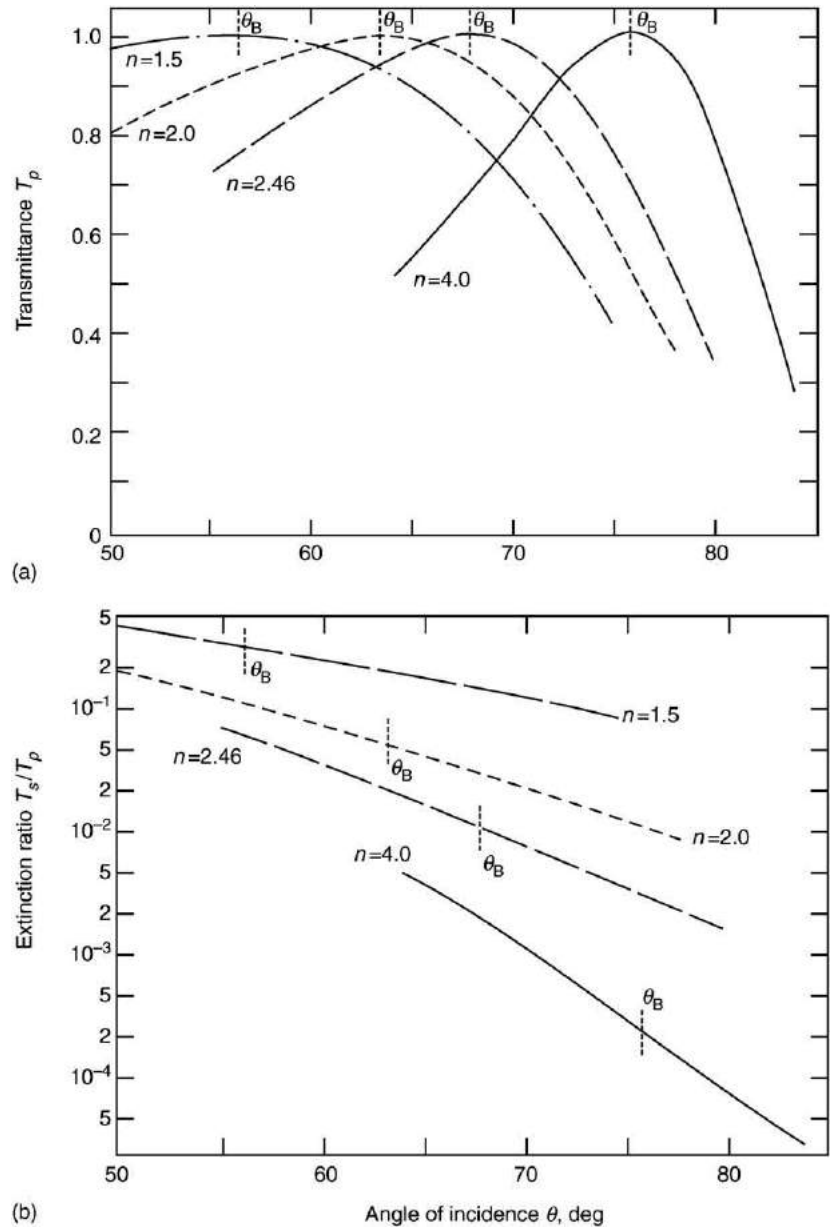


Fig. 8 (a) Transmittance and (b) extinction ratio of four plane-parallel plates of refractive index n as a function of angle of incidence, for angles near the Brewster angle. Assumptions are multiple reflections but no interference within each plate and no reflections between plates. Reproduced with permission from Bennett JM (1995) Polarization. In Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.

The main purpose of a quarter-wave plate is to change linearly polarized light into circularly polarized light. However, as **Fig. 11** shows, the state of polarization of the light will be quite different depending on the orientation of the plane of polarization of the incident beam relative to the fast axis of the quarter-wave plate.

A half-wave plate is used to rotate the plane of polarization of a linearly polarized beam. The plane of polarization is always rotated through an angle that is twice the angle the initial plane of polarization makes with the fast axis of the uniaxial crystal. A linearly polarized beam always remains linearly polarized.

Retardation plates are often made of mica, stretched polyvinyl alcohol, or crystal quartz, although they can also be made of other stretched plastics, sapphire, magnesium fluoride, and other materials.

Variable Retardation Plates and Compensators

Variable retardation plates are devices whose retardation can be varied in a variety of ways. They can be used to modulate or vary the phase of a beam of linearly polarized light or to analyze a beam of unknown polarization (often elliptically polarized light)

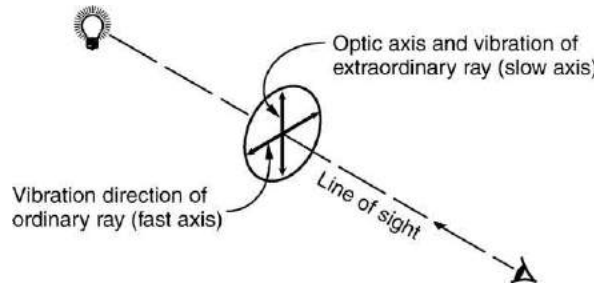


Fig. 9 Light incident normally on the front surface of a retardation plate showing the vibration directions of the ordinary and extraordinary rays. In a positive uniaxial crystal, the fast and slow axes are as indicated in parentheses; in a negative uniaxial crystal, the two axes are interchanged. Adapted with permission from Bennett JM (1995) Polarization. In Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.

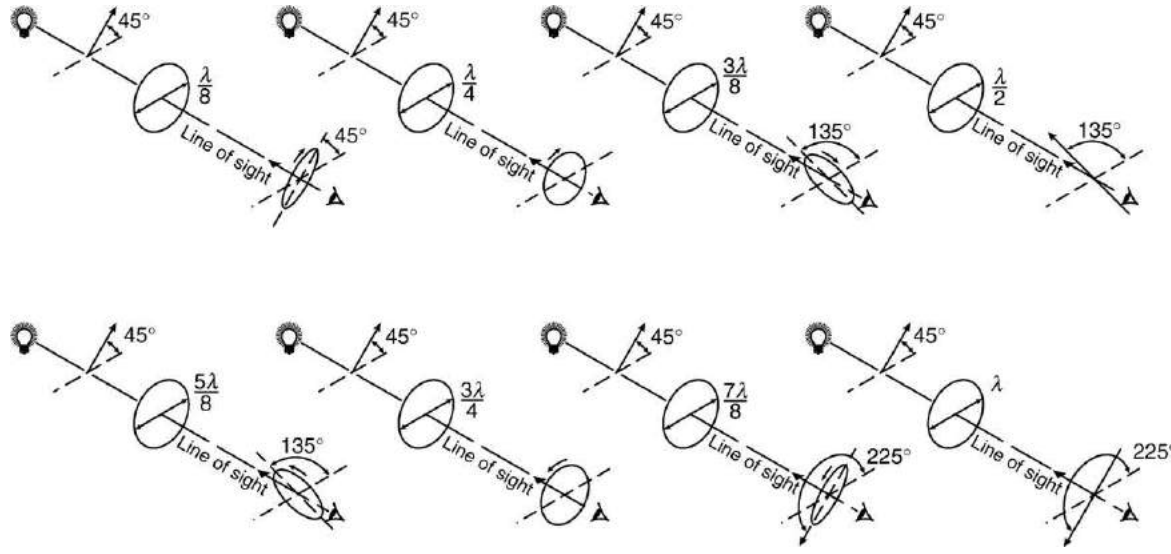


Fig. 10 State of polarization of a light wave after passing through a crystal plate whose retardation is indicated in fractions of a wavelength (phase retardation $2\pi/\lambda$ times these values) and whose fast axis is indicated by the double arrow. In all cases the incident light is linearly polarized at an azimuth of 45° to the direction of the fast axis. Adapted with permission from Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.

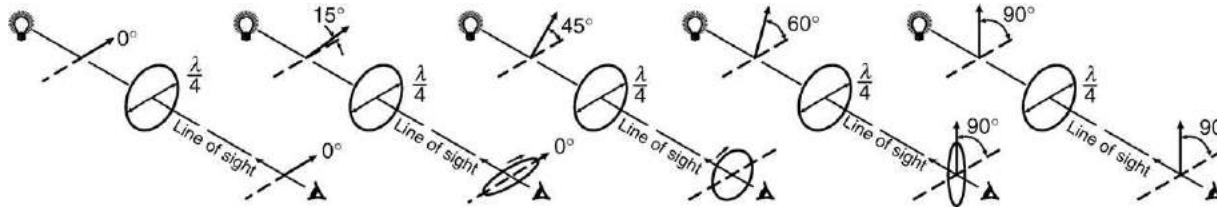


Fig. 11 State of polarization of a light wave after passing through a $\lambda/4$ plate (whose fast axis is indicated by the double arrow) for different azimuths of the incident linearly polarized beam. Adapted with permission from Bennett JM (1995) Polarization. In Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.

such as might be produced by transmission through a birefringent material or by reflection from a metal or film-covered surface. The term compensator is often applied to a variable retardation plate since it can be used to compensate for the phase retardation produced by a material. Common types of compensators include the Babinet and Soleil compensators, in which the total thickness of a birefringent material in the light path is changed, the Sénarmont compensator which consists of a fixed quarter-wave plate and rotatable analyzer to compensate for varying amounts of ellipticity in a light beam, and tilting-plate compensators, which change the thickness of birefringent material in the light beam by changing the angle of incidence. Electro-optic and piezo-optic modulators can also be used as high-frequency variable retardation plates since their birefringence can be changed by varying the electric field or pressure (see next section).

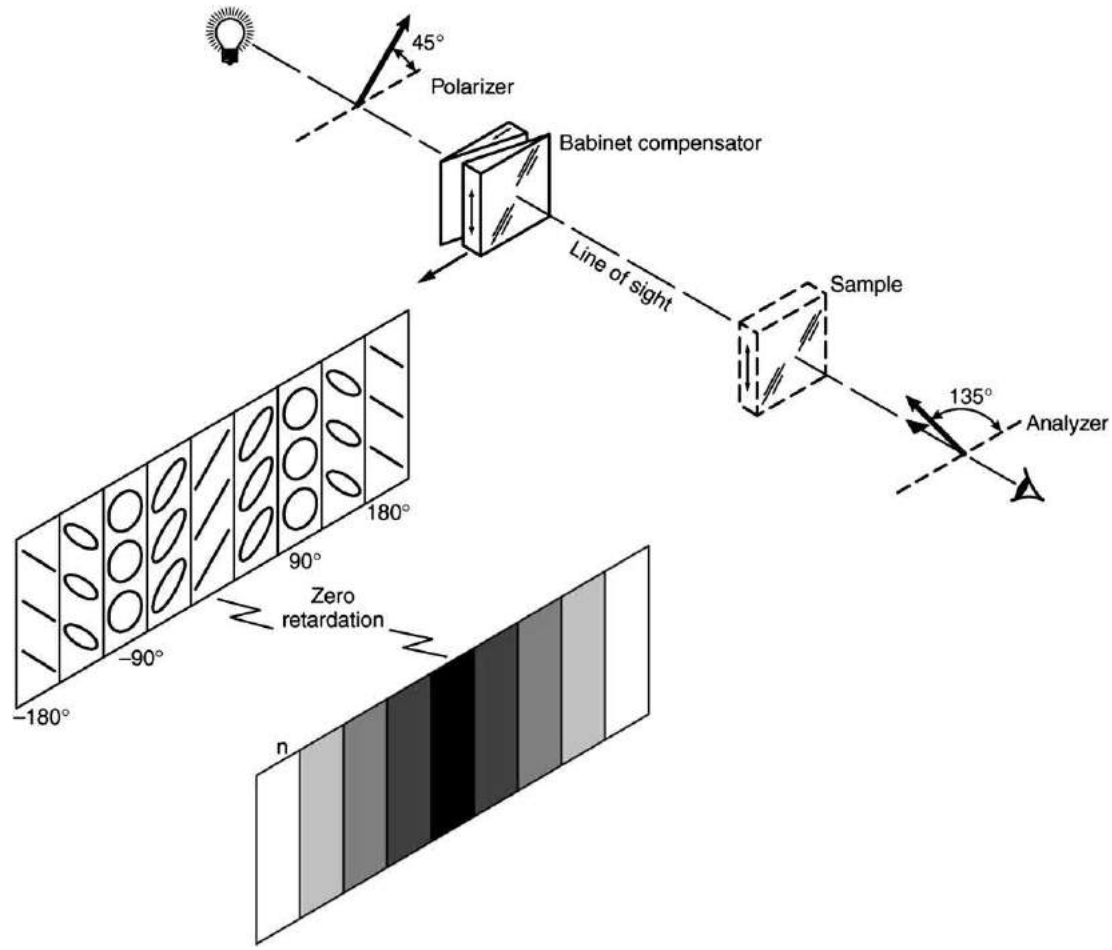


Fig. 12 Arrangement of a Babinet compensator, polarizer, and analyzer for measuring the retardation of a sample. The appearance of the field after the light has passed through the compensator is shown to the left of the sample position. Retardations are indicated for alternate regions. After the beam passes through the analyzer, the field is crossed by a series of dark bands, one of which is shown to the left of the analyzer. Adapted with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

The Babinet compensator, shown schematically in **Fig. 12** consists of two crystal quartz wedges, each with its axis in the plane of the face but with the axes at right angles to each other. One wedge is stationary, and the other can be moved in the direction indicated by the arrow, so that the total amount of quartz through which the light passes can be varied. The total retardation is proportional to the difference in thickness between the two wedges. This type of compensator was used extensively when the light source was a vertical slit and visual measurements were made of the state of polarization of a beam. However the bands representing different phase retardations were too narrow to be used effectively with a photoelectric detector, so the Babinet compensator was replaced by a Soleil compensator (**Fig. 13**). This device, sometimes called a Babinet–Soleil compensator, is similar to the Babinet compensator except that the field of view has a uniform tint if the compensator is constructed correctly. This is because the ratio of the thicknesses of the two crystal quartz blocks (a movable wedge and a fixed wedge attached to a plate with the two axes perpendicular to each other) is the same over the entire field. The Soleil compensator will produce light of varying ellipticity depending on the position of the movable wedge. It is used in the same way as a Babinet compensator with the uniformly dark field of the Soleil compensator corresponding to the black zero-retardation band in the Babinet compensator.

It is sometimes necessary to accurately measure the azimuth of a beam of linearly polarized light using a photoelectric detector with a rotatable analyzer (i.e., a polarizer) directly in front it. The obvious method is to rotate the analyzer until the detector signal is a minimum and then read the analyzer angle, which equals the extinction angle (perpendicular to the azimuth of the linearly polarized beam). However, the angle can be determined more precisely if the analyzer is offset by a small angle from the extinction angle and the transmittance noted. Then the analyzer is offset by a small angle on the other side of the extinction angle at the angle where the transmittance is the same. One half the difference between these two azimuthal angles gives a more accurate value of the extinction angle than can be obtained by measuring it directly.

Before the days of photoelectric detectors, half-shade devices were extensively used to determine the azimuth of a linearly polarized beam. The device consisted of two polarizers with their axes inclined at a small angle to each other. As the device was

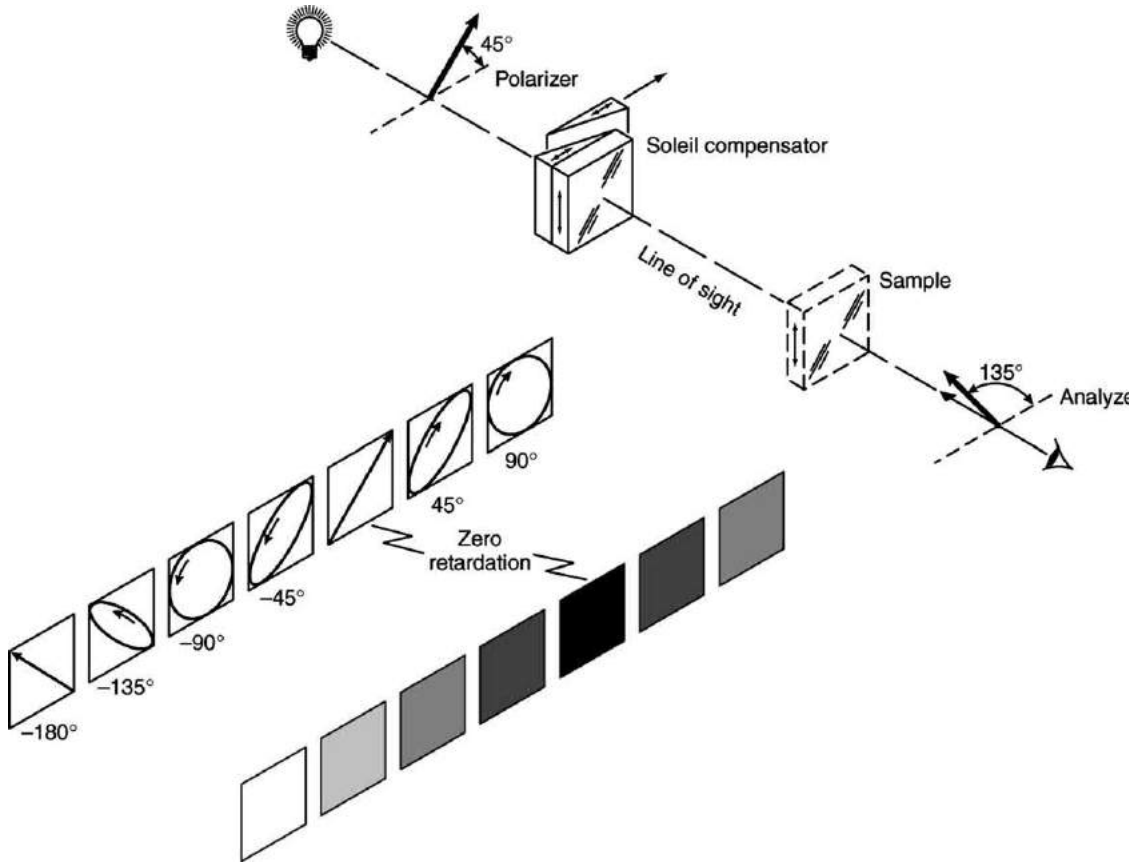


Fig. 13 Arrangement of a Soleil compensator, polarizer, and analyzer for measuring the retardation of a sample. The appearance of the field after the light has passed through the compensator is shown to the left of the sample position. After the beam passes through the analyzer, the field appears as one of the shades of gray shown to the left of the analyzer. Adapted with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

rotated, one part of the field became darker and the other part became lighter. At the match position, both parts of the field appeared equally bright. There were a variety of these devices as well as ellipticity half-shade devices that could detect very small amounts of ellipticity in a nominally linearly polarized beam and hence could verify when a compensator had completely converted elliptically polarized light into linearly polarized light.

Electro-Optic, Magneto-Optic, and Piezo-Optic Devices

The state of polarization of light can be rapidly altered by passing it through a material that has electro-optic, magneto-optic or piezo-optic properties. The voltage, magnetic field, or pressure are varied to make the material birefringent (see above). Some materials are isotropic without the applied field, i.e., the refractive index is the same in all directions. Other materials have an optic axis (uniaxial materials) or two optic axes (biaxial materials). In these cases, the applied field creates further asymmetries in the material.

The most common electro-optic, magneto-optic and piezo-optic effects are the Pockel's effect (depending on the electric field), the Kerr effect (depending on the square of the electric field), the Faraday effect (depending on the magnetic field), the Cotton–Mouton and Voigt effects (depending on the square of the magnetic field), and stress birefringence or the photoelastic effect (depending on pressure changes). These effects are shown in **Table 1** along with order of magnitude strengths needed to produce changes in the refractive index, and materials that most strongly exhibit the effects. The mathematical descriptions of the effects involve tensors and are too involved to present here. However, simple physical descriptions will be given.

If a varying electric field acts on an electro-optic material, electrons, ions, or permanent dipoles in the material are made to reorient, inducing an electric polarization. The induced polarization creates birefringence that modifies the optical polarization of a light beam passing through the material. The electric field strength determines how the polarization is changed (see **Fig. 10**).

Linearly polarized light passing through a magneto-optic material will be rotated in a direction parallel to the direction of the applied magnetic field. This phenomenon, called the Faraday effect, or Faraday rotation, is similar to what happens in optically

Table 1 Electro-optic, magneto-optic, and piezo-optic effects

Effect	*Device*	*Effect proportional to*	*Strength (Δn)*	*Materials*
Pockels (electro-optic)	Pockels cell	E	$10^{-12}E$	$BaTiO_2$, $LiNBO_3$
Kerr	Kerr cell	E^2	$10^{-19}E^2$	Nitrobenzene, Benzonitrile
Faraday (magneto-optic)	Faraday rotation	H	$10^{-14}H$	Zns, GaAs, CS_2
Cotton–Mounton	[a]	H^2	$10^{-25}H^2$	Chloroform, acetone
Photoelastic effect; stress birefringence	[b]	Pressure	–	Fused silica, polystyrene, Ge, KDP, ruby

[a]Too small to be of technological importance.

[b]Depends on the mounting or the processing of the material. Adapted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley and Sons.

active materials (see the next paragraph) except that the Faraday effect depends on the direction of the magnetic field, but not on the direction the light is traveling.

There are naturally optically active materials such as camphor, nicotine, sugar solutions, and crystal quartz that can rotate the plane of linearly polarized light passing through the material. For example, the Si–O bonds in crystal quartz form a helical path around the optic axis. This crystal structure interacts with an incoming linearly polarized beam traveling parallel to the optic axis and rotates it in a clockwise direction. A 1 mm-thick piece of quartz will rotate a linearly polarized beam by 21.7°. Other materials have asymmetric structures that have mirror images. A mirror image structure cannot be obtained by simply rotating the group of atoms in space. These materials are also optically active. The amount of rotation produced by an optically active material is proportional to the thickness (optical density) of the material that the light passes through. Dextrose (a form of glucose) and levulose (also known as fructose) rotate the plane of linearly polarized light in clockwise and counterclockwise directions, respectively. Many other organic molecules have D- (right handed, dextro-rotary) and L- (left handed, levo-rotary) mirror image forms.

An isotropic material can become anisotropic when stress or an induced strain is applied because of an elasto-optic interaction known as stress birefringence, or the photoelastic effect. This effect is usually bad because it reduces the performance of optical components and devices by introducing phase distortions caused by improper mounting or unequal thermal expansion between the mounts and components. Another source of distortion is strain that has been frozen into optical components during manufacture. French curves are excellent examples that show colored strain birefringence patterns when viewed between crossed polarizers. In one application, strain birefringence has been used constructively to produce a variable phase retarder made from crystal quartz and fused silica. A block of electro-optic crystal quartz, is cemented to a block of isotropic fused silica. When a variable electric field is applied to the crystal quartz, its length changes at the resonant frequency of the block. This produces strain in the fused silica block which in turn produces a variable retardation of light passing through the strained fused silica. Depending on the magnitude of the varying electric field, the device can act like a variable quarter-wave plate, a variable half-wave plate, or have other applications.

Matrix Methods for Computing Polarization

An optical system containing various polarizing and retarding devices can be modeled using matrices. In general, there is an incident beam (in matrix form), that has some state of polarization. It interacts with a device called an instrument (also in matrix form) that alters the state of polarization, so that the exiting beam (a third matrix, the product of the first and second matrices) has another state of polarization. There can be more than one instrument matrix acting on the same incident matrix to produce the final matrix. The two most common matrix methods are the Mueller calculus and the Jones calculus.

There is also a visual representation of this process, a Poincaré sphere, on which vectors represent different states of polarization. The polarization instrument moves the vector representing the incident polarization around the sphere in a prescribed manner. The vectors are called Stokes vectors and are used in the Mueller calculus. The various states of polarization are represented on the Poincaré sphere as follows: The equator represents various forms of linear polarization, the poles represent right- and left-circular polarization, and other points on the sphere represent elliptically polarized light. Every point on the sphere corresponds to a different polarization form. The radius of the sphere represents the incident irradiance of the light beam (which is usually assumed to be unity). The effects of various polarization instruments are determined by displacements on the sphere. Partially polarized light or absorption may be dealt with approximately by ignoring the irradiance factor, since the state of polarization is generally the quantity desired. The Poincaré sphere is most useful for visualizing problems involving nonabsorbing materials, various polarization instruments including polarizers, retarders, compensators, half-shade devices, and depolarizers, and has also been applied to ellipsometric problems and stress-optical measurements.

The Poincaré sphere is used with Stokes vectors, which are often designated S_0, S_1, S_2, and S_3. S_0 is the incident irradiance of the light beam, corresponding to the radius of the Poincaré sphere. S_1 is the difference in irradiances between the horizontal and vertical polarization components of the beam; for example, when S_1 is positive, the preference is for horizontal polarization. S_2 indicates preference for $+45^\circ$ or -45° polarization depending on whether it is positive or negative, and S_3 gives the preference for right or left circular polarization. Thus, the Stokes vectors S_1, S_2, and S_3 are simply the three Cartesian coordinates of a point on the

Poincaré sphere. S_1 and S_2 are perpendicular to each other in the equatorial plane, and S_3 points toward the north pole of the sphere. Any state of polarization of a light beam can be specified by these three vectors. The irradiance vector S_0 is related to the other three by the relation $S_0^2 = S_1^2 + S_2^2 + S_3^2$ when the beam is completely polarized. If the beam is partially polarized, $S_0^2 > S_1^2 + S_2^2 + S_3^2$.

In the Mueller calculus, the incident beam is represented by the four-component Stokes vector, written as a column vector. This vector has all real elements and gives information about irradiance properties of the beam. Thus it is not able to handle problems involving phase changes or combinations of two beams that are coherent. The instrument matrix is a 4×4 matrix with all real elements.

In the Jones calculus, the Jones vector representing the incident beam is a two-component column vector that generally has complex elements. It contains information about the *amplitude* properties of the beam and hence is well suited for handling problems involving the phases of light waves. However, it cannot handle problems involving depolarization, as the Mueller calculus can. The Jones instrument matrix is a 2×2 matrix whose elements are generally complex. The Jones calculus is well suited to problems involving a large number of similar devices arranged in series in a regular manner and makes it possible to obtain a result expressed explicitly in terms of the number of such devices. The Jones instrument matrix of a train of transparent or absorbing nondepolarizing polarizers and retarders contains no redundant information. The matrix contains four elements, each of which has two parts, so that there are a total of eight constants, none of which is a function of any other. The Mueller instrument matrix of such a train contains much redundancy; there are 16 constants but only 7 of them are independent. More information about the Poincaré sphere and the Mueller and Jones calculus is available in other references.

See also: Matrix Analysis

Further Reading

Auton, J.P., 1967. Infrared transmission polarizers by photolithography. Applied Optics 6, 1023–1027.
Azzam, Rasheed, M.A., Bashara, N.M., 1977. Ellipsometry and Polarized Light. Amsterdam: North-Holland Publishing Co (chaps. 1 and 2).
Azzam, Rasheed, M.A., 1995. In: Bass, M. (Ed.), Handbook of Optics, Ellipsometry, 2nd edn., vol. II. New York: McGraw-Hill (chap. 27).
Bennett, J.M., 1995. In: Bass, M. (Ed.), Handbook of Optics, Polarization, 2nd edn., vol. I. New York: McGraw-Hill (chap. 5).
Bennett, M., 1995. In: Bass, M. (Ed.), Handbook of Optics, Polarizers, 2nd edn., vol. II. New York: McGraw-Hill (chap. 3).
Bohren, C.F., Huffman, D.R., 1983. Absorption and Scattering of Light by Small Particles. New York: John Wiley.
Born, M., Wolf, E., 1999. Principles of Optics, 7th edn, Cambridge: Cambridge University Press (chap. 1).
Chipman, R.A., 1995. In: Bass, M. (Ed.), Handbook of Optics, Polarimetry, 2nd edn., vol. II. New York: McGraw-Hill (chap. 22).
Clark, J.R., 1956. New calculus for the treatment of optical systems. VIII. Electromagnetic theory. Journal of the Optical Society of America 46, 126–131.
Guenther, R.D., 1990. Modern Optics. New York: John Wiley and Sons (chaps. 2, 13, and 14).
Hass, M., O'Hara, M., 1965. Sheet infrared transmission polarizers. Applied Optics 4, 1027–1031.
Hecht, E., Zajac, A., 1997. Optics, 3rd edn. Reading, MA: Addison-Wesley Publishing Co (chap. 8).
Jackson, J.D., 1999. Classical Electrodynamics, 3rd edn. New York: John Wiley and Sons, Inc (chap. 7).
Jasperson, S.N., Schnatterly, S.E., 1969. An improved method for high reflectivity ellipsometry based on a new polarization modulation technique. Reviews of Scientific Instruments 40, 761–767.
Kerker, M., 1969. The Scattering of Light and Other Electromagnetic Radiation. New York: Academic Press.
Kliger, D.S., Lewis, J.W., Randall, C.E., 1990. Polarized Light in Optics and Spectroscopy. San Diego: Academic Press (chaps. 1–5).
Maldonado, T.A., 1995. In: Bass, M. (Ed.), Handbook of Optics, Electro-optic modulators, 2nd edn., vol. II. New York: McGraw-Hill, Inc. (chap. 13).
Markuvitz, N. (Ed.), 1951. Waveguide Handbook. New York: McGraw-Hill Book Co, pp. 218–229. 285–289.
Shurcliff, W.A., 1962. Polarized Light, Production and Use. Cambridge, MA: Harvard University Press.
van de Hulst, H.C., 1957. Light Scattering by Small Particles. New York: John Wiley.

Matrix Analysis

BD Guenther, Duke University, Durham, NC, USA

Introduction

Hooke postulated, in the seventeenth century, that light waves must be transverse but his idea was forgotten. Young and Fresnel put forward the same idea in the nineteenth century and accompanied their postulation with a theoretical description of light based on transverse waves. Forty years later, Maxwell proved that light must be a transverse wave and that **E** and **H**, for a plane wave in an isotropic medium with no free charge and no currents, are mutually perpendicular and lie in a plane normal to the direction of propagation, **k**.

Convention requires that we use the electric vector to label the direction of the electromagnetic wave's polarization. The direction of the displacement vector is called the direction of polarization and the plane containing the direction of polarization and the propagation vector is called the plane of polarization. The selection of the electric field is not completely arbitrary, except in relativistic situations, when $v \approx c$, the interaction of the electromagnetic wave with matter will be dominated by the electric field.

Assume that a plane wave is propagating in the z-direction. In complex notation, the plane wave is given by

$$\tilde{\mathbf{E}} = \mathbf{E}_0 e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi)} = \mathbf{E}_0 e^{i(\omega t - kz + \phi)} \tag{1}$$

The procedure used to decompose an arbitrary polarization into components parallel to two axes of a Cartesian coordinate system, is a technique used extensively in vector algebra to simplify mathematical calculations. According to the mathematical formalism associated with this technique, the polarization is described in terms of a set of basis vectors, $\mathbf{e}_i$. An arbitrary polarization would be expressed as

$$\mathbf{E} = \sum_{i=1}^{2} a_i \mathbf{e}_i \tag{2}$$

The set of basis vectors, $\mathbf{e}_i$, are orthonormal, i.e.:

$$\mathbf{e}_i \mathbf{e}_j^* = \delta_{ij} = \left\{ \begin{array}{ll} 1 & i = j \\ 0 & i \neq j \end{array} \right\} \tag{3}$$

where we have assumed that the basis vectors could be complex. We mention this mathematical formalism because an identical formalism is encountered in a description of spin.

In a Cartesian coordinate system, the $\mathbf{e}_i$'s are the unit vectors **i**, **j**, **k**. The summation in **Eq. (2)** extends over only two terms because the electromagnetic wave is transverse, confining **E** to a plane normal to the direction of propagation (according to the coordinate convention we have selected, the **E** field is in the x, y plane).

Polarization Ellipse

Following the formalism of **Eq. (2)**, a polarized wave can be written in terms of the x and y components of $\mathbf{E}_0$

$$\tilde{\mathbf{E}} = \mathbf{E}_{0x} e^{i(\omega t - kz + \phi_1)} \hat{\mathbf{i}} + \mathbf{E}_{0y} e^{i(\omega t - kz + \phi_2)} \hat{\mathbf{j}} \tag{4}$$

We will use only the real part of **E** for manipulation, to prevent errors. We divide each component of the electric field by its maximum value, so that the problem is reduced to one of the following two sinusoidal varying unit vectors:

$$\begin{aligned} \frac{E_x}{E_{0x}} &= \cos(\omega t - kz + \phi_1) \\ &= \cos(\omega t - kz)\cos\phi_1 - \sin(\omega t - kz)\sin\phi_1 \\ \frac{E_y}{E_{0y}} &= \cos(\omega t - kz)\cos\phi_2 - \sin(\omega t - kz)\sin\phi_2 \end{aligned} \tag{5}$$

When these unit vectors are added together, the resulting equation will describe the path taken by the tip of the resultant vector. The path will create a Lissajous figure.

To obtain the equation for the Lissajous figure, we eliminate the dependence of the unit vectors on $(\omega t - kz)$. First, multiply the equations by $\sin\phi$ and $\sin\phi_1$, respectively and then subtract the resulting equations. Second, multiply the two equations by $\cos\phi_2$ and $\cos\phi_1$ respectively and then subtract the new equations. These two operations yield the following pair of equations:

$$\begin{aligned} &\frac{E_x}{E_{0x}}\sin\phi_2 - \frac{E_y}{E_{0y}}\sin\phi_1 \\ &\quad = \cos(\omega t - kz)[\cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2] \end{aligned} \tag{6}$$

 doi:10.1016/B978-0-12-809283-5.00692-3

$$\frac{E_x}{E_{0x}}\cos\phi_2 - \frac{E_y}{E_{0y}}\cos\phi_1 = \sin(\omega t - kz)[\cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2] \tag{7}$$

The term in brackets can be simplified using the trig identity

$$\begin{aligned}\sin\delta &= \sin(\phi_2 - \phi_1)\\ &= \cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2\end{aligned} \tag{8}$$

After replacing the term in brackets by sin δ, the two equations are squared and added yielding the equation for the Lissajous figure:

$$\left(\frac{E_x}{E_{0x}}\right)^2 + \left(\frac{E_y}{E_{0y}}\right)^2 - \left(\frac{2E_xE_y}{E_{0x}E_{0y}}\right)\cos\delta = \sin^2\delta \tag{9}$$

The trig identity

$$\cos\delta = \cos(\phi_2 - \phi_1) = \cos\phi_1 \cos\phi_2 + \sin\phi_1 \sin\phi_2$$

was also used to further simplify **Eq. (9)**.

Eq. (9) has the same form as the equation of a conic

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

Geometry defines the conic as an ellipse because, from **Eq. (9)**

$$B^2 - 4AC = \frac{4}{E_{0x}^2E_{0y}^2}(\cos^2\delta - 1) < 0$$

This ellipse is called the polarization ellipse. The orientation of the ellipse, with respect to the x-axis, is

$$\tan 2\theta = \frac{B}{A - C} = \frac{2E_{0x}E_{0y}\cos\delta}{E_{0x}^2 - E_{0y}^2} \tag{10}$$

The tip of the resultant electric field vector obtained from **Eq. (4)** traces out the polarization ellipse in the plane normal to **k**, as predicted by **Eq. (9)**. The ratio of the length of the minor to the major axis of the ellipse is equal to the ellipticity, φ, i.e., the amount of deviation of the ellipse from a circle

$$\begin{aligned}\tan\varphi &= \pm\left(\frac{E_m}{E_M}\right)\\ &= \frac{E_{0x}\sin\phi_1 \sin\theta - E_{0y}\sin\phi_2 \cos\theta}{E_{0x}\cos\phi_1 \cos\theta + E_{0y}\cos\phi_2 \sin\theta}\end{aligned} \tag{11}$$

In particle physics, the light would be said to have a negative helicity if it rotated in a clockwise direction. If we look at the source, the electric vector seems to follow the threads of a left-handed screw, agreeing with the nomenclature that left-handed quantities are negative. However, in optics the light that rotates clockwise, as we view it traveling toward us from the source, is said to be right-circularly polarized. The counterclockwise rotating light is left-circularly polarized. The association of right-circularly polarized light, with 'right-handedness' in optics, came about by looking at the path of the electric vector in space at a fixed time, then tan $\psi = \tan(\phi - kz)$ (**Fig. 1**).

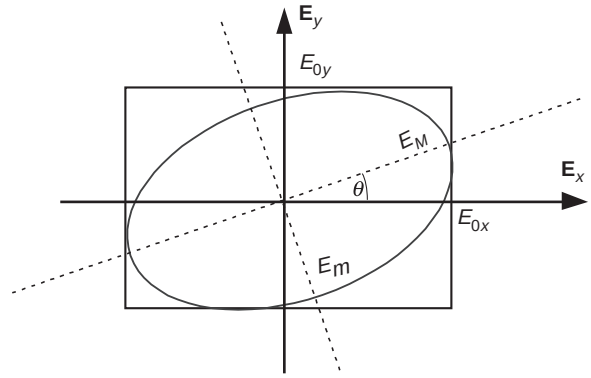


Fig. 1 General form of the ellipse described by **Eq. (9)**. Reproduced with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley and Sons.

Table 1 Typical Stokes vectors

Horizontal polarization	$\begin{bmatrix}1\\1\\0\\0\end{bmatrix}$	Vertical polarization	$\begin{bmatrix}1\\-1\\0\\0\end{bmatrix}$
$+45^\circ$ polarization	$\begin{bmatrix}1\\0\\1\\0\end{bmatrix}$	-45° polarization	$\begin{bmatrix}1\\0\\-1\\0\end{bmatrix}$
Right circular polarization	$\begin{bmatrix}1\\0\\0\\1\end{bmatrix}$	Left circular polarization	$\begin{bmatrix}1\\0\\0\\-1\end{bmatrix}$

Stokes Parameters

In the formalism associated with Eq. (2), the expansion coefficients, a_i, can be used to form a 2–2 matrix which, in statistical mechanics, is called the density matrix and in optics, the coherency matrix (Table 1). The elements of the matrix are formed by the rule

$$\rho_{ij} = \mathbf{a}_i \mathbf{a}_j^* \tag{12}$$

The matrix is Hermitian, so that $\rho_{ij} = \rho_{ij}^*$. We will not develop the theory of polarization using the coherency matrix, but simply use the coherency matrix to justify the need for four independent measurements to characterize polarization. There is no unique set of measurements required by the theory, but normally measurements made are of the Stoke's parameters which are directly related to the parameters of the polarization ellipse of Eq. (9).

The Stokes parameters of a light wave are measurable quantities, defined as:

$s_0 \Rightarrow$ Total flux density
$s_1 \Rightarrow$ Difference between flux density transmitted by a linear polarizer oriented parallel to the x-axis and one oriented parallel to the y-axis. The x and y-axes are usually selected to be parallel to the horizontal and vertical directions in the laboratory.
$s_2 \Rightarrow$ Difference between flux density transmitted by a linear polarizer oriented at 45° to the x-axis and one oriented at 135°.
$s_3 \Rightarrow$ Difference between flux density transmitted by a right-circular polarizer and a left-circular polarizer.

If the Stokes parameters are to characterize the polarization of a wave, they must be related to the parameters of the polarization ellipse. It is therefore important to establish that the Stokes parameters are variables of the polarization ellipse (Eq. (9)).

In its current form, Eq. (9) contains no measurable quantities and thus must be modified if it is to be associated with the Stokes parameters. The time average of the Poynting vector is the quantity observed when measurements are made of light waves. We must, therefore, find the time average of Eq. (9) if we wish to relate its parameters to observable quantities.

The time average Eq. (9) can now be written as

$$\frac{\langle E_x^2\rangle}{E_{0x}^2} + \frac{\langle E_y^2\rangle}{E_{0y}^2} - 2\frac{\langle E_x E_y\rangle}{E_{0x}E_{0y}}\cos\delta = \sin^2\delta \tag{13}$$

where time average is denoted by $\langle\ \rangle$

$$\langle E_x^2\rangle = \frac{1}{T}\int_{t_0}^{t_0+T} E_{0x}^2[\cos(\omega t - kz)\cos\phi_1 - \sin(\omega t - kz)\sin\phi_1]^2 \mathrm{d}t \tag{14}$$

With the time averages, Eq. (13) can be written as

$$\begin{aligned}(E_{0x}^2 + E_{0y}^2)^2 - (E_{0x}^2 - E_{0y}^2)^2 - (2E_{0x}E_{0y}\cos\delta)^2 \\ = (2E_{0x}E_{0y}\sin\delta)^2\end{aligned} \tag{15}$$

Each term in this equation can be identified with a Stokes parameter:

$$\begin{aligned}s_0 &= \langle E_{0x}^2\rangle + \langle E_{0y}^2\rangle \qquad s_1 = \langle E_{0x}^2\rangle - \langle E_{0y}^2\rangle \\ s_2 &= \langle 2E_{0x}E_{0y}\cos\delta\rangle \quad s_3 = \langle 2E_{0x}E_{0y}\sin\delta\rangle\end{aligned} \tag{16}$$

Eq. (15) can now be written as

$$s_0^2 - s_1^2 - s_2^2 = s_3^2 \tag{17}$$

For a polarized wave, only three of the Stokes parameters are independent. This agrees with the requirement placed upon elements of the Hermitian, coherency matrix, introduced above.

With this demonstration of the connection between the Stokes parameters and the polarization ellipse, the Stokes parameters can be written in terms of the parameters of the polarization ellipse.

$$\begin{aligned} s_1 &= s_0 \cos 2\varphi \cos 2\theta \\ s_2 &= s_0 \cos 2\varphi \sin 2\theta \\ s_3 &= s_0 \sin 2\varphi \end{aligned} \tag{18}$$

Mueller Calculus

Mueller pointed out that the Stokes parameters can be thought of as elements of a column matrix or a 4-vector (Table 2)

$$\mathcal{S} = \begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{pmatrix} \tag{19}$$

Table 2 Mueller matrices for polarizers

Polarizer	*Transmission axis*	*Mueller matrix*
Linear	Horizontal	$\frac{1}{2}\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
	Vertical	$\frac{1}{2}\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
	$+45°$	$\frac{1}{2}\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
	$-45°$	$\frac{1}{2}\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
Circular	Right	$\frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$
	Left	$\frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}$

Each element of the Stokes vector represents a measurable intensity. The vector can represent not only polarized 1ight, but unpolarized or partially polarized light.

In 1929, Soleillet discovered that an optical device performed a linear transformation on the input wave and, in 1942, Perrin put this fact into a matrix formalism involving the Stokes vectors:

$$\mathcal{S}_1 = \mathcal{M} \cdot \mathcal{S}_0 \tag{20}$$

Mueller used experimentally derived 4×4 matrices, the $\mathcal{M}$ in **Eq. (20)**, to describe the effect of an optical device on a light wave's polarization. The $\mathcal{S}$'s in **Eq. (20)** are column matrices whose elements are the Stokes parameters (**Eq. (16)**). The matrices are based upon an assumed linear relationship between the incident and transmitted beams (**Table 3**).

The analysis of the effect of a number of polarizers and retarders is made easier by the use of the Mueller–Stokes matrix calculus, coupled with the use of the Stokes vectors. To determine the Mueller matrices, one must measure the effect a device has on unpolarized, horizontally polarized, linearly polarized at $+45^\circ$, and right-circularly polarized light. It is then an algebraic exercise to generate the elements of the matrix. A few optical devices and their Mueller matrix are listed in **Tables 2** and **3**.

The Mueller matrix contains 16 parameters but there are only seven independent parameters. The matrix contains no information about absolute phase but it handles partially polarized and unpolarized light without modification (**Table 4**).

Table 3 Mueller matrices for retarders

Retarder	*Transmission axis*	*Mueller matrix*
Quarter-wave plate	Horizontal	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}$
	Vertical	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$
	$+45^\circ$	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$
	-45°	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}$
Half-wave plate	0° or 90°	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$
	$\pm 45^\circ$	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$

Table 4 Jones vectors

Horizontal polarization	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	Vertical polarization	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$
$+45°$ polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$-45°$ polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$
Right circular polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ i \end{bmatrix}$	Left circular polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -i \end{bmatrix}$

Table 5 Jones matrices for retarders

Retarder	*Transmission axis*	*Jones matrix*
	Horizontal	$\begin{bmatrix} e^{i\pi/4} & 0 \\ 0 & e^{-i\pi/4} \end{bmatrix}$
	Vertical	$\begin{bmatrix} e^{-i\pi/4} & 0 \\ 0 & e^{i\pi/4} \end{bmatrix}$
	$+45°$	$\frac{1}{2}\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$
Quarter-wave plate	$-45°$	$\frac{1}{2}\begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}$
Half-wave plate	$0°$ or $90°$	$\begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}$
	$\pm 45°$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

Jones Vector

There is one other representation of polarized light, complementary to the Stokes vector, developed by Clark Jones in 1941 and called the Jones vector. It is superior to the Stokes vector in that it handles light of a known phase and amplitude with a reduced number of parameters. However, it is inferior to the Stokes vector in that, unlike the Stokes representation, which is experimentally determined, the Jones representation cannot handle unpolarized or partially polarized light. The Jones vector is a theoretical construct that can only describe light with a well-defined phase and frequency. The density matrix formalism can be used to correct the shortcomings of the Jones vector, but then the simplicity of the Jones representation is lost.

Assuming the coordinate system is such that the electromagnetic wave is propagating along the z-axis, it was shown earlier that any polarization could be decomposed into two orthogonal **E** vectors, i.e., parallel to the x and y directions. The Jones vector is defined as a two-row, column matrix consisting of the complex components in the x and y direction:

$$\mathbf{E} = \begin{bmatrix} E_{0x}e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi_1)} \\ E_{0y}e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi_2)} \end{bmatrix} \tag{21}$$

If absolute phase is not an issue, then we may normalize the vector by dividing by that number (real or complex) that simplifies the components but keeps the sum of the square of the components equal to one. For example, assume that $E_{0x} = E_{0y}$, then

$$\mathbf{E} = E_{0x}e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi_1)}\begin{bmatrix} 1 \\ e^{i\delta} \end{bmatrix} \tag{22}$$

The normalized vector would be the terms contained within the bracket, each divided by $\frac{1}{\sqrt{2}}$.

Table 6 Jones matrices for polarizers

Polarizer	*Transmission axis*	*Jones matrix*
Linear	Horizontal	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$
	Vertical	$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$
	$+45°$	$\frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$
	$-45°$	$\frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$
Circular	Right	$\frac{1}{2}\begin{bmatrix} 1 & -i \\ i & 1 \end{bmatrix}$
	Left	$\frac{1}{2}\begin{bmatrix} 1 & i \\ -i & 1 \end{bmatrix}$

The general form of the Jones vector is

$$\mathbf{E} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \quad \mathbf{E}^* = [\mathbf{A}^* \mathbf{B}^*] \tag{23}$$

Some examples of Jones vectors are shown in **Table 4**.

Jones Calculus

The Jones calculus is complementary to the Mueller calculus and operates on the Jones vector, (**Eq. (23)**), similar to the way the Mueller matrix operates on the Stokes vector (**Table 5**).

The Jones matrix contains eight independent parameters with no redundancy, making it simpler than the Mueller calculus. However, the Jones calculus only applies to polarized light. The Jones calculus can be extended, using the density matrix formalism to allow manipulation of unpolarized light, but with a loss of simplicity. The matrix equation for Jones calculus is

$$\mathbf{E}_{\text{out}} = \mathcal{W}\mathbf{E}_{\text{in}} \tag{24}$$

A few optical devices and their Jones matrices, $\mathcal{W}$, are listed in **Tables 5** and **6**.

For every matrix in Jones calculus, there is a matrix in Mueller calculus, but the converse is not true. For example, it is possible to construct a depolarizer, by using a thick piece of opal glass in the visible or by using gold covered sandpaper in the infrared. Such a device can be described in Mueller calculus, but there is no matrix for such a device in Jones calculus (**Tables 5** and **6**).

See also: Polarization Introduction

Further Reading

Born, M., Wolf, E., 1970. Principles of Optics. New York: Pergamon Press.
Clark Jones, R., 1941. A new calculus for the treatment of optical systems. Journal of the Optical Society of America 32, 488–503.
Clark Jones, R., 1942. A new calculus for the treatment of optical systems. Journal of the Optical Society of America 31, 486–493.
Guenther, R.D., 1990. Modern Optics. New York: Wiley.
Hecht, E., 1998. Optics, 3rd edn. Reading, MA: Addison-Wesley.
Klein, M.V., 1970. Optics. New York: Wiley.
Kliger, D.S., Lewis, J.W., Randall, C.E., 1990. Polarized Light in Optics and Spectroscopy. Boston: Academic Press.
McMaster, W.H., 1954. Polarization and the Stokes parameters. American Journal of Physics 22, 351–362.
Mueller, H., 1948. The foundation of optics. Journal of the Optical Society of America 38, 661.

Electromagnetic Theory

SG Johnson and JD Joannopoulos, Massachusetts Institute of Technology, Cambridge, MA, USA

Introduction

Photonic crystals are periodically structured electromagnetic media, generally possessing photonic bandgaps: ranges of frequency in which light cannot propagate through the structure. This periodicity, whose lengthscale is proportional to the wavelength of light in the bandgap, is the electromagnetic analog of a crystalline atomic lattice, where the latter acts on the electron wavefunction to produce the familiar band gaps, semiconductors, etc., of solid-state physics. The study of photonic crystals is likewise governed by the Bloch–Floquet theorem, and intentionally introduced defects in the crystal (analogous to electronic dopants) give rise to localized electromagnetic states: linear waveguides and point-like cavities. The crystal can thus form a kind of perfect optical 'insulator', which can confine light around sharp bends, in lower-index media, and within wavelength-scale cavities, among other novel possibilities for control of electromagnetic phenomena. Below is introduced the basic theoretical background of photonic crystals in one, two, and three dimensions (schematically depicted in Fig. 1), as well as hybrid structures that combine photonic-crystal effects in some directions with more-conventional index guiding in other directions. (Line and point defects in photonic crystals are discussed in another article.)

Electromagnetic wave propagation in periodic media was first studied by Lord Rayleigh in 1887, in connection with the peculiar reflective properties of a crystalline mineral with periodic 'twinning' planes (across which the dielectric tensor undergoes a mirror flip). These correspond to one-dimensional photonic crystals, and he identified the fact that they have a narrow bandgap prohibiting light propagation through the planes. This bandgap is angle-dependent, due to the differing periodicities experienced by light propagating at non-normal incidences, producing a reflected color that varies sharply with angle. (A similar effect is responsible for many other iridescent colors in nature, such as those of butterfly wings and abalone shells.) Although multilayer films received intensive study over the following century, it was not until 100 years later, when Yablonovitch and John, in 1987, joined the tools of classical electromagnetism and solid-state physics, that the concepts of omnidirectional photonic bandgaps in two and three dimensions was introduced. This generalization, which inspired the name 'photonic crystal', led to many subsequent developments in their fabrication, theory, and application, from integrated optics to negative refraction to optical fibers that guide light in air.

Maxwell's Equations in Periodic Media

The study of wave propagation in three-dimensionally periodic media was pioneered by Felix Bloch in 1928, unknowingly extending an 1883 theorem in one dimension by Gaston Floquet. Bloch proved that waves in such a medium can propagate without scattering, their behavior described by a periodic envelope function multiplied by a planewave. Although Bloch studied quantum mechanics, leading to the surprising result that electrons in a conductor scatter only from imperfections and not from the periodic ions, the same techniques can be applied to electromagnetism by casting Maxwell's equations as an eigenproblem in analog with Schrödinger's equation. By combining the source-free Faraday's and Ampere's laws at a fixed (angular) frequency ω, i.e., time dependence$e^{-i\omega t}$one can obtain an equation in only the magnetic field $\vec{H}$:

$$\vec{\nabla} \times \frac{1}{\varepsilon} \vec{\nabla} \times \vec{H} = \left(\frac{\omega}{c}\right)^2 \vec{H} \tag{1}$$

where ε is the dielectric function $\varepsilon(x,y.z)$ and c is the speed of light. This is an eigenvalue equation, with eigenvalue (ω/c^2) and an eigen-operator $\vec{\nabla} \times (1/\varepsilon)\vec{\nabla}\times$ that is Hermitian (acts the same to the left and right) under the inner product $\int \vec{H}^* \cdot \vec{H}'$ between two

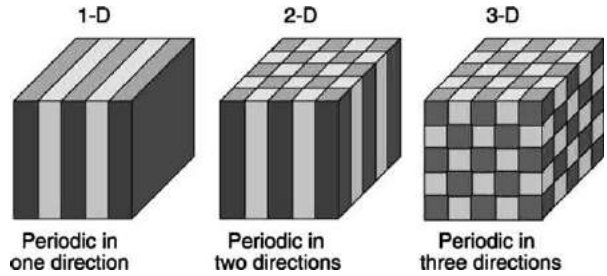


Fig. 1 Schematic depiction of photonic crystals periodic in one, two, and three directions, where the periodicity is in the material (typically dielectric) structure of the crystal. Only a 3d periodicity, with a more complex topology than is shown here, can support an omnidirectional photonic bandgap.

fields $\vec{H}$and $\vec{H}$. (The two curls correspond roughly to the 'kinetic energy' and $1/\varepsilon$ to the 'potential' compared to the Schrödinger Hamiltonian $\nabla^2 + V$.) It is sometimes more convenient to write a generalized Hermitian eigenproblem in the electric field $\vec{E}, \vec{\nabla} \times \vec{\nabla} \times \vec{E} = (\omega/c)^2 \varepsilon \vec{E}$ which separates the kinetic and potential terms. Electric fields that lie in higher ε i.e., lower potential, will have lower ω this is discussed further in the context of the variational theorem of **Eq. (3)**.

Thus, the same linear-algebraic theorems as those in quantum mechanics can be applied to the electromagnetic wave solutions. The fact that the eigen-operator is Hermitian and positive-definite (for real $\varepsilon > 0$) implies that the eigenfrequencies ω are real, for example, and also leads to orthogonality, variational formulations, and perturbation–theory relations that are discussed further below. An important difference compared to quantum mechanics is that there is a transversality constraint: one typically excludes $\vec{\nabla} \cdot \vec{H} \neq 0$ (or $\vec{\nabla} \cdot \varepsilon \vec{E} \neq 0$) eigensolutions, which lie at $\omega = 0$; i.e., static-field solutions with free magnetic (or electric) charge are forbidden.

Bloch Waves and Brillouin Zones

A photonic crystal corresponds to a periodic dielectric function $\varepsilon(\vec{x}) = \varepsilon(\vec{x} + \vec{R}_i)$ for some primitive lattice vectors $\vec{R}_i (i = 1,2,3$, for a crystal periodic in all three dimensions). In this case, the Bloch–Floquet theorem for periodic eigenproblems states that the solutions to **Eq. (1)** can be chosen of the form $\vec{H}(\vec{x}) = e^{i\vec{k}\cdot\vec{x}} \vec{H}_{n,\vec{k}}(\vec{x})$ with eigenvalues $\omega_n(\vec{k})$, where $\vec{H}_{n,\vec{k}}$ is a periodic envelope function satisfying

$$\left(\vec{\nabla} + i\vec{k}\right) \times \frac{1}{\varepsilon} \left(\vec{\nabla} + i\vec{k}\right) \times \vec{H}_{n,\vec{k}} = \frac{\omega_n(\vec{k})_2}{c} \vec{H}_{n,\vec{k}} \tag{2}$$

yielding a different Hermitian eigenproblem over the primitive cell of the lattice at each Bloch wavevector $\vec{k}$. This primitive cell is a finite domain if the structure is periodic in all directions, leading to discrete eigenvalues labeled by $n = 1,2,\ldots$ These eigenvalues $\omega_n(\vec{k})$, are continuous functions of $\vec{k}$, forming discrete 'bands' when plotted versus the latter, in a 'band structure' or dispersion diagram – both $\omega$and $\vec{k}$ are conserved quantities, meaning that a band diagram maps out all possible interactions in the system. (Note also that $\vec{k}$. is not required to be real; complex $\vec{k}$. gives evanescent modes that can exponentially decay from the boundaries of a finite crystal, but which cannot exist in the bulk.)

Moreover, the eigensolutions are periodic functions of $\vec{k}$ as well: the solution at $\vec{k}$ is the same as the solution at $\vec{k} + \vec{G}_j$ where $\vec{G}_j$ is a primitive reciprocal lattice vector defined by $\vec{R}_i \cdot \vec{G}_j = 2\pi\delta_{i\cdot j}$. Thanks to this periodicity, one need only compute the eigensolutions for $\vec{k}$ within the primitive cell of this reciprocal lattice – or, more conventionally, one considers the set of inequivalent wavevectors closest to the $\vec{k} = 0$ origin, a region called the first Brillouin zone. For example, in a one-dimensional system, where $R_1 = a$ for some periodicity a and $G = 2\pi/a$, the first Brillouin zone is the region $k = -\pi/\ldots\pi/a$; all other wavevectors are equivalent to some point in this zone under translation by a multiple of G_1. Furthermore, the first Brillouin zone may itself be redundant if the crystal possesses additional symmetries such as mirror planes; by eliminating these redundant regions, one obtains the irreducible Brillouin zone, a convex polyhedron that can be found tabulated for most crystalline structures. In the preceding one-dimensional example, since most systems will have time-reversal symmetry ($k \to -k$), the irreducible Brillouin zone would be $k = 0\ldots\pi/a$.

The familar dispersion relations of uniform waveguides arise as a special case of the Bloch formalism: such translational symmetry corresponds to a period $a \to 0$.In this case, the Brillouin zone of the wavevector k(also called β) is unbounded, and the envelope function $\vec{H}_{n,\vec{k}}$is a function only of the transverse coordinates.

The Origin of the Photonic Bandgap

A complete photonic bandgap is a range of ω in which there are no propagating (real $\vec{k}$) solutions of Maxwell's **Eq. (2)** for any $\vec{k}$ surrounded by propagating states above and below the gap. There are also incomplete gaps, which only exist over a subset of all possible wavevectors, polarizations, and/or symmetries. Both sorts of gaps are discussed in the subsequent sections, but in either case, their origins are the same and can be understood by examining the consequences of periodicity for a simple one-dimensional system.

Consider a one-dimensional system with uniform $\varepsilon = \vec{\varepsilon}$ which has planewave eigensolutions $\omega(k) = ck$ as depicted in **Fig. 2(left)**. This ε has trivial periodicity a for any $a \geq 0$, with $a = 0$ giving the usual unbounded dispersion relation. One is free, however, to label the states in terms of Bloch envelope functions and wavevectors for some $a \neq 0$, in which case the bands for $|k| > \pi/a$ are translated (folded) into the first Brillouin zone, as shown by the dashed lines in **Fig. 2(left)**. In particular, the $k = -\pi/a$ mode in this description now lies at an equivalent wavevector to the $k = \pi/a$ mode, and at the same frequency; this accidental degeneracy is an artifact of the 'artificial' period that has been chosen. Instead of writing these wave solutions with electric fields $\vec{E}(x) \sim e^{\pm i\pi x/a}$, one can equivalently write linear combinations $e(x) = \cos(\pi x/a)$ and $o(x) = \sin(\pi x/a)$ as shown in **Fig. 3**, both at $\omega = c\pi/a\sqrt{\vec{\varepsilon}}$. Now, however, suppose that one perturbs ε so that it is nontrivially periodic with period a for example, a sinusoid $\varepsilon(x) = \vec{\varepsilon} \cdot [1 + \Delta \cdot \cos(2\pi x/a)]$, or a square wave as in the inset of **Fig. 2(right)**. In the presence of such an oscillating 'potential', the accidental degeneracy between $e(x)$ and $o(x)$ is broken: supposing $\Delta > 0$, then the field $e(x)$ is more concentrated in the higher-ε regions than $o(x)$ and so lies at a lower frequency. This opposite shifting of the bands away from the mid-gap frequency $\omega \cong c\pi/a\sqrt{\vec{\varepsilon}}$ creates a bandgap, as depicted in **Fig. 2(right)**. (In fact, from the perturbation theory described subsequently, one can show that for $\Delta \ll 1$ the bandgap, as a fraction of mid-gap frequency, is $\Delta\omega/\omega \cong \Delta/2$.) By the same arguments, it follows that any

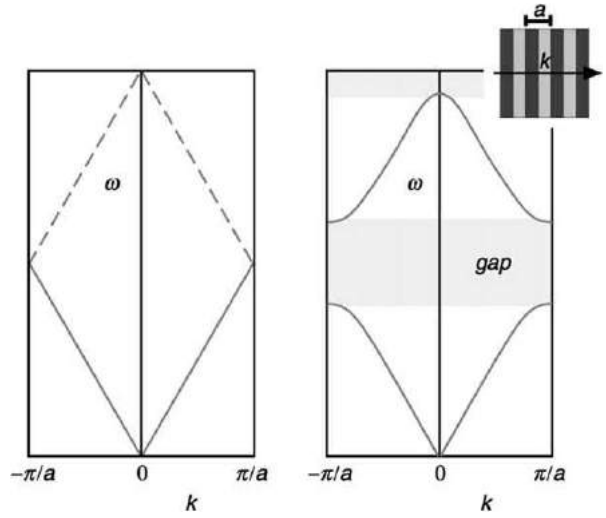


Fig. 2 Left: Dispersion relation (band diagram), frequency ω versus wavenumber k of a uniform one-dimensional medium, where the dashed lines show the 'folding' effect of applying Bloch's theorem with an artificial periodicity a. Right: Schematic effect on the bands of a physical periodic dielectric variation (inset), where a gap has been opened by splitting the degeneracy at the $k = \pm\pi/a$ Brillouin-zone boundaries (as well as a higher-order gap at $k=0$).

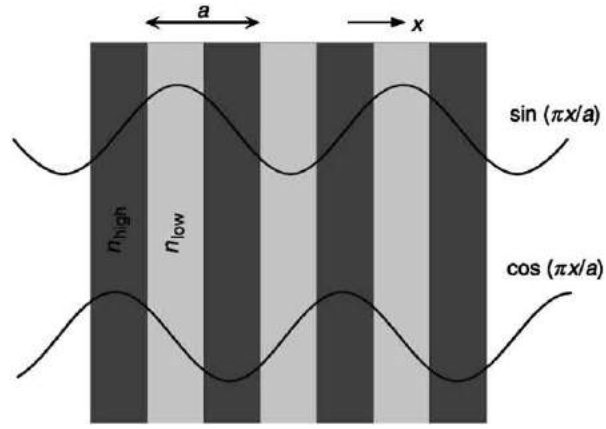


Fig. 3 Schematic origin of the band gap in one dimension. The degenerate $ka = \pi/a$ planewaves of a uniform medium are split into $\cos(\pi x/a)$ and $\sin(\pi x/a)$ standing waves by a dielectric periodicity, forming the lower and upper edges of the bandgap, respectively – the former has electric-field peaks in the high dielectric (n_{high}) and so will lie at a lower frequency than the latter (which peaks in the low dielectric).

periodic dielectric variation in one dimension will lead to a bandgap, albeit a small gap for a small variation; a similar result was identified by Lord Rayleigh in 1887.

More generally, it follows immediately from the properties of Hermitian eigensystems that the eigenvalues minimize a variational problem:

$$\omega^2_{n,\vec{k}} = \min_{\vec{E}_{n,\vec{k}}} \frac{\int |\left(\vec{\nabla} + i\vec{k}\right) \times \vec{E}_{n,\vec{k}}|^2}{\int \varepsilon |\vec{E}_{n,\vec{k}}|^2} c^2 \tag{3}$$

in terms of the periodic electric field envelope $\vec{E}_{n,\vec{k}}$ where the numerator minimizes the 'kinetic energy' and the denominator minimizes the 'potential energy' Here, the $n>1$ bands are additionally constrained to be orthogonal to the lower bands:

$$\int \vec{H}^*_{m,\vec{k}} \cdot \vec{H}_{n,\vec{k}} = \int \varepsilon \vec{E}^*_{m,\vec{k}} \cdot \vec{E}_{n,\vec{k}} = 0 \tag{4}$$

for $m<n$. Thus, at each $\vec{k}$ there will be a gap between the lower 'dielectric' bands concentrated in the high dielectric (low potential) and the upper 'air' bands that are less concentrated in the high dielectric: the air bands are forced out by the orthogonality condition, or otherwise must have fast oscillations that increase their kinetic energy. (The dielectric/air bands are analogous to the valence/conduction bands in a semiconductor.)

In order for a complete bandgap to arise in two or three dimensions, two additional hurdles must be overcome. First, although in each symmetry direction of the crystal (and each $\vec{k}$ point) there will be a bandgap by the one-dimensional argument, these bandgaps will not necessarily overlap in frequency (or even lie between the same bands). In order that they overlap, the gaps must be sufficiently large, which implies a minimum ε contrast (typically at least 4/1 in 3d). Since the 1d mid-gap frequency $\sim c\pi/a\sqrt{\varepsilon}$ varies inversely with the period a it is also helpful if the periodicity is nearly the same in different directions – thus, the largest gaps typically arise for hexagonal lattices in 2d and fcc lattices in 3d, which have the most nearly circular/spherical Brillouin zones. Second, one must take into account the vectorial boundary conditions on the electric field: moving across a dielectric boundary from ε to some $\varepsilon' < \varepsilon$, the inverse 'potential' $\varepsilon|\vec{E}|^2$ will decrease discontinuously if $\vec{E}$ is parallel to the interface ($\vec{E}_{\|}$ is continuous) and will increase discontinuously if $\vec{E}$ is perpendicular to the interface ($\varepsilon\vec{E}_{\perp}$ is continuous). This means that, whenever the electric field lines cross a dielectric boundary, it is much harder to strongly contain the field energy within the high dielectric, and the converse is true when the field lines are parallel to a boundary. Thus, in order to obtain a large bandgap, a dielectric structure should consist of thin, continuous veins/membranes along which the electric field lines can run – this way, the lowest band(s) can be strongly confined, while the upper bands are forced to a much higher frequency because the thin veins cannot support multiple modes (except for two orthogonal polarizations). The veins must also run in all directions, so that this confinement can occur for all $\vec{k}$ and polarizations, necessitating a complex topology in the crystal.

Ultimately, however, in two or three dimensions there are only rules of thumb for the existence of a bandgap in a periodic structure, since no rigorous criteria have yet been determined. This made the design of 3d photonic crystals a trial and error process, with the first example by Ho *et al.* of a complete 3d gap coming three years after the initial 1987 concept. As is discussed by the final section below, a small number of families of 3d photonic crystals have since been identified, with many variations thereof explored for fabrication.

Computational Techniques

Because photonic crystals are generally complex, high index-contrast, two- and three-dimensional vectorial systems, numerical computations are a crucial part of most theoretical analyses. Such computations typically fall into three categories: time-domain 'numerical experiments' that model the time-evolution of the fields with arbitrary starting conditions in a discretized system (e.g., finite-difference); definite-frequency transfer matrices wherein the scattering matrices are computed in some basis to extract transmission/reflection through the structure; and frequency-domain methods to directly extract the Bloch fields and frequencies by diagonalizing the eigenoperator. The first two categories intuitively correspond to directly measurable quantities such as transmission (although they can also be used to compute e.g., eigenvalues), whereas the third is more abstract, yielding the band diagrams that provide a guide to interpretation of measurements as well as a starting-point for device design and semi-analytical methods. Moreover, several band diagrams are included in the following sections, and so the frequency-domain method used to compute them is briefly outlined here.

Any frequency–domain method begins by expanding the fields in some complete basis, $\vec{H}_{\vec{k}}(\vec{x}) = \sum_n h_n \vec{b}_n(\vec{x})$, transforming the partial differential **Eq. (2)** into a discrete matrix eigenvalue problem for the coefficients h_n. Truncating the basis to N elements leads to $N \times N$ matrices, which could be diagonalized in $O(N^3)$ time by standard methods. This is impractical for large 3d systems, however, and is also unnecessary – typically, one only wants the few lowest eigenfrequencies, in which case one can use iterative eigensolver methods requiring only $\sim O(N)$ time. Perhaps the simplest such method is based directly on the variational theorem **Eq. (3)**: given some starting coefficients h_n one iteratively minimizes the variational 'Rayleigh' quotient using e.g., preconditioned conjugate–gradient descent. This yields the lowest band's eigenvalue and field, and upper bands are found by the same minimization while orthogonalizing against the lower bands ('deflation'). There is one additional difficulty, however, and that is that one must at the same time enforce the $(\vec{\nabla} + i\vec{k}) \cdot \vec{H}_{\vec{k}} = 0$ transversality constraint, which is nontrivial in three dimensions. The simplest way to maintain this constraint is to employ a basis that is already transverse, for example planewaves $\vec{h}_{\vec{G}} e^{i\vec{G}\cdot\vec{x}}$ with transverse amplitudes $\vec{h}_{\vec{G}} \cdot (\vec{G} + \vec{k}) = 0$. (In such a planewave basis, the action of the eigen-operator can be computed via a fast Fourier transform in $O(N \log N)$ time.)

Semi-analytical Methods: Perturbation Theory

As in quantum mechanics, the eigenstates can be the starting point for many analytical and semi-analytical studies. One common technique is perturbation theory, applied to small deviations from an ideal system – closely related to the variational **Eq. (3)**, perturbation theory can be exploited to consider effects such as nonlinearities, material absorption, fabrication disorder, and external tunability. Not only is perturbation theory useful in its own right, but it also illustrates both old and peculiarly new features that arise in such analyses of electromagnetism compared to scalar problems such as quantum mechanics.

Given an unperturbed eigenfield $\vec{E}_{n,\vec{k}}$ for a structure ε the lowest-order correction $\Delta\omega_n^{(1)}$ to the eigenfrequency from a small perturbation $\Delta\varepsilon$ is given by

$$\Delta\omega_{(1)}^{n} = -\frac{\omega_n(\vec{k})}{2} \frac{\int \Delta\varepsilon |\vec{E}_{n,\vec{k}}|^2}{\int \varepsilon |\vec{E}_{n,\vec{k}}|^2} \tag{5}$$

where the integral is over the primitive cell of the lattice. A Kerr nonlinearity would give $\Delta\varepsilon \sim |\vec{E}|^2$, material absorption would produce an imaginary frequency correction (decay coefficient) from a small imaginary $\Delta\varepsilon$ and so on. Similarly, one can compute the shift in frequency from a small $\Delta\vec{k}$ in order to determine the group velocity $d\omega/dk$; this variation of perturbation theory is also called $k \cdot p$ theory. All such first-order perturbation corrections are well known from quantum mechanics, and in the limit of infinitesimal perturbations give the exact Hellman–Feynman expression for the derivative of the eigenvalue. However, in the limit where $\Delta\varepsilon$ is a small shift Δh of a dielectric boundary between some ε_1 and ε_2 an important class of geometric perturbation, **Eq. (5)** gives a surface integral of $|\vec{E}|^2$ on the interface, but this is ill-defined because the field there is discontinous. The proper derivation of perturbation theory in the face of such discontinuity requires a more careful limiting process from an anisotropically smoothed system, yielding the surface integral:

$$\Delta\omega^n_{(1)} = -\frac{\omega_n(\vec{k})}{2} \frac{\int\int \Delta h \left(\Delta\varepsilon_{12} |\vec{E}_{||}|^2 - \Delta\varepsilon^{12}_{-1} |D_\perp|^2 \right)}{\int \varepsilon |\vec{E}_{n,\vec{k}}|^2}$$

where $\Delta\varepsilon_{12} = \varepsilon_1 - \varepsilon_2, \Delta\varepsilon^{-1}_{12} = \varepsilon^{-1}_1 - \varepsilon^{-1}_2$, and $\vec{E}_{||}/D_\perp$ denotes the (continuous) interface parallel/perpendicular components of the unperturbed electric/displacement eigenfield. A similar expression is required in high index-contrast systems to employ, e.g., coupled-mode theory for slowly-varying waveguides or Green's functions for interface roughness.

Standard perturbation–theory techniques also provide expressions for higher-order corrections to the eigenvalue and eigenfield, based on an expansion in the basis of the unperturbed eigenfields. This approach, however, runs into immediate difficulty because the eigenfields are also subject to the transversality constraint, $(\vec{\nabla} + i\vec{k}) \cdot \varepsilon \vec{E}_{\vec{k}} = 0$ and this constraint varies with ε and $\vec{k}$ – the eigenfields $\vec{E}_{\vec{k}}$ are not a complete basis for the eigenfields constrained at a different ε or $\vec{k}$. For ε perturbations, this problem can be eliminated by using the $\vec{H}$ or $\vec{D}$ eigenproblems, whose constraints are independent of ε. For $\vec{k}$ perturbations, one can employ a transformation by Sipe to derive a corrected higher-order perturbation theory (for e.g., the group-velocity dispersion), based on the fact that all of the non-transverse fields lie at $\omega = 0$. Such completeness issues also arise applying the variational **Eq. (3)**, as was noted in the previous section: in order for a useful variational bound to apply, one must operate in the constrained (transverse) subspace.

Two-Dimensional Photonic Crystals

After the identification of one-dimensional bandgaps, it took a full century to add a second dimension, and three years to add the third. It should therefore come as no surprise that 2d systems exhibit most of the important characteristics of photonic crystals, from nontrivial Brillouin zones to topological sensitivity to a minimum index contrast, and can also be used to demonstrate most proposed photonic–crystal devices. The key to understanding photonic crystals in two dimensions is to realize that the fields in 2d can be divided into two polarizations by symmetry: *TM* (transverse magnetic), in which the magnetic field is in the (xy) plane and the electric field is perpendicular (z); and *TE* (transverse electric), in which the electric field is in the plane and the magnetic field is perpendicular.

Corresponding to the polarizations, there are two basic topologies for 2d photonic crystals, as depicted in **Fig. 4(top)**: high index rods surrounded by low index (top) and low-index holes in high index (bottom). Here, a hexagonal lattice is used because, as noted earlier, it gives the largest gaps. Recall that a photonic band gap requires that the electric field lines run along thin veins: thus, the rods are best suited to TM light ($\vec{E}$ parallel to the rods), and the holes are best suited to TE light ($\vec{E}$ running around the holes). This preference is reflected in the band diagrams, shown in **Fig. 4**, in which the rods/holes (top/bottom) have a strong TM/TE band gap. For these diagrams, the rod/hole radius is chosen to be $0.2a/0.3a$, where a is the lattice constant (the nearest-neighbor periodicity) and the high/low ε is 12/1. The TM/TE bandgaps are then 47%/28% as a fraction of mid-gap frequency, but these bandgaps require a minimum ε contrast of 1.7/1 and 1.9/1, respectively. Moreover, it is conventional to give the frequencies ω in units of $2\pi c/a$, which is equivalent to a/λ(λ being the vacuum wavelength) – Maxwell's equations are scale-invariant, and the same solutions can be applied to any wavelength simply by choosing the appropriate a. For example, the TM mid-gap ω in these units is 0.36, so if one wanted this to correspond to $\lambda = 1.55$ μm one would use $a = 0.36 \cdot 1.55\ \mu\text{m} = 0.56\ \mu\text{m}$.

The Brillouin zone (a hexagon) is shown at left-center, with the irreducible Brillouin zone shaded (following the sixfold symmetry of the crystal); the corners (high symmetry points) of this zone are given canonical names, where Γ always denotes the origin $\vec{k} = 0$, K is the nearest-neighbor direction, and M is the next-nearest-neighbor direction. The Brillouin zone is a two-dimensional region of wavevectors, so the bands $\omega_n(\vec{k})$ are actually surfaces, but in practice the band extrema almost always occur along the boundaries of the irreducible zone (i.e., the high-symmetry directions). So, it is conventional to plot the bands only along these zone boundaries in order to identify the bandgap, as is done in **Fig. 4**.

Actually, the hole lattice can display not only a TE gap, but a complete photonic bandgap (for both polarizations) if the holes are sufficiently large (nearly touching). In this case, the thin veins between nearest-neighbor holes induce a TE gap, while the interstices between triplets of holes form 'rod-like' regions that support a TM gap overlapping the TE gap.

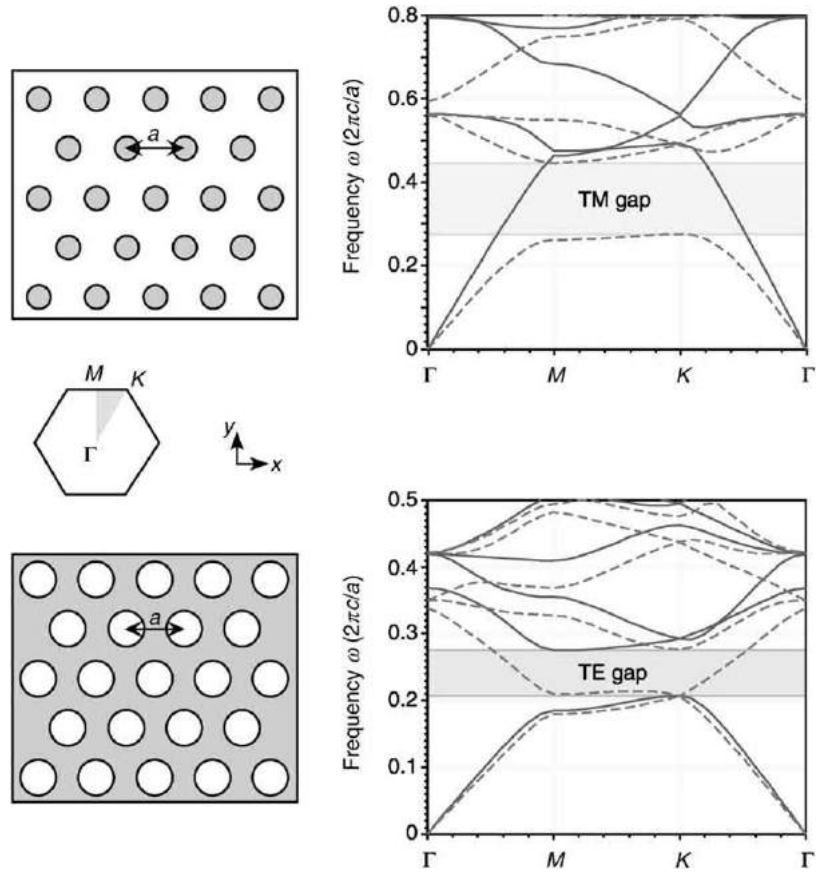


Fig. 4 Band diagrams and photonic band gaps for hexagonal lattices of high dielectric rods ($\varepsilon = 12, r = 0.2a$) in air (top), and air holes ($r = 0.3a$) in dielectric (bottom), where a is the center–center periodicity. The frequencies are plotted around the boundary of the irreducible Brillouin zone (shaded triangle, left center), with solid/dashed lines denoting TE/TM polarization (electric field parallel/perpendicular to plane of periodicity). The rods/holes have a gap in the TM/TE bands.

Photonic-Crystal Slabs

In order to realize 2d photonic-crystal phenomena in three dimensions, the most straightforward design is to simply fabricate a 2d-periodic crystal with a finite height: a photonic-crystal slab, as depicted in Fig. 5. Such a structure can confine light vertically within the slab via index guiding, a generalization of total internal reflection – this mechanism is the source of several new tradeoffs and behaviors of slab systems compared to their 2d analogs.

The key to index guiding is the fact that the 2d periodicity implies that the 2d Bloch wavevector $\vec{k}_{||}$ is a conserved quantity, so the projected band structure – all states in the bulk substrate/superstrate (the uniform regions far below/above the slab) versus their in-plane wavevector component (projected wavevector) – creates a map of which states can radiate vertically. If the slab is suspended in air, for example, then the eigensolutions of the bulk air are $\omega = c\sqrt{|\vec{k}_{||}|^2 + k_\perp^2}$ which when plotted versus $\vec{k}_{||}$ forms the continuous light cone $\omega \geq c|\vec{k}_{||}|$ shown as a shaded region in Fig. 5. Because the slab has a higher $\varepsilon(12)$ than the air (1), and frequency goes as $1/\sqrt{\varepsilon}$ discrete guided bands are 'pulled down' in frequency from this continuum – these bands, lying beneath the light cone, cannot couple to any vertically radiating mode by the conservation law and so are confined to the slab (exponentially decaying away from it). If the horizontal mid-plane of the slab is a mirror symmetry plane, then just as there were TM and TE states in 2d, here there are two categories of modes: even (TE-like) and odd (TM-like) modes under reflections through the mirror plane (which are purely TE/TM in the mirror plane itself). Because the slab here is based on the 2d hole crystal, which had a TE gap, here there is a 26% 'bandgap' in the even modes: a range of frequencies in which there are no guided modes. It is not a complete photonic bandgap, not only because of the odd modes, but also because there are radiating (light cone) modes at every ω. The presence of these radiating modes means that if all in-plane translational symmetry is broken by a localized change in the structure, say a waveguide bend or a resonant cavity, then vertical radiation losses are inevitable; there are various strategies to minimize the losses to tolerable levels, however. On the other hand, if only one direction of translational symmetry is broken, as in a linear-defect waveguide, ideally lossless guiding can be maintained.

Photonic-crystal slabs have two new critical parameters that influence the existence of a gap. First, it must have vertical mirror symmetry in order that the gaps in the even and odd modes be treatable separately – such symmetry is broken by the presence of a

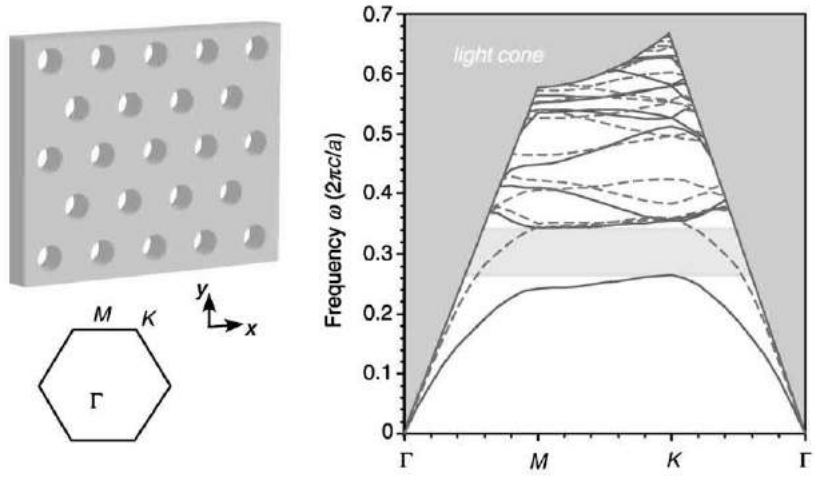


Fig. 5 Projected band diagram for a finite-thickness (0.5*a*) slab of air holes in dielectric (cross section as in Fig. 4 bottom), with the irreducible Brillouin zone at lower left. The shaded region is the light cone: the projection of all states that can radiate in the air. Solid/dashed lines denote guided modes (confined to the slab) that are even/odd with respect to the horizontal mirror plane of the slab, whose polarization is TE-like/TM-like, respectively. There is a 'bandgap' (region without guided modes) in the TE-like guided modes only.

substrate that is not the mirror image of the superstrate, but in practice the symmetry breaking can be weak if the index contrast is sufficiently high (so that the modes are strongly confined in the slab). Second, the height of the slab must not be too small (or the modes will be weakly confined) or too large (or higher-order modes will fill the gap); the optimum height is around half a wavelength λ/n_{eff} (relative to an average/effective index n_{eff} that depends on the polarization). In Fig. 5, a height of $0.5a$ is used, which is near the optimum (with holes of radius $0.3a$ and $\varepsilon = 12$ as in the previous section).

Three-Dimensional Photonic Crystals

Photonic-crystal slabs are one way of realizing 2d photonic-crystal effects in three dimensions; an example of another way, lifting the sacrifices imposed by the light cone, is depicted in Fig. 6. This is a 3d-periodic crystal, formed by an alternating hole-slab/rod-slab sequence in an ABCABC stacking of bilayers – equivalently, it is an fcc (face-centered cubic) lattice of air cylinders in dielectric, stacked and oriented in the 111 direction, where each overlapping layer of cylinders forms a rod/hole bilayer simultaneously. Its band diagram is shown in Fig. 6 along the boundaries of its irreducible Brillouin zone (from a truncated octahedron, inset), and this structure has a $>20\%$ complete gap ($\Delta\omega$ as a fraction of mid-gap frequency) for $\varepsilon = 12/1$, forbidding light propagation for all wavevectors (directions) and all polarizations. Not only can this crystal confine light perfectly in 3d, but because its layers resemble 2d rod/hole crystals, it turns out that the confined modes created by defects in these layers strongly resemble the TM/TE states created by corresponding defects in two dimensions. One can therefore use this crystal to directly transfer designs from two to three dimensions while retaining omnidirectional confinement. Its fabrication, of course, is more complex than that of photonic-crystal slabs (with a minimum ε contrast of 4/1), but this and other 3d photonic crystal structures have been constructed even at micron (infrared) lengthscales, as described below.

There are three general dielectric topologies that have been identified to support complete 3d gaps for $\varepsilon = 12/1$ (e.g., Si:air) contrast: diamond-like arrangements of high dielectric 'atoms' surrounded by low dielectric, which can lead to $>20\%$ gaps between the 2nd and 3rd bands; fcc 'inverse opal' lattices of nearly-touching low dielectric spheres (or similar) surrounded by high dielectric, giving gaps around 10% between the 9th and 10th bands; and cubic 'scaffold' lattices of rods along the cube edges, giving $\sim 7\%$ gaps between the 2nd and 3rd bands. It is notable that the first two topologies correspond to fcc lattices, which have the most nearly spherical Brillouin zones in accordance with the rules of thumb given above. Many variations on these topologies continue to be proposed – for example, the structure of Fig. 6 is diamond/graphite-like – mainly in conjunction with different fabrication strategies, such as the following three successful approaches. First, layer-by-layer fabrication, in which individual crystal layers (typically of constant cross-section) are deposited one-by-one and etched with a 2d pattern via standard lithographic methods (giving fine control over placement of defects, etc.); Fig. 6 can be constructed in this fashion (as well as other diamond-like structures with large gaps). Second, colloidal self-assembly, in which small dielectric spheres in a fluid automatically arrange themselves into close-packed (fcc) crystals by surface forces – these crystals can be back-filled with a high-index material, out of which the original spheres are dissolved in order to form inverse–opal crystals with a complete gap. Third, holographic lithography, in which a variety of 3d crystals can be formed by an interference pattern of four laser beams to harden a light-sensitive resin (which is then back-filled and dissolved, as with colloids, to achieve the requisite index contrast). The second and third techniques are notable for their ability to construct large-scale 3d crystals (thousands of periods) in a short time.

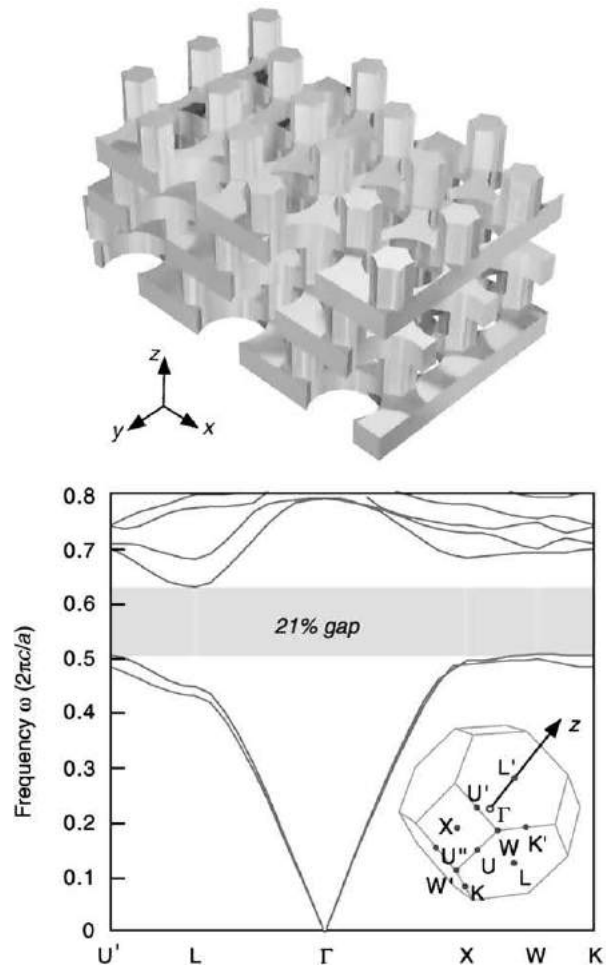


Fig. 6 Band diagram (bottom) for 3d-periodic photonic crystal (top) consisting of an alternating stack of rod and hole 2d-periodic slabs (similar to **Fig. 4**), with the corners of the irreducible Brillouin zone labeled in the inset. This structure exhibits a $\Delta\omega/\omega_{\text{midgap}} = 21\%$ omnidirectional bandgap.

Further Reading

Ashcroft, N.W., Mermin, N.D., 1976. Solid State Physics. Philadelphia: Holt Saunders.
Cohen-Tannoudji, C., Din, B., Laloë, F., 1977. Quantum Mechanics. Paris: Hermann.
Joannopoulos, J.D., Meade, R.D., Winn, J.N., 1995. Photonic Crystals: Molding the Flow of Light. Princeton: Princeton University Press.
Johnson, S.G., Joannopoulos, J.D., . Photonic Crystals: The Road from Theory to Practice. Boston: Kluwer.
Johnson, S.G., Ibanescu, M., Skorobogatiy, M., Weisberg, O., Joannopoulos, J.D., Fink, Y., 2002. Perturbation theory for Maxwell's equations with shifting material boundaries. Physical Review E 65, 066611.
Sakoda, K., 2001. Optical Properties of Photonic Crystals. Berlin: Springer.
Sipe, J.E., 2000. Vector $k \cdot p$ approach for photonic band structures. Physical Review E 62, 5672–5677.

Nonlinear Optics in Photonic Crystal Fibers

JE Sharping, Cornell University, Ithaca, NY, USA
P Kumar, Northwestern University, Evanston, IL, USA

Introduction

Photonic crystal fibers (PCFs) are very similar to normal optical fibers in that they consist of a core surrounded by cladding, such that light is guided within the core of the fiber. The primary difference between PCF and standard optical fibers is that PCFs feature an air–silica cross-section, whereas standard optical fibers have an all-glass cross-section. An electron micrograph of a typical PCF is shown in **Fig. 1**. The air holes extend along the axis of the fiber for its entire length and the core of the fiber is formed by a defect, or missing hole, in the periodic structure. The core is formed of solid glass, whose refractive index is that of pure silica (or whatever other glass is chosen), and the cladding is formed by the air–glass mixture, whose effective refractive index depends on the ratio of air-to-glass, also known as the air-fill fraction, that comprises the structure. The resulting effective-index of the cladding will be lower compared with that of the core and, as such, will provide the refractive index variation necessary to support total internal reflection at the core-cladding boundary, and guide light in a manner similar to that of standard optical fibers. The fiber design (i.e., size, shape, and the air-fill fraction) dictates solutions to Maxwell's equations for light propagating within the fiber. Valid solutions are referred to as 'modes' which propagate along the fiber in a known manner, and have a well-defined shape in the transverse direction (i.e., they have a well-defined transverse mode structure).

Nonlinear-optical effects in fibers result from the interaction of optical fields with the glass via the $\chi^{(3)}$,or Kerr nonlinearity. The phenomenon of nonlinear refractive index is a manifestation of a light–material interaction mediated by $\chi^{(3)}$. The magnitudes of the components of the third-order susceptibility tensor in glass, $\chi^{(3)}$, are generally quite small compared with the analogous second-order $\chi^{(2)}$ terms for materials exhibiting such nonlinearities (e.g., lithium niobate, beta-barium borate (BBO), etc.). The relatively small $\chi^{(3)}$ nonlinearity in optical fibers makes them ideal for wavelength-division multiplexed optical communication where light propagation subject to a minimum of nonlinear effects is critical. Nonlinearity does, however, eventually become an issue in wavelength-division multiplexed systems as the launched optical power increases and as the channel spacing decreases. On the other hand, one can utilize nonlinear-optical effects in soliton communication systems and to build useful photonic devices. Despite the weak $\chi^{(3)}$ nonlinearity, the net nonlinear-optical effect in fibers can be large due to the ability to tightly confine intense fields within the core of an optical fiber and maintain the interaction over a long distance as the guided fields propagate through the fiber.

The study of nonlinear-fiber optics has benefited from dramatic improvements in optical fiber and fiber-optic device fabrication. The importance of understanding nonlinear-fiber optics is driven by the need to develop fiber-integrated devices, and also by the need to understand and mitigate the problems that these nonlinearities cause in optical communication systems.

This section introduces the unique linear- and nonlinear-optical properties of PCFs in order to understand the reasons why nonlinear-optical effects are often enhanced in such fibers. These discussions pertain to PCFs which are 'highly nonlinear'.

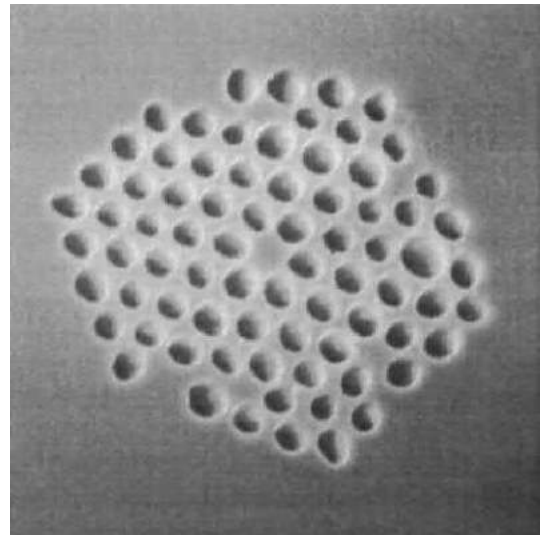


Fig. 1 An electron micrograph showing the periodic microstructure of a typical PCF. The core is formed by the 'missing hole' in the center of the microstructure. Reproduced with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800nm. *Optical Letters* 25: 25–27. ©2000 Optical Society of America, courtesy of OFS.

It is essential to clarify that 'highly nonlinear' in this context does not mean that the $\chi^{(3)}$ is any larger than that of standard telecommunication fibers, rather that the effect of this nonlinearity is enhanced due to the fiber's very small core.

PCF Properties

Photonic crystal fibers feature a variety of interesting properties. From the standpoint of nonlinear-fiber optics there are four very useful fundamental properties of PCFs:

- a mechanically robust optical fiber can be fabricated with an extremely small core (a few μm^2);
- a fiber can be made to guide in a single transverse mode over an extremely broad wavelength range (370 nm–1600 nm);
- there are new degrees of freedom that allow one to manipulate the fiber's group-velocity dispersion (GVD) properties; and
- many, but not all, PCFs are polarization maintaining as a result of form birefringence present in the core.

The fact that small-core PCFs can be fabricated is clear from **Fig. 1** by taking note of the fact that the center defect region which comprises the core is about 1.7 μm in diameter. Photonic crystal fibers with even smaller cores have been fabricated.

Transverse Mode Structure

A widely accepted model used to describe the transverse modal behavior of PCFs is called the effective-index model. The effective-index model can be used to understand why some PCFs are 'endlessly single mode', meaning that the fiber guides in a single transverse mode over an exceptionally wide wavelength range (370 nm–1600 nm). In the effective-index model, the refractive index of the core, $n_{co}(\lambda)$, is that of glass, and the refractive index of the cladding, $n_{eff}(\lambda)$, assumes a value in between that of glass and air. In the context of PCFs, one makes a modification to the standard expression describing single-mode behavior in step-index fibers:

$$V_{\text{eff}} = \frac{2\pi\Lambda}{\lambda}\sqrt{n_{\text{co}}(\lambda)^2 - n_{\text{eff}}(\lambda)^2} < V_{\text{cufoff}} \tag{1}$$

where Λ is the spacing between air holes, λ is the wavelength of light, and V_{cutoff} is the cutoff condition for the PCF. A similar expression for the V parameter is commonly used to understand the modal behavior of standard fibers where the larger the V is, the more transverse modes are supported within the fiber. In standard fibers the cutoff condition below which only a single mode can propagate within the core of a fiber is given by $V_{\text{cutoff}} < 2.405$. In the case of PCFs, a numerical method should be used to determine V_{cutoff}. Mechanically robust PCFs can be fabricated where the dispersion in $n_{\text{eff}}(\lambda)$ (i.e., the variation of n_{eff} with λ) offsets dispersion in $n_{\text{co}}(\lambda)$ and compensates for the $2\pi\Lambda/\lambda$ coefficient in **Eq. (1)**. Therefore, the light within the fiber propagates in a single, Gaussian-like mode because for all wavelengths $V_{\text{eff}} < V_{\text{cutoff}}$. A graph of the variation of V_{eff} with Λ/λ is shown in **Fig. 2**, where d represents the size of an air hole.

Conceptually, the effective index model can be understood by noting that at short wavelengths the mode field is confined well within the all-silica core, but as λ increases the mode field extends further into the air–glass cladding and V_{eff} and $n_{\text{eff}}(\lambda)$ both decrease.

Dispersion in PCF

Other critical differences between PCFs and standard optical fibers lie in the dispersion properties. When light propagates through a fiber its behavior depends on the light's optical frequency:

$$E(t,z) = A(t,z)e^{i[\omega t - \beta(\omega)z]} \tag{2}$$

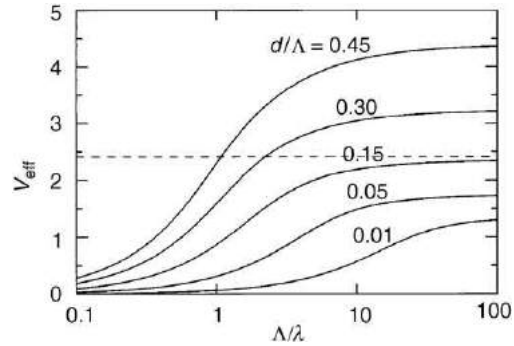


Fig. 2 Variation of V_{eff} for different relative hole diameters d/Λ.The calculation assumes a fiber with an air–glass cross-section where the refractive index of air and glass was taken to be 1 and 1.45, respectively. The dashed line marks $V_{\text{eff}} = 2.405$, the cutoff value for a step-index fiber. Reproduced with permission from Birks TA, Knight JC and Russell PStJ (1997) Endlessly single-mode photonic crystal fiber. *Optical Letters* 22: 961–963. ©1997 Optical Society of America.

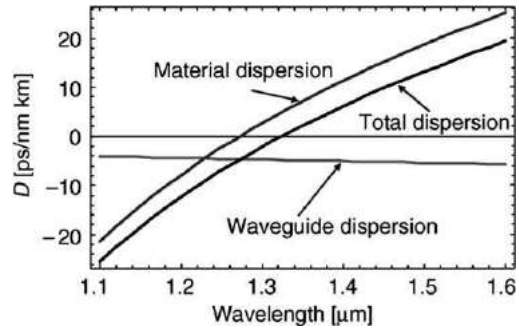


Fig. 3 Plots of the theoretical dispersion coefficient, D as a function of wavelength for a standard optical communication fiber.

Eq. (2) describes the mode as it propagates through the fiber. It is decomposed into a slowly varying envelope, $A(t,z)$, and a rapidly varying exponential component where ω is the frequency of the mode, t is time, z is the position along the length of the fiber, and $\beta(\omega)$ is called the mode-propagation constant. The general term used in describing the frequency dependence of β is chromatic dispersion, which includes contributions from the material as well as the waveguide. Other types of dispersion present in optical fibers include multi-modal (arising from multiple guided transverse modes) and polarization-mode dispersion.

One way to understand the chromatic dispersion of a mode propagating through an optical fiber is to study the Taylor series expansion of the mode-propagation constant, β, about the center frequency of the field, ω_0:

$$\beta(\omega) = \beta_0 + \beta_1(\omega - \omega_0) + \frac{1}{2}\beta_2(\omega - \omega_0)^2 + \cdots \tag{3}$$

where $\beta_i = \mathrm{d}^i\beta/\mathrm{d}\omega^i$. The physical significance of the various β_i in Eq. (3) are as follows: the phase-fronts of the electric field move at a speed given by $\omega/\beta_0 = v_p$, and the envelope, $A(t,z)$, moves at a group velocity given by $1/\beta_1$. A GVD term, which governs temporal spreading of the envelope, is given by β_2. Higher-order β terms are usually negligible for propagation of pulses of ≥ 1 ps duration in optical fibers and are lumped into the category of 'higher-order chromatic dispersion'.

The notation 'β_2', as defined above, is often used in the literature with dimensions of ps^2/km. However, another expression is frequently used because of its direct relationship to measured quantities. It is straightforward to measure the relative delay, T, between two pulses having different center wavelengths. Choosing a particular wavelength as a reference, one can then measure relative delay as a function of an injected pulse's center wavelength. The first derivative with respect to λ of the relative delay curve gives the GVD according to

$$D = \frac{\partial\left(\frac{1}{v_g}\right)}{\partial\lambda} = \frac{1}{L}\frac{\mathrm{d}T}{\mathrm{d}\lambda} = -\frac{2\pi c}{\lambda^2}\beta_2 \tag{4}$$

where v_g is the group velocity, and L is the length of the fiber under test. The dimension commonly used for D is ps/(nm km).

Chromatic dispersion in single-mode optical fibers results from two different wavelength-dependent fiber parameters. The medium itself, glass in this case, has a wavelength-dependent refractive index. This 'material' contribution has the same magnitude regardless of the various parameters associated with the waveguide. A second contribution has to do with the design of the optical fiber. This 'waveguide' contribution to dispersion arises from the fact that the wavelength-dependent mode depends on the properties of the waveguide (i.e., the core size and refractive index contrast between the core and cladding). Empirical models can be used to describe the material, waveguide, and total GVD for standard communication fibers. Such a set of curves is given in Fig. 3 where it can be seen that it is possible to have positive, negative, or zero values for D. For historical reasons, the regions where D is negative (β_2 is positive) exhibit 'normal GVD', while those where D is positive (β_2 is negative), exhibit 'anomalous GVD'. The wavelength corresponding to $D=0$ is referred to as the zero-dispersion wavelength (λ_0), which for most silica glass fibers is about 1,300 nm.

In contrast with standard optical fibers, where the waveguide contribution to D is always less than zero, small-core PCFs can be fabricated where the waveguide contribution to GVD is positive and quite large. As such, in PCFs, λ_0 can be shifted to wavelengths shorter than the intrinsic dispersion zero of glass. Control over the GVD is essential for phase matching certain nonlinear-optical interactions involving light of different colors co-propagating within a fiber. Indeed, several exciting applications of nonlinear optics in PCF require a fiber with a $\lambda_0 \sim 800$ nm. The GVD is also of great importance when working with pulsed light in PCF, because GVD results in temporal pulse broadening. It also governs pulse temporal walkoff effects, limiting the effective interaction length between pulses of different colors. This new flexibility to manipulate the GVD curve, by varying waveguide design parameters, is a key advantage associated with using PCFs for nonlinear optics.

Birefringence in PCF

The core of an optical fiber often exhibits some amount of anisotropy. The core may be elliptical in form (shape) which leads to a phenomenon referred to as form birefringence. Since mode propagation depends on the fiber structure, a fiber with an elliptical

core will exhibit mode propagation that depends on the electric field's polarization with respect to the axes of the elliptical core. As a result of birefringence, the polarization of the mode varies as it propagates through the fiber (unless care is taken to align the polarization of the injected light with respect to a principal axis of the birefringence).

Polarization-maintaining (PM) fibers are designed to include birefringence in a particular axis of the fiber. By including a well-defined birefringence throughout the length of a fiber that is larger than that induced by external perturbations, fast and slow axes of the optical fiber are created for all guided wavelengths, giving two orthogonal 'polarization modes'. If light is injected into one of the polarization modes (i.e., with its linear polarization along one of the axes) it remains linearly polarized along that axis as it propagates along the fiber. The two polarization modes generally have different group velocities, so pulsed light in each mode will take a different amount of time to propagate through a given segment of fiber. Most PCFs exhibit strong birefringence due to a slightly elliptical core combined with a large core-cladding index difference, and so they behave similarly to PM fibers. Special care must be taken when working with PCF to be sure that the polarization of the light launched into the fiber is aligned with one of the birefringent axes.

In practice, there are a few other features of PCF that are of importance when discussing nonlinear-optic interactions:

- propagation losses are generally larger in PCFs than in standard optical fibers; and
- free-space coupling and splicing are difficult and can result in large coupling losses.

Nonlinear Phenomena

The basic principles determining nonlinear effects in PCFs are the same as those for standard optical fibers. It is the new flexibility in PCFs to obtain transverse-modal and GVD behavior different from that of standard optical fibers that makes PCFs truly interesting for nonlinear optics. The relevant nonlinear-optical effects are: self-phase modulation (SPM); cross-phase modulation (CPM); third-harmonic generation (3HG); four-wave mixing (FWM); Raman scattering; and Brillouin scattering.

Self-phase modulation (also known as the optical Kerr effect) refers to the self-induced phase shift experienced by an optical field as it propagates through a fiber. It becomes particularly important for the case of pulses of light propagating through optical fibers. In small core PCFs, SPM is enhanced due to the high-intensity light propagating within the core. Self-phase modulation can lead to substantial spectral broadening of pulsed light propagating along an optical fiber.

When a pulse of light experiences normal GVD (i.e., $D<0$) as it propagates, the longer-wavelength components travel faster than the shorter-wavelength components. Anomalous GVD (i.e., $D>0$) leads to the opposite, short-wavelength components traveling faster than the long-wavelength components. Group-velocity dispersion generally leads to temporal broadening of pulses as they propagate along a fiber. Under ideal conditions, however, SPM in combination with anomalous GVD, leads to pulses which propagate without any temporal or spectral broadening. These self-sustaining pulses are called 'optical solitons'.

When waves of light having different wavelengths co-propagate along a fiber, CPM can occur. It can be understood as a phase shift induced on one wave, due to the presence of the other wave. Cross-phase modulation also leads to spectral broadening and solitonic pulse propagation.

In 3HG and FWM, one or more photons are destroyed and others are created. In 3HG, three 'fundamental' photons are destroyed to create one with three times the energy of the fundamental photons. In FWM, two fundamental photons are destroyed while two others are created. While it is straightforward to conserve energy in 3HG and FWM, these interactions must be 'phase-matched', meaning that the interacting waves must be made to propagate in-phase over a meaningful length. Such phase-matching conditions need to be carefully considered when studying 3HG and FWM. Nevertheless, 3HG and FWM can be used to obtain frequency shifts and all-optical amplification. In comparison with other types of optical fibers, PCFs are particularly useful for 3HG and FWM applications. The small core of the PCF allows interactions to occur at much lower input powers, and the new flexibility associated with the GVD properties permits phase matching in cases which are not possible using standard optical fibers. Finally, the fiber's endlessly single-mode behavior permits very good transverse mode overlap between interacting waves having widely different center wavelengths.

Self-phase modulation, CPM, 3HG, and FWM are photon–photon interactions wherein no energy is exchanged with the medium itself. In contrast, Raman scattering and Brillouin scattering result from photon–phonon interactions. The differences between Raman and Brillouin scatterings lie in the energy of the phonons involved and the direction in which the interactions occur. Raman scattering is an interaction between a photon and an optical–phonon mode of the molecules making up the material. In the case of Raman scattering in glass, the energy shift, associated with molecular vibrational (Raman) modes, corresponds to frequencies of 1–12 THz. Raman scattering occurs in the forward and backward directions. With pulsed light, stimulated Raman scattering can occur when the lower-frequency spectrum of the pulse overlaps with the spectrum of the Raman resonances excited by the higher-frequency spectrum. When this happens, energy can be efficiently shifted in spectrum towards the peak of the Raman resonance. In the forward direction, this 'Raman self-frequency shift' builds up. Additionally, if the Raman self-frequency shift occurs in the presence of anomalous dispersion, a 'Raman soliton self-frequency shift' can result. As the injected power is increased, the spectral shift between the injected pulse and the resulting Raman soliton increases. The principal advantage associated with PCF is the ability to generate Raman solitons for a broader range of wavelengths than was previously possible.

For Brillouin scattering, interaction with the acoustical phonons results in frequency shifts of about 10 GHz and the interaction only occurs in the backward direction. Brillouin scattering is generally a nuisance in fiber-based devices, leading to intensity noise

and other problems. The interesting feature of Brillouin scattering in PCFs is that the threshold intensity where problems begin to occur is higher for PCFs than for standard optical fibers. The higher threshold permits further optimization of fiber-based devices wherein Brillouin scattering limits the performance.

Experiment Examples

In the following subsections, a selection of experiments, demonstrating a few of the relevant nonlinear-optical phenomena, are briefly described.

Supercontinuum Generation

One of the most exciting demonstrations of nonlinear optics in PCF is that of supercontinuum generation. In a typical experiment, 100 femtosecond pulses from a mode-locked Ti:Sapphire laser operating at a wavelength of 800 nm were injected into the PCF. As the injected power was increased, a broad continuum of spectrum was generated from wavelengths of 400 nm up to 1,600 nm. Typical data showing the input and output optical spectra are given in Fig. 4, where it can be seen that well over an octave of frequency spectrum is generated from an input pulse whose spectral width is only about 10 nm.

It is widely accepted that supercontinuum generation results from a combination of linear and nonlinear optical effects conspiring to generate the broad spectrum. As the pulse propagates through the PCF, SPM, FWM, and Raman scattering are all likely to occur with relative efficiencies depending heavily on the input pulses spectral and temporal characteristics, as well as the fibers properties. Indeed, efficient supercontinuum generation can occur in PCFs for pump wavelengths lying near the zero-dispersion point in the normal or anomalous dispersion regime of the PCF, and for fibers as short as a few millimeters in length.

Spectral broadening, due to SPM, is common in standard optical fibers, but what makes this particular experiment interesting is the remarkable width of spectrum generated with a short piece of fiber ($\sim$1 m) and with comparatively small optical powers.

Optical Switching in PCF

Cross-phase modulation can be used to build an all-optical switch. Such a switch can be implemented as a three-port device where the output port for a given optical bit (the signal pulse), is determined by the presence of an optical control (the pump pulse). Switching can then be achieved by dividing the signal pulse equally on two arms of an interferometer and injecting the strong pump pulse only on one arm. Because the pump pulse co-propagates with only one of the two signal pulses, there exists a CPM-induced phase difference $\varphi_{NL} = 2\gamma P_p L$ at the output of the interferometer, where P_p is the peak power of the pump pulse and L is the interaction length. By varying the intensity of the pump pulses one can vary the magnitude of this phase difference. If a π-phase shift is achieved, one can switch the interference from destructive to constructive, or vice-versa, thus realizing an all-optical switch.

Fig. 5 shows an experimental setup used to observe switching near 1,550 nm (a similar apparatus using bulk-optic rather than fiber-optic components can be used to conduct experiments near 780 nm). The pump and signal are synchronous few-ps-duration pulses with a tunable wavelength separation of 5–15 nm. The switching characteristics for this implementation are shown in Fig. 6. The apparatus has the advantage of requiring short fiber lengths, low switching powers, and allows switching of weak pulses. It demonstrates the feasibility of using nonlinear optics in PCF to perform essential functions in high-speed all-optical processing.

Parametric (Mixing) Processes

The first set of experiments with controlled FWM in a PCF achieved nondegenerate parametric gains over a 30 nm range of pump wavelengths near the λ_0 of the PCF. The experiments also confirmed the wavelength dependence of the GVD coefficient of the PCF

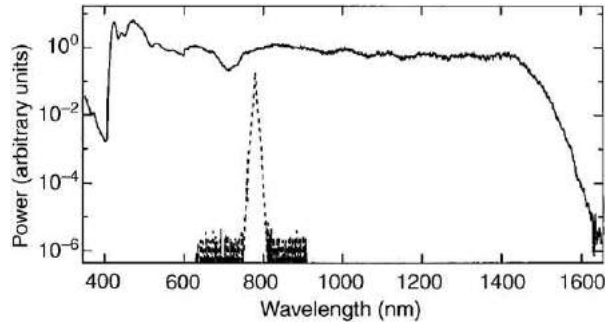


Fig. 4 In supercontinuum generation one observes a broad continuum generated after short pulses of light from a Ti:Sapphire laser propagate through a 75 cm section of PCF. The spectrum of the input pulse is shown as a dashed curve, while the output is a solid curve. Reproduced with permission from Ranka JK, Windeler RS and Strentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optical Letters* 25: 25–27. ©2000 Optical Society of America.

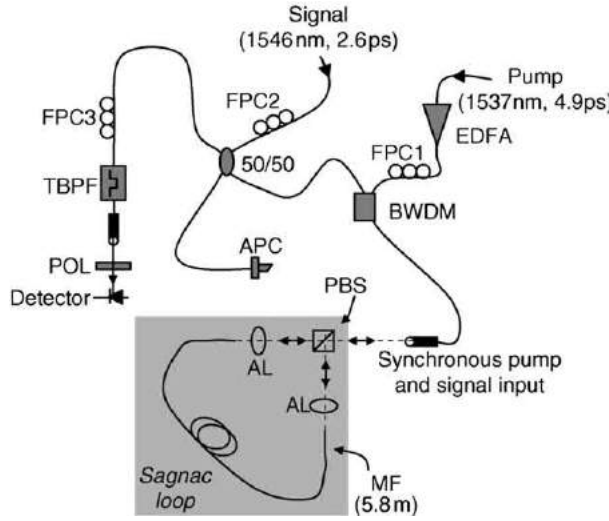


Fig. 5 Experimental setup used to demonstrate all-optical switching near 1,550 nm. (EDFA, erbium-doped fiber amplifier; BWDM, bandpass wavelength-division multiplexor; PBS, polarization beamsplitter; FPC, fiber polarization controller). Reproduced with permission from Sharping JE, Fiorentino M, Kumar P and Windeler RS (2002) All-optical switching based on cross-phase modulation in microstructure fiber. *IEEE Photonics Technology Letters* 14: 77–79. ©2002 IEEE.

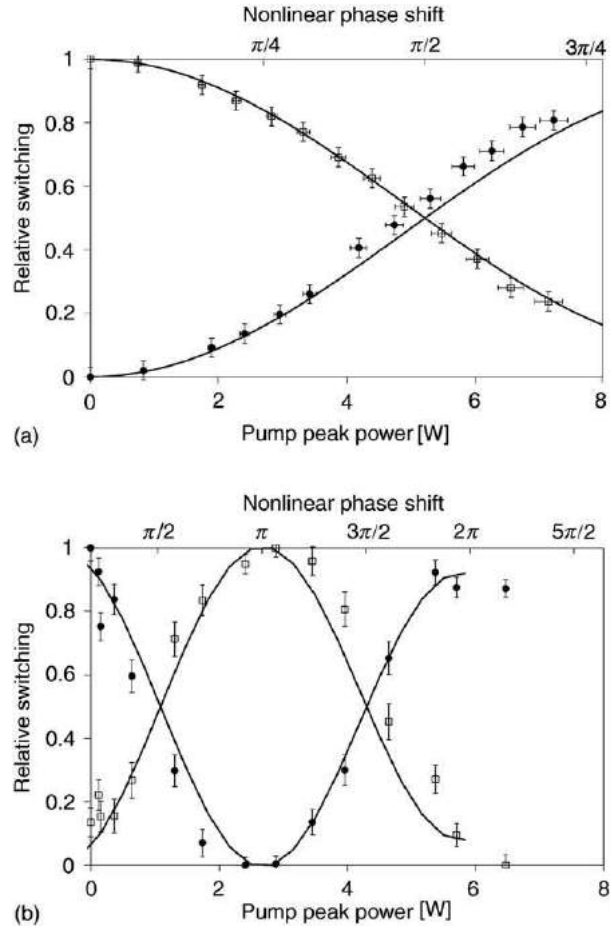


Fig. 6 Switching curves (open boxes and filled circles), showing the relative power measured in each port of the switch vs. the pump peak power, for experiments conducted near (a) 1,550 nm and (b) 780 nm. The curves accompanying the data are generated from numerical solutions of coupled wave equations for CPM. Reproduced with permission from Sharping JE, Fiorentino M, Kumar P and Windeler RS (2002) All-optical switching based on cross-phase modulation in microstructure fiber. *IEEE Photonics Technology Letters* 14: 77–79. ©2002 IEEE.

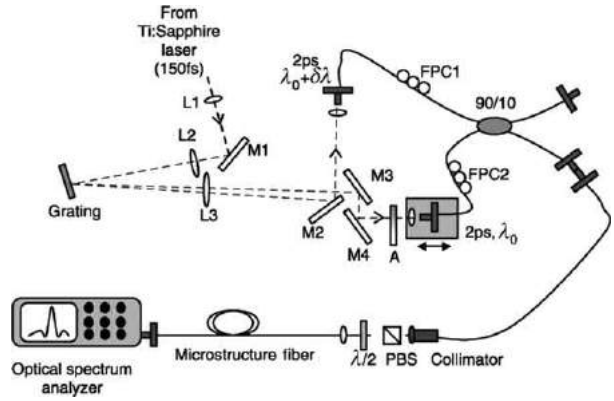


Fig. 7 A schematic of the experimental setup used to investigate FWM in PCFs. Reproduced with permission from Sharping JE, Fiorentino M, Coker A, Kumar P and Windeler RS (2001) Four-wave mixing in microstructure fiber. *Optical Letters* 26: 1048–1050. ©2001 Optical Society of America.

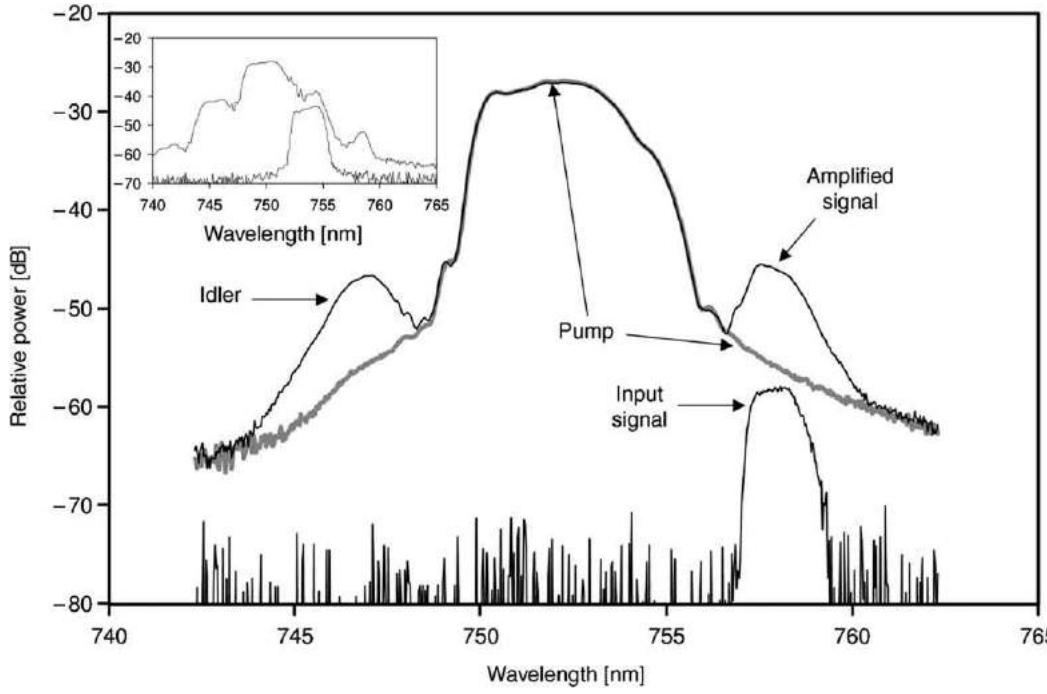


Fig. 8 A typical FWM spectrum observed at the output of the microstructure fiber. The inset shows a spectrum where higher-order cascaded mixing is evident. Reproduced with permission from Sharping JE, Fiorentino M, Coker A, Kumar P and Windeler RS (2001) Four-wave mixing in microstructure fiber. *Optical Letters* 26: 1048–1050. ©2001 Optical Society of America.

near λ_0. Since the dispersion characteristics of these fibers can be adjusted during the fabrication process, the experiments demonstrate the potential for the use of PCFs in broadband parametric amplifiers, wavelength shifters, and other optical communication devices.

The experimental setup used to demonstrate phase-matched FWM in PCF is shown in Fig. 7. The pump and the input signal are two synchronous pulsed beams having 3–5 nm wavelength separation with the center wavelength tunable over a 720–850 nm range. The maximum peak power of the pump pulses is $\simeq$ 12 W. The two synchronous beams are then combined and injected into the PCF. The pump and signal's optical paths are adjusted to obtain temporal overlap in the PCF and their polarizations are aligned by fiber polarization controllers.

Fig. 8 shows a typical FWM optical spectrum at the output of a 6.1 m long PCF. Here the strong pump beam and the weak signal beam have wavelengths of 753 nm and 758 nm, respectively. The spectrum shows the undepleted pump, the amplified signal, and the generated idler at 747 nm. The spectra in Fig. 8 show that large gain is achievable for a pump-to-signal spacing of 5 nm. Gain values of more than 20 (13 dB) were obtained.

Conclusion

In summary, the advantages of using photonic crystal fibers for demonstrating nonlinear-fiber optical effects arise from four novel properties:

- the nonlinear coefficient is enhanced in small-core PCFs (core area of a few μm^2);
- PCFs can support a single transverse mode over an extremely broad wavelength range (370 nm–1600 nm);
- PCF design parameters allow one to manipulate the fiber's GVD properties; and
- nonlinear interactions are enhanced due to the polarization maintaining properties of PCFs which result from form birefringence present in the core.

These four properties combine to allow efficient interactions to occur in PCFs which are either inefficient or not possible at all in standard optical fibers.

See also: Microstructure Fibers

Further Reading

Agrawal, G.P., 2000. Nonlinear Fiber Optics, 3rd edn. San Diego, CA: Academic Press.
Birks, T.A., Knight, J.C., Russell, P.S.J., 1997. Endlessly single-mode photonic crystal fiber. Optical Letters 22, 961–963.
Hansryd, J., Andrekson, P.K., Westlund, M., Li, J., Hedekvist, P., 2002. Fiber-based optical parametric amplifiers and their applications. IEEE Journal of Selected Topics in Quantum Electronics 8, 506–520.
Monro, T.M., Richardson, D.J., Broderick, N.G.R., Bennett, P.J., 2000. Modeling large air fraction holey optical fibers. Journal of Lightwave Technology 18, 50–56.
Ranka, J.K., Windeler, R.S., Stentz, A.J., 2000. Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. Optical Letters 25, 25–27.
Russell, P.St.J., 2003. Photonic crystal fibers. Science 299, 358–362.
Sharping, J.E., Fiorentino, M., Coker, A., Kumar, P., Windeler, R.S., 2001. Four-wave mixing in microstructure fiber. Optical Letters 26, 1048–1050.
Sharping, J.E., Fiorentino, M., Kumar, P., Windeler, R.S., 2002. All-optical switching based on cross-phase modulation in microstructure fiber. IEEE Photonics Technology Letters 14, 77–79.

Photonic Crystal Lasers, Cavities and Waveguides

J O'Brien and W Kuang, University of Southern California, Los Angeles, CA, USA

Introduction

Photonic bandstructure engineering, in which the electromagnetic dispersion relations are intentionally modified, can be accomplished by creating one-, two-, and three-dimensionally periodic dielectric materials. These periodic materials, called photonic crystals, can be used to confine and guide light. This article describes the formation of optical resonant cavities and waveguides in these materials. Most of the discussion will focus on two-dimensional photonic crystals. One-dimensionally periodic systems are treated elsewhere in this volume. This article will begin with a short review of the relevant electromagnetic properties of photonic crystals. The formation and properties of resonant cavities will follow this, and the article will conclude with waveguides formed by linear defects in two-dimensional photonic crystals.

Electromagnetics of Photonic Crystals

A two-dimensional photonic crystal consists of a dielectric material in which the dielectric constant is periodic in two dimensions. The period of these materials is on the order of a half-wavelength of the operating optical wavelength. In this section, we briefly review the electromagnetic properties of two-dimensional photonic crystals. For more details, see the article entitled **Spectroscopy:** Raman Spectroscopy in this volume.

There are five Bravais lattices in two dimensions. Most of the research on photonic crystal resonant cavities and waveguides has focused on square and triangular lattices, and in this article we will consider examples based on the triangular lattice. Since these structures are usually defined lithographically however, there is, in principle, no need to confine the investigation to naturally occurring Bravais lattices. Due to the limitations of current nanofabrication technology, much of the research has focused on photonic crystals that have a finite thickness. The important feature of photonic crystals is that the bandgaps in the electromagnetic spectrum opened up by all of the Bragg planes overlap spatially and spectrally, so that a frequency region exists in which no electromagnetic propagation is allowed. These electromagnetic bandgaps can be used to confine optical modes in very small-volume resonant cavities and in waveguides. In high-contrast dielectric systems, such as a semiconductor/air periodic system, only a few lattice periods may be necessary to confine an electromagnetic mode. In the case of finite thickness photonic crystals, this periodicity has the effect of limiting the bandgap formed to the in-plane directions and in the case of a two-dimensional photonic crystal in a high-index dielectric slab to the formation of a bandgap in the guided modes of the slab. In this case there are still radiation modes of the slab.

Consider a two-dimensional photonic crystal that is periodic in the $x - y$ plane. To start, we can consider materials that are uniform and infinitely extended in z. Electromagnetic fields propagating in the $x - y$ plane are Bloch states. These can be written as:

$$\vec{H}_{k_z,k_\parallel}(\vec{\rho},z,t) = e^{i(\omega t-(\vec{k}_\parallel\cdot\vec{\rho}+k_z z))}\vec{u}_{k_z,k_\parallel}(\vec{\rho}) \tag{1}$$

where $\vec{\rho}$ labels the in-plane coordinates and $\vec{k}_\parallel$ labels the in-plane wavevectors. $\vec{u}_{k_z,k_\parallel}$ is a periodic function in the $x - y$ plane. Modes propagating in the $x - y$ plane in such a system can be classified as transverse electric (TE) waves or transverse magnetic (TM) waves. TE waves have nonzero E_x, E_y, and H_z field components, and TM waves have nonzero E_z, H_x, and H_y field components. A triangular lattice of air holes patterned into a high refractive index dielectric can be used to create a bandgap for both the guided TE and TM modes of the membrane for a range of hole radii. The TE and TM bandgaps that are formed can be over different spectral ranges, however. For laser applications this is unimportant since in practice the emission occurs from electron-hole recombination in a semiconductor quantum well, and for unstrained or compressively strained quantum well materials this emission is TE polarized. In most cases, waveguides have also been designed to work for a single optical polarization. It is generally true that a photonic lattice that consists of a connected high dielectric region is likely to exhibit a TE bandgap, while a lattice formed by disconnected high dielectric regions is more likely to exhibit a TM bandgap. The bandgap in the TE modes is formed between the first and second bands. The first band is called the dielectric band because the field at the Brillouin zone boundary is a standing wave with its intensity concentrated in the high dielectric regions. The second band is called the air band because at the Brillouin zone boundary this field is a standing wave with its intensity localized in the low dielectric regions. Over a reasonable range of lattice parameters, the bandwidth of the TE bandgap can be changed by changing r/a where r is the radius of the holes and a is the lattice constant of the triangular lattice. As r/a gets larger the dielectric band moves up in frequency. This can be thought of as being due to the fact that this mode has a decreasing effective index as r/a increases. The air band frequency increases as r/a increases. Since more of the electric field of the air band is located in the low dielectric regions than the field of the dielectric band, the air band increases in frequency with increasing r/a faster than the dielectric band and the bandgap therefore increases with increasing r/a. This trend holds in photonic crystals of finite thickness as well.

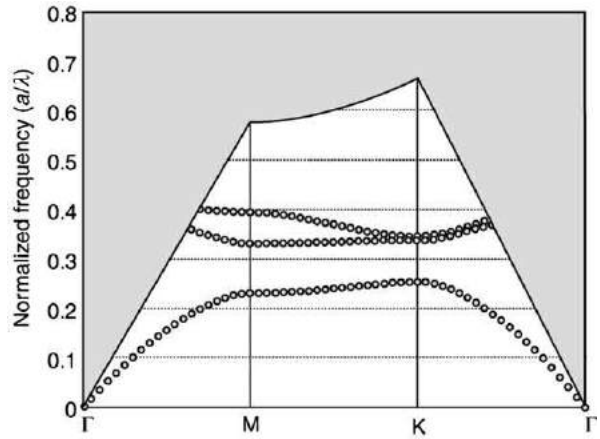


Fig. 1 The photonic band diagram for a photonic crystal slab with a triangular lattice of air holes perforating a high-index dielectric slab. The slab, which assumes a refractive index $n=3.4$, is suspended in the air with a thickness of $d/a=0.6$. The air holes have radii of $r/a=0.3$, where a is the lattice constant. Only the lowest three eigenmodes are shown in the plot.

For photonic crystals of a finite thickness, the ability to simply classify modes as either TE or TM is lost. In the most simple finite thickness photonic crystals, that of a high index dielectric slab in which a two-dimensional photonic crystal has been patterned surrounded above and below by the same low index material which may be air of sapphire, or silicon dioxide, the modes can be classified as being either even or odd modes with respect to the mid-plane of the high index slab. In this mid-plane, however, modes can be classified as being either TE or TM, with each containing the same nonzero vector field components as in the infinite case. Away from this mid-plane, however, modes have in general six nonzero electromagnetic field components. The photonic bandgap that is created in this case is in the guided mode spectrum of the high index dielectric slab. There is no gap formed in the radiation modes of the slab. The radiation modes consist of modes with a ray vector that is not totally internally reflected at the high index slab/low index cladding interfaces. **Fig. 1** shows the dispersion relation, plotted over the irreducible Brillouin zone, of the even modes for dielectric slab with a refractive index of 3.4 in which a two-dimensional triangular lattice of holes has been patterned. The surrounding material is air. The figure shows an electromagnetic bandgap formed between the first and the second guided modes of the slab. Radiation modes of the slab have real valued propagation vectors in the z-direction. These modes lie above the in-plane dispersion relation for air and exist inside the shaded region in the figure. This dividing line between the guided modes and the radiation modes is referred to as the light line.

If the upper and lower cladding layers of the high index slab are different and therefore have different indices of refraction, then the ability to classify modes as being either even or odd about the mid-plane of the slab disappears and it is possible to lose the bandgap in the guided modes altogether. In cases where the refractive index of the bottom cladding is not significantly different from that of the top cladding, such as a case where air serves as the top cladding layer and a material such as sapphire or silicon dioxide forms the bottom cladding, it is found that the even and odd modes describe the field reasonably accurately. To emphasize the approximation involved in the model, these modes are most often referred to in the literature as even-like or odd-like. Overall, there are three primary effects of having asymmetric cladding layers such as an air top cladding and sapphire bottom cladding case. These are a reduction in the effective bandgap width, an increase in the area of the radiation modes on the dispersion relation, and the loss of a rigorous bandgap in the guided modes of the slab. The first two of these effects are the most serious for device designers.

Photonic Crystal Resonant Cavities and Lasers

Photonic crystal resonant cavities can be formed at frequencies inside the bandgap or at the bandedges. Modes in the bandgap are formed as a result of a defect in the lattice. Modes at the bandedge are also sometimes used because the fields with wavevectors at Brillouin zone boundary are standing waves. These standing waves are analogous to the resonant modes that are formed in distributed feedback (DFB) laser structures. Here we will focus on modes confined in the bandgap. Allowed modes in electromagnetic bandgaps in photonic crystal slabs can be introduced by introducing one or more defects into the lattice. By perturbing a single lattice site, we can permit a localized mode or modes that have a frequency in the gap. This phenomenon is similar to the formation of deep levels in an electronic bandgap of a solid due to impurities and identical to the formation of resonant modes in vertical cavity surface emitting lasers (VCSELs) and phase-shifted DFB lasers. In each of these last cases, a defect in the periodicity of a precise thickness occurs between two one-dimensional distributed Bragg reflectors. Defect modes in photonic crystal resonant cavities can be engineered to radiate in a particular direction by trading-off the in-plane versus out-of-plane losses, and the resonant frequency is determined by the parameters of the lattice. Very small mode optical mode volumes can also be obtained in these resonant cavities leading to the possibility of modifying spontaneous emission.

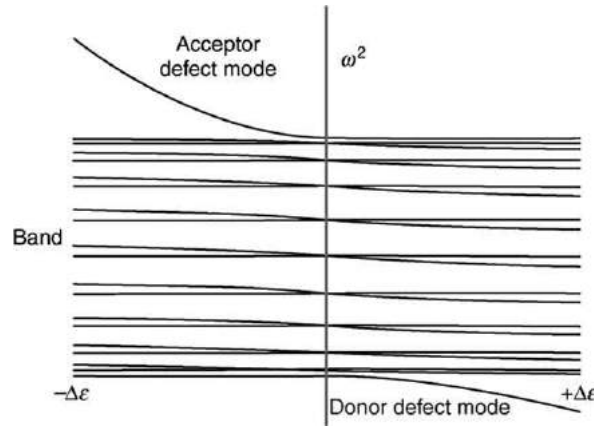


Fig. 2 A one-dimensional Slater–Koster model for a single defect in a 10-unit-cell chain, showing the allowed states as a function of the defect perturbation.

Simple models exist which illustrate the formation of localized modes inside the bandgap as a result of defects in the otherwise periodic lattice. One of these was applied to the formation of deep levels in solids. This model can also be applied to the electromagnetics of defects in periodic structure. For a simple one-dimensional case, assuming that the defect exists at a single unit cell and that the dispersion of the band is dominated by nearest neighbor interactions, this can be solved exactly. Fig. 2 shows the allowed frequencies of a ten unit cell chain with a single defect located at the center unit cell. Notice that one mode, called a donor mode, drops out of the band for positive values of $\Delta\varepsilon$ and one mode, called an acceptor mode, rises out of the band for negative values of $\Delta\varepsilon$. This is true in general. A perturbation of the lattice in which the defect has a higher index than the unperturbed lattice will cause a mode to drop out of the air band while a perturbation that has a lower dielectric constant than the background lattice will cause a mode to rise into the bandgap out of the dielectric band. In practice, resonant modes and frequencies in photonic crystals are modeled numerically.

A great deal of variety exists in resonant modes formed by defects in two-dimensional photonic crystal slabs. The simplest examples consist of a single missing hole in a two-dimensional square or triangular lattice patterned into a high-index dielectric slab. However, the resonant mode size, shape, frequency, and polarization can be engineered by tailoring the local dielectric in the region of the resonant mode. Losses in these cavities can be conceptually separated into losses in-plane through the photonic crystal and radiation losses out-of-plane. The in-plane losses are a result of having only a finite number of lattice periods surrounding the localized mode. A finite number of periods results in a finite loss due to tunneling of the fields through the lattice. This loss can in principle be made arbitrarily small by increasing the number of lattice periods surrounding the resonant mode. For a fixed number of lattice periods, the in-plane radiation loss will be reduced as the bandwidth of the photonic bandgap is increased because the magnitude of the in-plane decay constant increases with increasing bandgap width. A resonant mode formed by a defect in a truly three-dimensional photonic crystal will have losses limited by the number of achievable lattice periods and the bandwidth of the photonic crystal bandgap.

Photonic crystal cavities formed by two-dimensional photonic crystals of finite thickness suffer from an out-of-plane radiation loss. Optical confinement in the direction perpendicular to the two-dimensional photonic crystal is due to total internal reflection of the mode that occurs in the high index photonic crystal slab surrounded by lower index materials. Out-of-plane losses are due to the presence of wavevector components in the resonant mode that lie above the light line inside the radiation cone of the cladding. Fig. 3 shows the Fourier transform of the magnetic field at the mid-plane of photonic crystal slab in which the resonant mode is formed by a single missing hole in a two-dimensional triangular lattice. In this cavity geometry, a doubly degenerate pair of modes is introduced into the photonic bandgap. The figure shows the Fourier transform of the field for one mode of this double degenerate pair. The boundary of the radiation cone in the figure is marked by the solid circle. Components of the mode inside this circle are not confined to the slab and contribute to out-of-plane radiation losses. These losses can be reduced by engineering the mode so that it has a smaller overlap with the radiation cone. This can be accomplished by allowing the mode to expand in real space in directions in which the wavevectors of the mode extend into the radiation cone. This reduces the spread of the Fourier transform of the mode and reduces the overlap of the mode with the radiation cone.

Symmetry can also be used to reduce the coupling to the radiation fields of the slab. Reducing the resonant frequency of the mode can also reduce out-of-plane radiation losses. This has the effect of reducing the size of the radiation cone in k-space. One way to accomplish this is to reduce r/a in the photonic crystal lattice. This will generally reduce frequencies of all of the photonic crystal modes. It will also, however, generally be accompanied by a reduction in the bandwidth of the photonic bandgap which will increase the in-plane losses. As a result, care must be taken in designing photonic crystal resonant cavities. Nevertheless, in the smallest mode volume photonic crystal resonant cavities, the optical losses will generally be dominated by out-of-plane radiation loss.

Quality factors have been predicted to exceed 100 000 in carefully designed photonic crystal resonant cavities in which the optical mode volume is on the order of a few cubic half-wavelengths in the material. Larger resonant cavities formed by

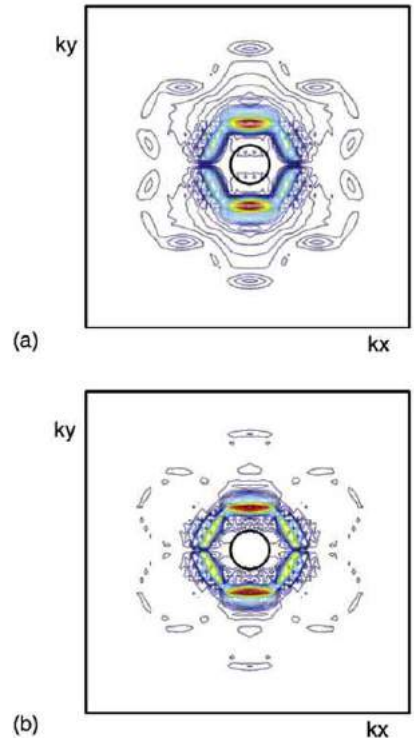


Fig. 3 (a) The spatial Fourier transform of the magnetic field component, Hz, of a defect cavity mode at the mid-plane of a triangular-lattice suspended membrane single-defect photonic crystal microcavity. The inner circle indicates the light cone inside which mode components radiate vertically. (b) The spatial Fourier transform of a modified defect cavity mode. The overlap of the mode with light cone has been reduced by modifying the defect cavity.

introducing multiple defects in the lattice are often less well confined in the real space lattice leading to a reduction in the components overlapping the radiation cone in k-space and a reduction in out-of-plane radiation loss. These cavities also are much more likely to support multiple resonant modes for a given index perturbation.

In analyzing resonance modes of photonic crystal resonant cavities, one of the most important parameters is the quality factor Q of the cold cavity. Practically, a theoretical prediction of the quality factor requires the use of numerical methods and finite-difference time-domain and finite element techniques are commonly used. There are basically three numerical approaches that can be used to calculate the quality factor of the cavity modes. Two of these methods calculate the quality factor from the time domain fields while the third is a frequency domain approach. The first method is to calculate the slope of the exponential decay of the energy of a given cavity mode with time. This decay of energy from the cavity is described by $\exp(-t/t_{\mathrm{ph}})$ where t_{ph} is the photon lifetime which is related to the quality factor Q, by $Q=\omega t_{\mathrm{ph}}$. This method is most useful for relatively low Q modes where the slope of energy decay is visibly greater than zero.

Another method is to calculate the ratio of the full width at half magnitude (FWHM) of the cavity resonance in the frequency domain, $\Delta\omega$, to the center frequency, ω_0. However, distortion to the spectrum is often introduced because the numerical simulation terminates before the impulse response is fully evolved. This has the effect of viewing the true time-domain response through a rectangular window, which translates mathematically into the convolution of the true spectrum with a sinc function. This must be accounted for in the determination of the quality factor.

A third method calculates the ratio of cycle-averaged power absorbed in the boundary to the total energy in the cavity mode. This last method has the advantage of by being able to separate the radiation losses into different directions:

$$\frac{1}{Q}=\frac{\omega P}{U}=\frac{\omega(P_{\perp}+P_{\parallel})}{U}=\frac{1}{Q_{\perp}}+\frac{1}{Q_{\parallel}} \tag{2}$$

in which, the effective vertical quality factor $Q_{\perp}$ is given by the ratio of power lost to the absorber at the top and bottom $P_{\perp}$ to the total cavity energy $U(t)$, and the effective in-plane quality factor $Q_{\parallel}$ is similarly given by the ratio of in-plane power loss $P_{\parallel}$ to the total cavity energy.

Photonic crystal lasers and resonant cavities have been demonstrated experimentally. The first demonstration of lasing in a two-dimensional photonic crystal laser occurred in a cavity formed by a single defect in a triangular lattice that was patterned into an InGaAsP membrane. The laser operated under optically pumped conditions at 143K. Room temperature pulsed operation of photonic crystal defect lasers was reported shortly thereafter. Fig. 4 shows an electron micrograph of such a resonant cavity. In the figure, the resonant cavity is formed in a suspended membrane in which a triangular lattice photonic crystal with a defect has been

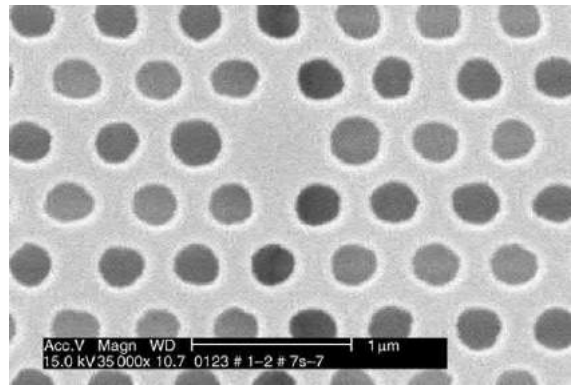


Fig. 4 The scanning electron micrograph of a triangular-lattice suspended membrane photonic crystal single defect cavity.

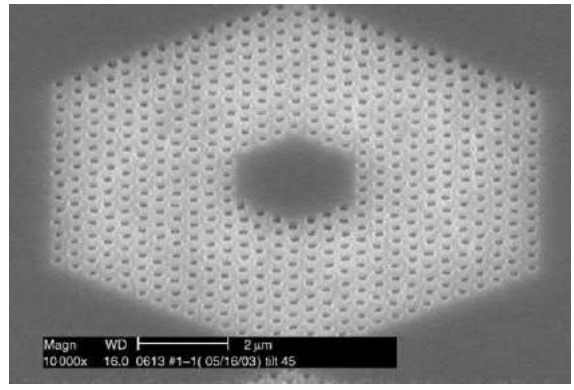


Fig. 5 The scanning electron micrograph of a triangular-lattice, 37-missing-hole, photonic crystal defect cavity, bonded on sapphire which serves as lower cladding to improve heat dissipation. CW operation of the laser cavity was demonstrated.

patterned. The fabrication of these devices involves pattern definition by electron beam lithography and pattern transfer by a series of wet and dry etching steps. The observed quality factors are, of course, sensitive to fabrication imperfections. Many other cavity designs and demonstrations have also been reported.

To operate a photonic crystal laser continuously at room temperature, it is necessary to design a high-Q resonant cavity that dissipates heat well. Poor heat dissipation is a major drawback of suspended membrane resonant cavities. One strategy for improving the heat dissipation in photonic crystal laser cavities is to form the two-dimensional photonic crystal membrane cavity on top of a low-index high-thermal-conductivity substrate. Fig. 5 shows an electron micrograph of a photonic crystal cavity formed by leaving out 37 holes from a triangular lattice patterned into an InGaAsP membrane. This membrane is bonded to a sapphire substrate, which facilitates heat dissipation. This cavity is capable of room temperature continuous wave operation.

It is believed that these lasers are capable of electrically-pumped room temperature operation. This is based on the fact that large quality factors have been demonstrated in optically-pumped photonic crystal lasers and some quality factor reduction due to free carrier absorption and absorption at metal contacts can be tolerated and still lead to lasing. A reasonable assumption would be that the performance of these electrically-pumped photonic crystal lasers would have much in common with the performance of VCSELs. Both VCSELs and photonic crystal cavities have small mode volumes with the cavity formed in some directions by distributed Bragg reflection. The emission direction of photonic crystal lasers can be engineered to be vertical or in-plane, but the basic resonant cavities do have important similarities. The free carrier absorption loss in photonic crystal laser cavities may be expected to be larger than in VCSELs since a photonic crystal laser mode lacks the standing wave behavior in the vertical direction that allows doping at the nodes of the standing wave to reduce the absorption loss. If free carrier absorption loss leads to larger internal losses for photonic crystal lasers than for VCSELs, then some reduction in the achievable slope efficiencies will also occur for photonic crystal lasers.

Photonic Crystal Waveguides

Another important optical element for integrated photonic circuitry is a linear waveguide. Conventional dielectric waveguides confine propagating beams by an index of refraction difference between the waveguide core and the waveguide cladding. Photonic crystal waveguides are most often formed by a linear defect, which consists of a row of modified unit cells of the lattice, patterned

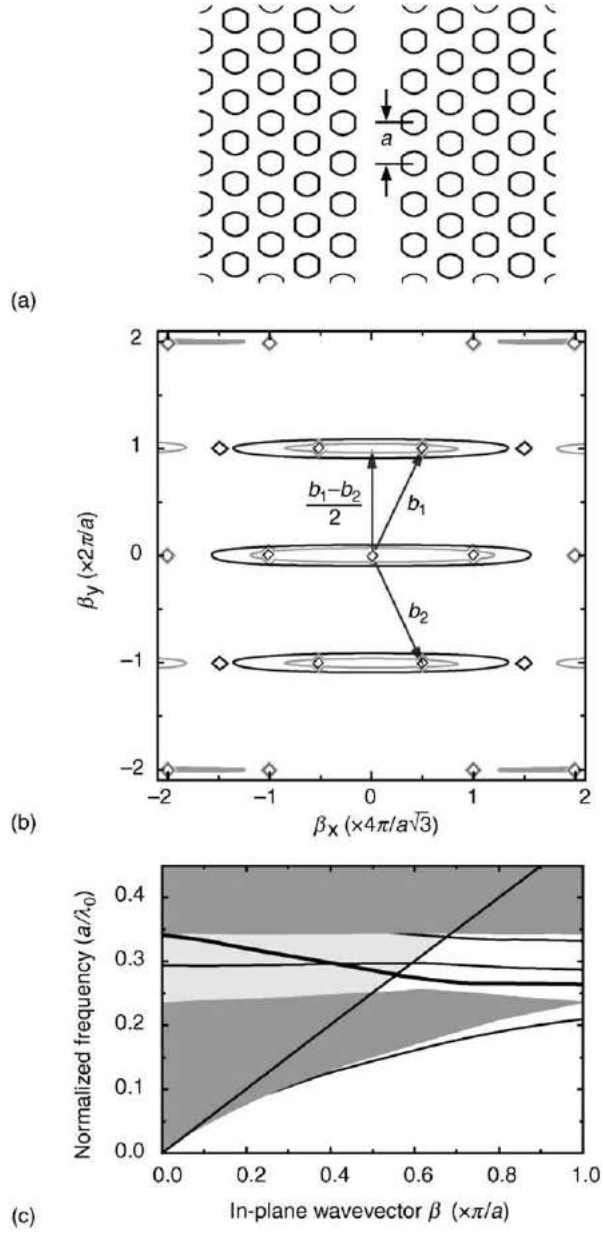


Fig. 6 (a) The top view of a triangular lattice photonic crystal single-line defect waveguide with air hole radius $r/a=0.3$. (b) The 2D spatial Fourier transform of the waveguide dielectric distribution. (c) The photonic band diagram for a suspended membrane single-line photonic crystal waveguide. The membrane has a refractive index of $n=3.4$ and its thickness is $d/a=0.6$. The surrounding material is air. The light gray and dark gray area indicate the regimes where field radiates through the air and photonic crystals, respectively.

into a high index dielectric membrane. The guiding in this case is due to a mixture of total internal reflection at the high index membrane/low index cladding interfaces and distributed reflections from the photonic crystal in-plane. Because this confinement does not exclusively dependent on total internal reflection, the transmission loss through small bending radius waveguide branches and bends can be made very small. Predictions and experimental demonstrations exist which show that photonic crystal waveguides are capable of supporting small bending radius waveguide bends with almost total transmission. This makes photonic crystal waveguides a candidate for waveguides in densely integrated photonic circuits.

Fig. 6(a) shows an illustration of the top view of a two-dimensional photonic crystal waveguide. This is a triangular lattice in which one row of unit cells has been modified. Shown in Fig. 6(b) is the Fourier transform of index of refraction for the situation in which the photonic crystal consists of low index holes in a high index semiconductor. The presence of the linear defects modifies and broadens the reciprocal lattice from a simple reciprocal triangular lattice, before the creation of linear defect, to the distribution shown in the figure. This Fourier transform has components along the direction of propagation at $(\vec{b}_1-\vec{b}_2)/2$ as shown in the figure where $\vec{b}_1$ and $\vec{b}_2$ are reciprocal lattice unit vectors. These reciprocal lattice components at $(\vec{b}_1-\vec{b}_2)/2$ couple a

planewave with wavevector β to other planewaves with wavevectors $\beta \pm (\vec{b}_1 - \vec{b}_2)/2$. The result of this coupling is that the electric field of the waveguide mode is a Bloch wave. In a photonic crystal waveguide in which the waveguide axis is the z-direction, the Bloch wavefield in the waveguide can be written in the form:

$$\vec{E}(x,y,z,t) = \sum_n c_n \vec{f}_\beta(x,z) \times \exp\left[\mathrm{i}\left(\omega t - \left[\beta + n\left(\frac{\vec{b}_1 - \vec{b}_2}{2}\right)\right]y\right)\right] \tag{3}$$

In this expression the periodic function of the Bloch wave has been explicitly expanded as a Fourier series and each term in the Fourier series is called a spatial harmonic. Note that these fields differ from that of a photonic crystal fiber because the cladding of a photonic crystal waveguide has a periodicity along the direction of a waveguide. A photonic crystal fiber does not.

Most of the realizations of photonic crystal waveguides occur in two-dimensional photonic crystals formed in high index dielectric slabs. The membrane is usually located a few microns above a high index substrate. Because the fields of the guided modes decay exponentially outside the membrane, the effect of a substrate several microns away is very nearly negligible. The photonic band structure for $r/a = 0.3$, normalized membrane thickness $d/a = 0.6$ and refractive index of $n = 3.4$ is shown in **Fig. 6 (c)**. Only the modes with even symmetry along the mid-plane of the membrane are shown. A bandgap opens for modes with odd symmetry only for large low index filling factors in the triangular lattice. Filling factors large enough to create a bandgap for the odd modes are often impractical so that the odd modes are rarely considered. The even and odd modes are orthogonal, so that plotting the dispersion relation for the even modes only is justified. In reality, even and odd modes may be weakly coupled as a result of fabrication imperfections. The horizontal axis of the dispersion relation covers the in-plane wavevector β along the propagation direction of the irreducible Brillouin zone. The vertical radiation light cone and transverse radiating region of the photonic crystal are mapped as light gray and dark gray areas, respectively. They correspond to the regimes in which light radiates through the air cladding and photonic crystal cladding, respectively. Waveguide dispersion relations can be calculated using two-dimensional approaches in which case the guiding due to the high index slab is accounted for by using an effective index of refraction for the mode. The effective index is the ratio of the propagation coefficient to the free space wavevector of the guided

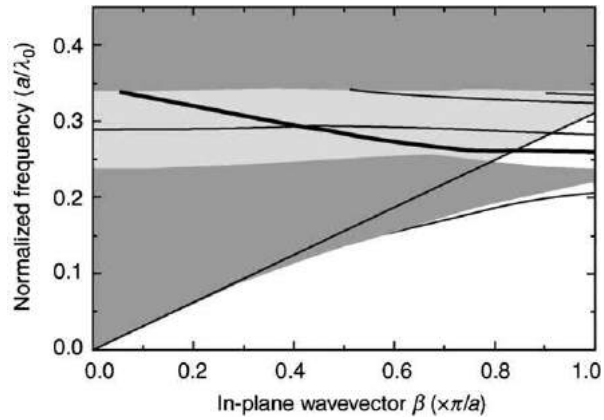


Fig. 7 The photonic band diagram for a triangular-lattice sapphire-bonded single-line photonic crystal defect waveguide. The dielectric membrane has a refractive index of $n = 3.4$ and thickness of $d/a = 1.0$. The index of the sapphire bottom cladding is assumed to be 1.6.

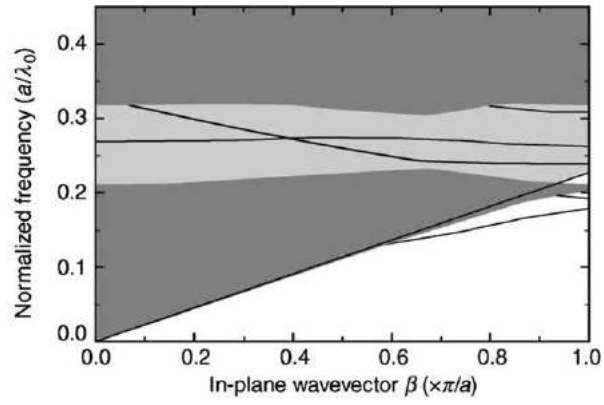


Fig. 8 The photonic band diagram for a triangular-lattice deeply etched single-line photonic crystal defect waveguide. The top cladding, guiding membrane, and bottom cladding have an index of refraction of 3.0, 3.4, and 3.0, respectively. Their normalized thicknesses d/a are 1.0, 1.0, and 6.0, respectively.

mode of the slab. The two-dimensional calculation then uses the effective index of the mode as the high index of material in the photonic crystal. To obtain accurate results, the effective index for each mode calculated separately and the two-dimensional calculation of the photonic crystal waveguide dispersion must be repeated for each waveguide mode.

Examples of photonic crystal waveguide dispersion relations are shown in Figs. 7 and 8 for waveguides formed by photonic crystals in dielectric slabs on high index lower cladding layers. Fig. 7 is the dispersion relation for an oxide lower cladding layer. The area of the light gray region, which is the projection of the oxide cladding light cone onto waveguide propagation direction, increases as a result of higher index of refraction. Fig. 8 shows the dispersion relation for a waveguide with an even larger lower cladding index. In this case the air holes of the two-dimensional photonic crystal are patterned all the way through the lower cladding layer, with an index of 3.0, on a high index substrate. The defect modes lying within the in-plane photonic bandgap are all above the cladding light line in this case. Propagation losses are predicted to increase by over two orders of magnitude from the structure with very low index cladding layers shown in Fig. 6 to the high index lower cladding structure of Fig. 8.

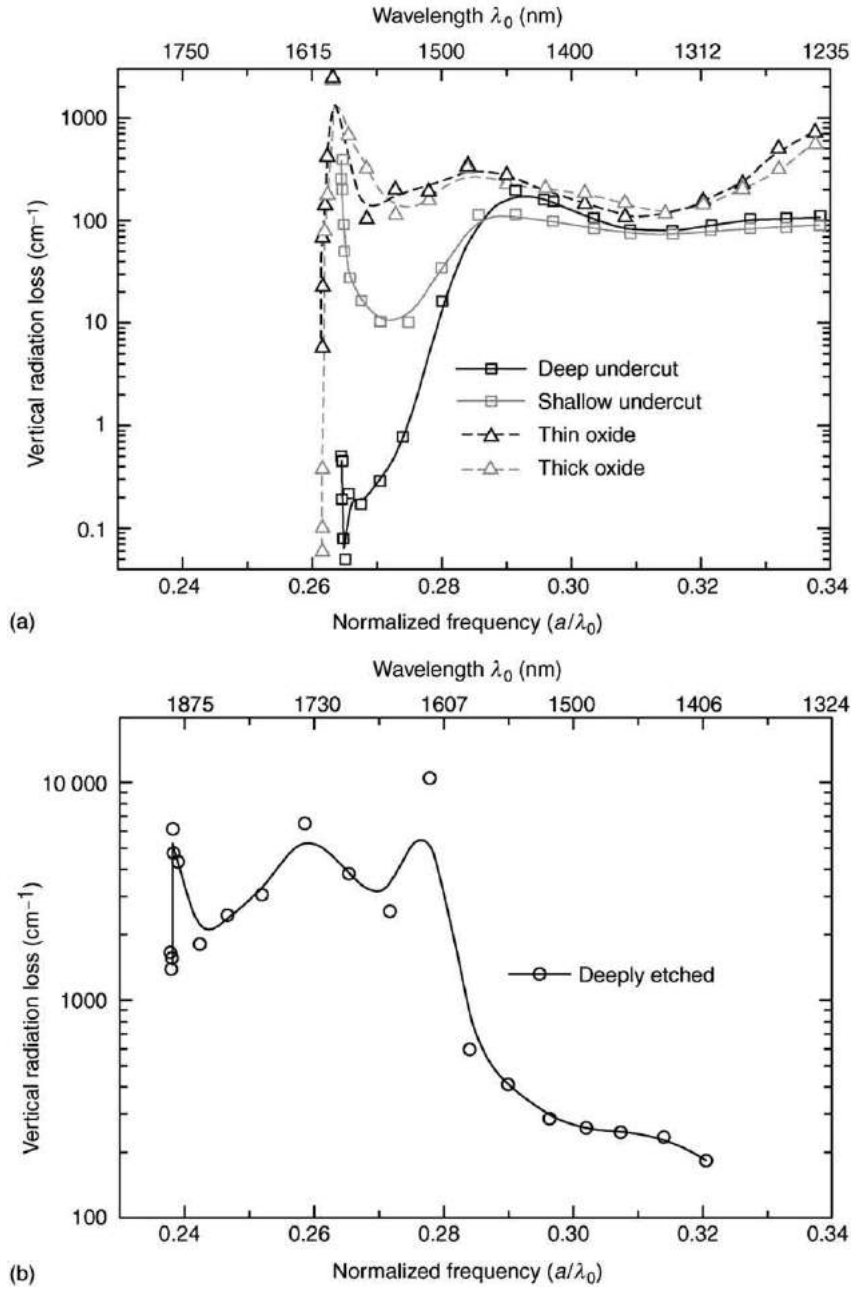


Fig. 9 The out-of-plane radiation loss as a function of normalized frequency for the photonic crystal defect slab waveguides illustrated in Figs. 6–8. The lattice constant $a = 450$ nm for deeply etched waveguide and 420 nm for the rest of the cases. Points are calculated values and lines are B-spline curve fits.

Table 1 Photonic crystal defect slab waveguides considered in the calculations

Waveguide structures	*Suspended membrane*		*Oxidized lower cladding*		*Deeply etched*
	Deep undercut	*Shallow undercut*	*Thin oxide*	*Thick oxide*	
Layer description: Material (normalized thickness *d*/*a*)			Air (>3.0)		
	–	–	–	–	AlGaAs (1.0)
	GaAs (0.6)	GaAs (0.6)	GaAs (0.6)	GaAs (0.6)	GaAs (1.0)
	Air (3.0)	Air (1.0)	Al_xO_y (2.0)	Al_xO_y (5.0)	AlGaAs (6.0)
			GaAs substrate (>3.0)		

Refractive index of 3.4 is assumed for GaAs, 3.0 for AlGaAs, 1.6 for Al_xO_y, and 1.0 for air.

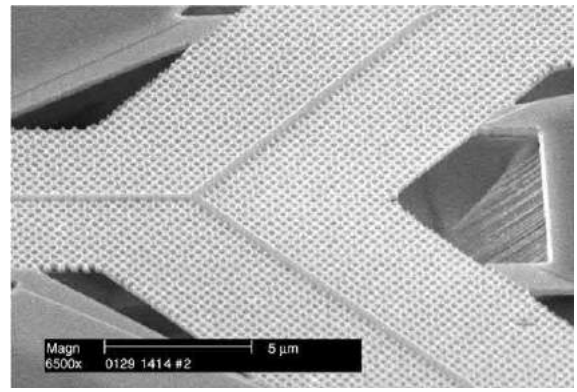


Fig. 10 The scanning electron micrograph of a photonic crystal Y-branch fabricated on InP material system.

Fig. 9(a) and (b) shows the results of a fully three-dimensional calculation for the propagation loss of the three structure shown in Figs. 6–8. The details of these structures are shown in Table 1. This calculated loss is due to coupling between the guided mode and the radiation modes of the high index substrate. A wavelength scale is also included using $a = 420$ nm in Fig. 9(a) and 450 nm in Fig. 9(b). The suspended membrane suffers the least vertical radiation loss among all of the waveguides considered and has a minimum loss around 0.2 cm^{-1}. Low-loss waveguides formed on high index cladding layers may be possible, but care should be taken to eliminate coupling between the guided modes and the radiation modes of the cladding. In this case, the design strategy for reducing the optical loss is to reduce the magnitude of the Fourier component of the electromagnetic field inside the light cone. It should also be noted that deeply-etched waveguides formed by removing multiple rows of holes have been demonstrated with significantly lower radiation loss than is predicted for the single line defect waveguides. However, these waveguides are multi-moded at all frequencies. It is also important to remember that the nanofabricated waveguides will suffer additional losses due to fabrication imperfections that were not included in the numerical model. Fig. 10 shows a nanofabricated photonic crystal waveguide. The image is an electron micrograph of a y-branch component of a Mach–Zehnder interferometer formed in a two-dimensional photonic crystal waveguide patterned into a suspended InGaAsP membrane.

Finally, it is worth noting that the group velocity of photonic crystal waveguide modes, where the group velocity is given by the gradient of the dispersion surface:

$$v_g = \nabla_{\vec{\beta}}\omega \tag{4}$$

is very small near the zone boundary. This can be seen from the dispersion relations shown in Figs. 6–8. In fact, the slope of the dispersion relations is likely to be a parameter which can be engineered by careful waveguide design. This property may allow photonic crystal waveguides to find applications in optical processing functions such as delay lines.

See also: Electromagnetic Theory

Further Reading

Baba, T., 1997. Photonic crystals and microdisk cavities based on GaInAsP–InP system. IEEE Journal of Selected Topics in Quantum Electronics 3, 808–830.
Joannopoulos, J.D., Meade, R.D., Winn, J.D., 1995. Photonic Crystals: Molding the Flow of Light. Princeton: Princeton University Press.
Mekis, A., Chen, J.C., Kurland, I., Fan, S., Villeneuve, P.R., Joannopalos, J.D., 1996. High transmission through sharp bends in photonic crystal waveguides. Physics Review Letters 77, 3787–3790.

Painter, O., Lee, R.K., Yariv, A., *et al*., 1999. Two-dimensional photonic crystal defect laser. Science 284, 1819–1821.
Sakoda, K., 2001. Optical Properties of Photonic Crystals. New York: Springer.
Slater, J.C., 1967. Insulators, Semiconductors, and Metals, vol. 3. New York: McGraw-Hill Book Company.
Srinivasan, K., Painter, O., 2002. Momentum space design of high-Q photonic crystal optical cavities. Optics Express 10, 670–684.
Villeneuve, P.R., Fan, S., Joannopoulos, J.D., 1996. Microcavities in photonic crystals: mode symmetry, tenability, and coupling efficiency. Physics Review B 54, 7837–7842.
Yariv, A., Yeh, P., 1984. Optical Waves in Crystals: Propagation and Control of Laser Radiation. New York: John Wiley & Sons.

Microstructure Fibers

RS Windeler, OFS Laboratories, Murray Hill, NJ, USA

Introduction

Microstructure fibers have unique properties that cannot be obtained from traditional fibers (i.e. all glass, doped silica fibers) and can deliver functionalities superior to many of today's best transmission and specialty fibers. Their unique properties are obtained from an intricate cross-section of high and low index regions that traverse the length of the fiber. The vast majority of fibers consist of silica for the high-index region and air for the low-index region. These fibers are known by several different names including microstructure fiber, holey fiber, and photonic crystal fiber. The term microstructure fiber is used in this chapter because it is the most generic, encompassing all the fiber types.

Microstructure fiber properties vary greatly and are determined by the size and arrangement of air holes in the fiber. For example, fibers have been fabricated such that the numerical aperture of the core or cladding approaches one. They have been fabricated with unique dispersion profiles, such as a zero-group velocity dispersion in the near visible regime or an ultraflat dispersion over hundreds of nanometers wide. They have been fabricated to generate a super continuum over two octaves wide. Some guide light with an air core via bandgap guidance. In addition, the air holes in these microstructure fibers can be filled with highly tunable materials, giving one the capability of controlling the fiber properties for use in energy efficient devices.

Microstructure fibers can have doped regions like traditional fibers. These hybrid fibers combine the benefits of traditional and microstructure fibers. The doped core typically guides most of the light, but its guidance properties can be strongly influenced by the air regions. Applications for hybrid fibers include dispersion compensation, polarization maintaining, and bend insensitive fibers.

Loss is always an important factor when determining whether a microstructure fiber will compete with or replace a traditional optical fiber. The index profiles that make these fibers so special can also lead to fibers that have an inherently high loss at connections or along the length of the fiber. Loss at connections typically occurs due to mode mismatch between the two fibers, undesired hole collapse, or poor alignment. Loss along the fiber length can occur due to impurities, hole surface roughness, or poor confinement.

The dispersion profile (zero dispersion wavelength, dispersion slope, normal or anomalous) of microstructure fibers can be optimized more than a traditional fiber, because the hole size and placement can be arranged to make the cladding index strongly wavelength dependent. Dispersion causes a pulse to spread because the phase velocity is wavelength dependent. Normal dispersion is when longer wavelengths travel faster than shorter wavelengths. Anomalous dispersion is just the opposite–shorter wavelengths travel faster than longer wavelengths. Dispersion determines the amount of interaction between different wavelengths. For applications such as transmission fiber, one wants little interaction and for high nonlinear applications, one wants significant interaction.

This chapter examines the methods for fabricating microstructure fibers and reviews the different types of microstructure fibers and their properties. Fibers that guide by total internal reflection are reviewed separately from bandgap guided fibers because their properties are significantly different. Lastly, a description and example of an active device is presented in which the air regions of a microstructure fiber are filled with controllable material.

Microstructure Fiber Fabrication

At first glance, fabrication of microstructure fibers looks similar to traditional fibers in that the fibers are fabricated in a two-step process. First a preform is fabricated and then it is drawn (stretched) into a fiber. However, when the process is examined in more detail, both the preform fabrication and draw depart significantly from traditional methods. First, a novel preform fabrication method is used to incorporate air holes that run the length of the preform. Second, novel drawing procedures are used to keep the air holes open as the preform diameter is reduced by several orders of magnitude during draw.

Preform Fabrication

Microstructure preforms are cylinders of amorphous material usually less than a few centimeters in diameter that have an index profile running the length of the preform similar to the desired fiber. The most common preform consists of air and a single material such as silica, polymer, or high nonlinear glass. These air-containing preforms are relatively easy to fabricate but are difficult to draw because the structures can deform.

The basic parameters of the fiber are usually determined by the geometry of the holes in the preform. These include hole diameter (d), hole position, pitch (center-to center hole spacing, Λ), core diameter (a), and number of layers. However, changes in

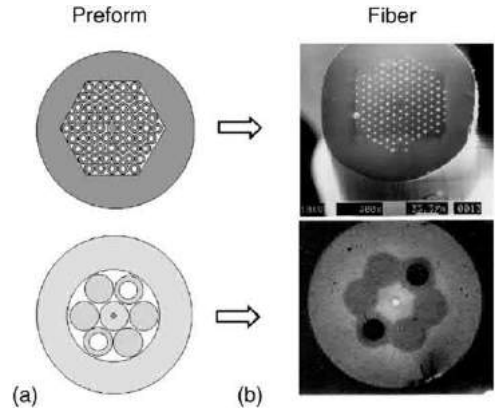


Fig. 1 (a) Tubes, rods, and core rods are stacked together in a close packed arrangement and held together with an overclad tube. (b) During draw the desired air region stay open forming a microstructure fiber.

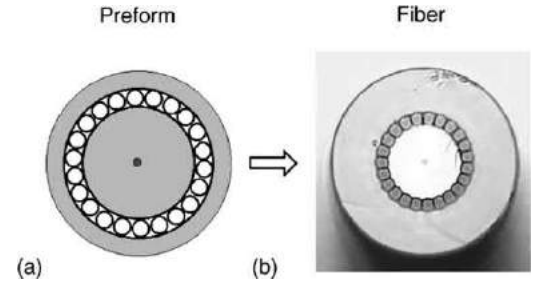


Fig. 2 (a) An air-clad preform is fabricated by stacking a ring of small, thin walled tubes around a large core rod. (b) Photograph of the resulting fiber.

the size and shape of the holes in the fiber can be made purposely or accidentally during draw, which will cause deviations between the fiber and preform profiles.

Several methods are used to fabricate microstructure preforms including stacking, casting, extrusion, and drilling. Each process has its advantages and disadvantages. By far, the most common method is stacking in which tubes, rods, and core rods (rods containing doped cores fabricated by traditional optical fiber techniques) are stacked in a close packed pattern such as a triangular or hexagonal lattice (**Fig. 1**). The bundle of tubes and rods is then bound together with a large tube, called an overclad tube. Fibers can be made with complicated or asymmetric index profiles by strategically placing tubes with the same outer diameter (for good stacking) but a different inner diameter to change the index in that location. For example, a fiber with smaller holes on opposite sides of the core produces high birefringence.

The main advantages of stacking are that no special equipment is needed for fabricating microstructure preforms and that doped cores are easily added. However, there are several disadvantages. First, the stacking method is limited to simple geometries of air holes because the tubes are stacked in a close pack arrangement. Second, unless hexagonal tubes are used, interstitial areas are created between the tubes that may not be desired in the final fiber. Third, the method is labor intensive and requires significant glass handling.

A variation of the stacking method is used for fabricating air-clad fibers. Air-clad preforms are created by placing a layer of small tubes around the perimeter of a large core rod. The assembly is then overcladed for strength (**Fig. 2**).

Extrusion and casting of glass powder, polymer, or sol–gel slurry are also used for making single material microstructure preforms. The main advantage of extrusion and casting is that complicated structures can be fabricated in which the position, size, and shape of the air regions are independent of one another. The disadvantage of this method is that preform fabrication is more difficult compared to the other fabricate methods. This method may become more common as complex air structures are needed to make advanced microstructure fibers.

The last method consists of drilling holes in a preform or rod. Drilling is well understood and used for other specialty fibers. The advantage of this method is that it is easy to drill holes of various sizes in any position in a preform, including doped regions. The disadvantages of drilling are that the holes cannot be drilled very deep compared to a traditional preform length; the distance between holes may be limited due to cracking; and the fiber may experience high loss due to surface roughness of the holes and impurities incorporated during drilling.

Fiber Draw

It is significantly more difficult to draw a microstructure preform with air holes than a traditional preform. This is because the air holes tend to close due to surface tension. This force can be overcome or minimized using two different techniques. The first is to draw the preform under very high tension (low temperature). Minimal hole collapse occurs when the draw tension is significantly larger than the surface tension. The problem with this method is that the break rate increases substantially at higher tension.

The second method is to apply an external pressure to the holes to counteract the surface tension. If a single pressure source is used, the holes can be made larger or smaller during draw by changing the pressure as needed. However, the process can become unstable because a large hole needs a lower pressure to maintain its size than a small hole. If holes of different sizes exist, a larger hole will grow at the expense of a smaller one regardless of the pressure used. To minimize the instability, this method is typically performed in conjunction with high-tension drawing.

To alleviate the draw problem, a second solid material can be used in the preform to obtain the desired index profile. With the air regions removed from the preform, the drawing process becomes significantly easier. In addition, when a preform consists of two solid materials, one of the materials does not have to be continuous, as in single material preforms. This allows for simpler designs such as concentric rings, which may be easier to fabricate and model (Fig. 13(c)). The disadvantage of this approach is finding two materials that have compatible physical and chemical properties, have low loss, and have a large index of refraction difference.

Microstructure Fiber Types

Microstructure fibers are categorized by their method of guidance, properties, and function. Index guided microstructure fibers most closely resemble traditional fibers and are described first. These fibers guide light by total internal reflection and have a core index of refraction greater than the cladding. Bandgap fibers are examined next. The unique guidance of bandgap fibers allows the core index to be lower than the cladding index. Tunable microstructure fibers are described last. These microstructure fibers are filled with tunable materials so that the fiber's properties can be actively manipulated.

Index Guided Fibers

Most microstructure fibers guide light by total internal reflection. The size and spacing of air holes determine the guiding properties. For example, a solid silica core surrounded by a cladding consisting of small air holes that are spread apart creates a small core-to-cladding index difference similar to what can be achieved in traditional fibers. At the other extreme, a solid silica core surrounded by a cladding of large holes, which are closely spaced together, forms a large core-to-cladding index difference. This structure creates a large numerical aperture in the core and optically isolated regions within the fiber. The ability to create regions of significantly different indices in various parts of the fiber distinguishes the index guided microstructure fibers from traditional fiber. Different types of microstructure fiber are formed by varying the size, spacing and pattern of the air holes.

Different techniques are used to model these fibers, depending on spacing and size of the air holes relative to the wavelength of the guided light. If the hole diameter and spacing is much smaller than the wavelength of the guided light, the cladding index can be modeled as an average index weighted by the volume fraction of the two materials (typically air and silica). As the hole diameter or spacing approaches the wavelength of the guided light, complicated models are used such as finite difference method, finite element method, beam propagation method, or multipole method.

Endlessly single-mode fiber

The endlessly single-mode fiber consists of a solid core surrounded by a two dimensional array of small air holes in the cladding (Fig. 3). The size and position of the air holes are arranged such that only the fundamental mode exists in the core regardless of the wavelength of the guided light.

The prevention of higher order modes can be understood by examining the$V$ number. The V number is a dimensionless parameter that describes the guiding nature of the waveguide. In traditional fiber, when the V number is less than 2.405, the fiber will guide only the fundamental mode. The V number is given by

$$V = \frac{2\pi a}{\lambda}(n_{\text{core}}^2 - n_{\text{clad}}^2)^{1/2} \tag{1}$$

where a is the core diameter, λ is the wavelength of the guided light, and n_{core} and n_{clad} are the index of refractions of the core and cladding, respectively. In traditional fiber, the core and cladding index (n_{core} and n_{clad}) are for the most part constant, so the V number increases as the wavelength decreases.

However, for the endlessly single mode microstructure fiber, the V number does not increase indefinitely with decreasing wavelength, but approaches a constant value. Obtaining a V number independent of wavelength is understood by examining the distance the light propagates into the cladding as a function of wavelength. As the wavelength becomes shorter, light penetrates a shorter distance into the cladding, interacting with less of the air regions. This causes the effective cladding index (n_{clad}) to increase with decreasing wavelength, such that V approaches a constant value.

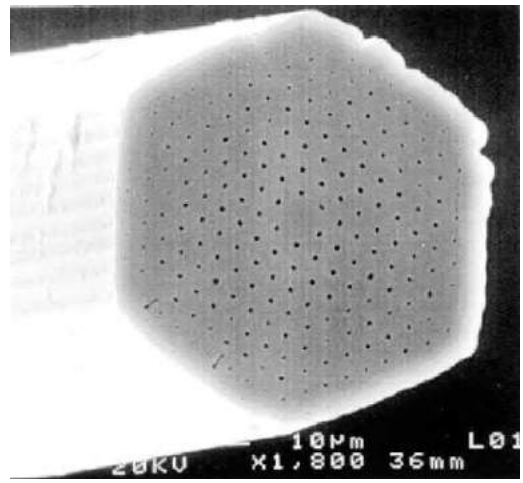


Fig. 3 Photograph of an endlessly single mode fiber. Reprinted with permission from Birks TA, Knight JC and Russell P (1997) Endlessly single mode photonic crystal fiber. *Optics Letters* 22: 961–963.

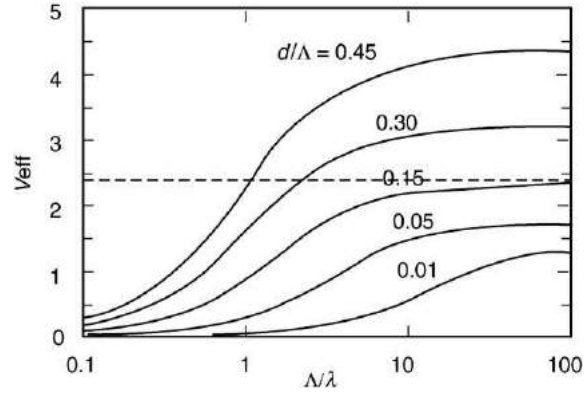


Fig. 4 The effective V number as a function of Λ/λ for various d/Λ. Reprinted with permission from Birks TA, Knight JC and Russell P (1997) Endlessly single mode photonic crystal fiber. *Optics Letters* 22: 961–963.

Microstructure fibers are made endlessly single-moded over all wavelengths by properly designing the hole diameter-to-pitch ratio (d/Λ). When the d/Λ is low enough, the microstructure fiber guides light only in the single mode regardless of wavelength and core size (**Fig. 4**). However, these large core diameter fibers are limited by long wavelength bend loss in the same way as large core single mode traditional fibers.

High delta core fiber

High delta core microstructure fibers (HDCMF) consist of a small (typically less than 2 microns) solid silica core surrounded by one or more layers of large air holes closely spaced together (**Fig. 5**). This creates a very large index difference between the core and cladding. A core-to-cladding index difference of 0.4 is easily achieved in a HDCMF, which is an order of magnitude larger than can be achieved with traditional fiber.

Even though the core is very small, these fibers are almost always multimoded due to the large index difference between the core and cladding. However, because of its small core, it is difficult to launch anything other than the fundamental mode into the fiber. Once a mode is guided in the fiber it remains in that mode due to the large difference in effective indices between modes (**Fig. 6**). When higher-order modes are purposely generated in the fiber, they also guide over long lengths without coupling to other modes (**Fig. 7**). This strong guidance makes the fiber extremely bend insensitive.

HDCMF can have a unique dispersion profile because it can guide the fundamental mode exclusively in the multimode regime. This allows greater flexibility in controlling the shape of the dispersion profile, especially the ability to move the zero dispersion to shorter wavelengths. The HDCMF is the first silica fiber with anomalous dispersion of single-mode light below 1270 nm (the wavelength at which the material dispersion of silica is zero). These fibers are often made to have a zero-group velocity dispersions around 770 nm for use with Ti-sapphire lasers.

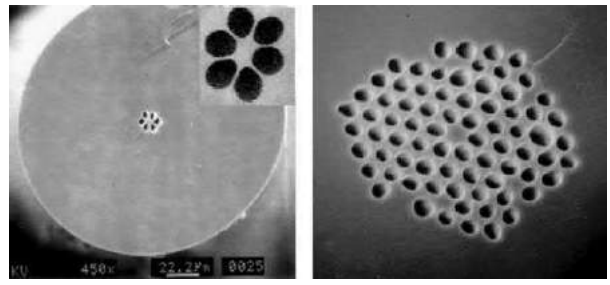


Fig. 5 High delta core fibers consist of a small solid core surrounded by at least one layer of large air holes closely spaced together. Right-hand panel reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798.

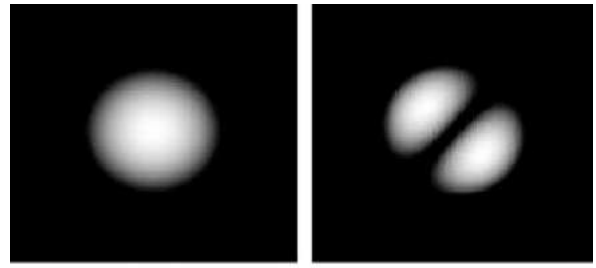


Fig. 6 Individual modes do not couple well due to large differences in effective indices between modes. The figure shows the mode profile and effective index of the fundamental and the next higher mode for a fiber with a two-micron diameter core with a core-to-cladding index difference of 0.45. Reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798.

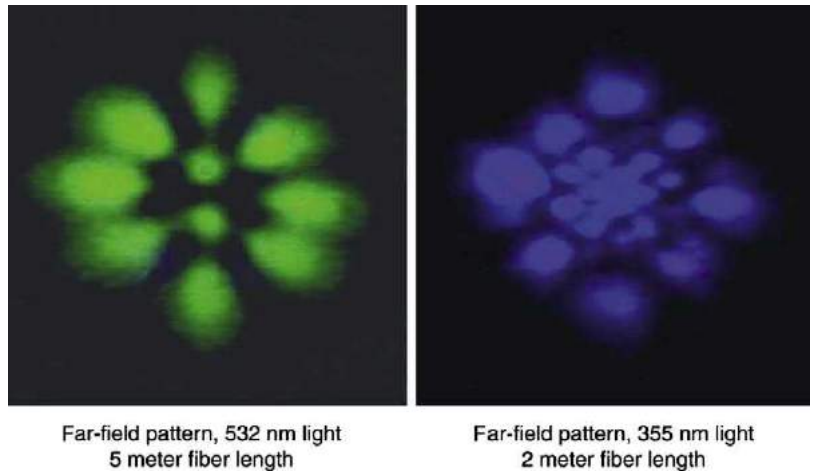


Fig. 7 Far-field patterns show that higher order modes propagate long distances in the HDCMF without coupling to other modes. Reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798.

HDCMF is ideal for performing nonlinear experiments in the near-visible regime because the fiber's small, high-index core creates high intensity light, and the fiber's low dispersion creates long interaction times. For example, a broadband continuum can be generated over two octaves using less than a meter of HDCMF fiber (**Figs. 8** and **9**).

Tapered microstructure fiber (TMF)

Coupling light in and out of HDCMF is difficult and results in a large loss due to its small core. When free space optics are used to couple the light, frequent realignment of the beam is required to minimize loss due to drifts in the equipment. Splicing HDCMF

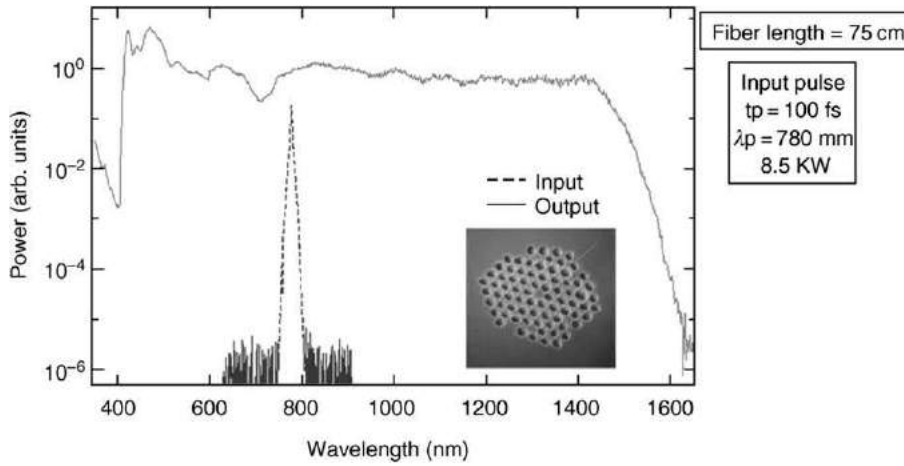


Fig. 8 Plot of a supercontinuum spectrum over two octaves wide generated from a 75 cm piece of HDCMF. Reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optics Letters* 25: 25–27.

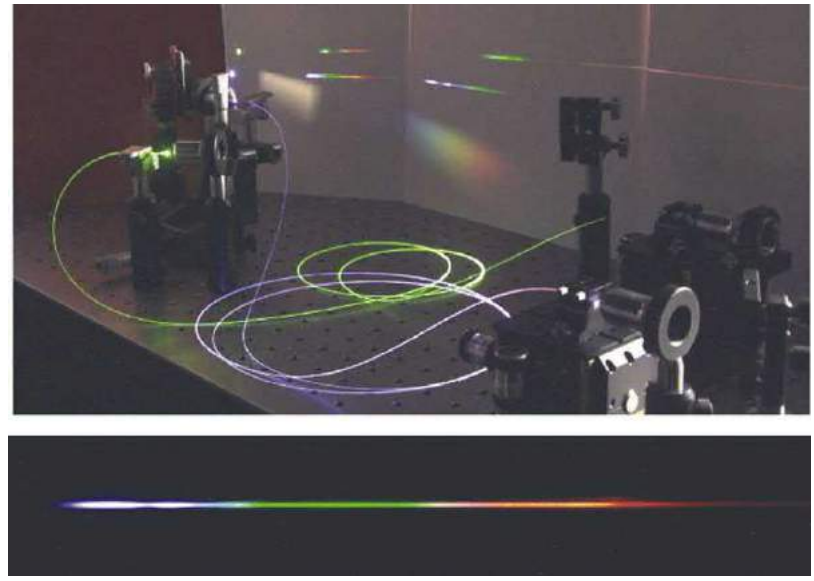


Fig. 9 Photograph of fiber generating the supercontinuum. The bottom photo shows the continuum after passing through a prism.

fiber to traditional fiber also results in high loss due to a large modal mismatch. These coupling problems are solved with tapered microstructure fibers.

Tapered microstructure fibers (TMF) consist of a traditional germanium single-mode core surrounded by a small silica cladding (**Fig. 10**). The inner cladding is surrounded by a layer of air holes and then a protective outer silica cladding for strength. Since the core of the TMF is very similar to a traditional single-mode fiber core, the splice loss is low. Splice losses below 0.075 dB are typical.

Properties similar to the HDCMF are obtained by adiabatically tapering the TMF while maintaining the cross-sectional profile of the fiber. In the taper region, the fundamental mode is no longer guided by the germanium core and evolves into the fundamental mode of the silica region, where it is confined by the ring of air holes. The taper waist is typically ten to twenty centimeters long and has properties identical to the HDCMF. When the fiber is adiabatically expanded back to its original size, the light is guided back into the standard diameter germanium core, making it easy to splice to traditional fiber with low loss.

The same effect can be obtained by stretching a traditional fiber from 125 microns down to 2 microns. However, there are several disadvantages compared to the tapered microstructure fiber. The taper region is much longer in the traditional fiber; the optimal diameter of the silica core is harder to maintain; the taper region must remain clean and uncoated; and the taper region is much weaker.

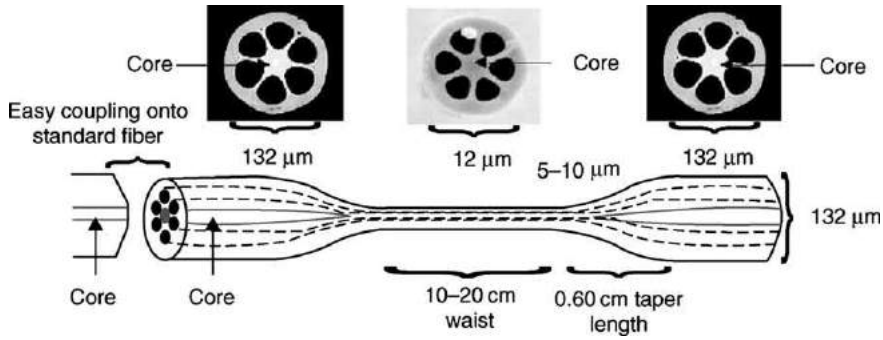


Fig. 10 Schematic drawing of the tapered microstructure fiber. Reprinted with permission from Chandalia JK, Eggleton BJ, Windeler RS, Kosinski SG, Liu X and Xu C (2001) Adiabatic coupling in tapered air-silica microstructures optical fiber. *IEEE Photonics Tech. Letters* 13(1) (©2001 IEEE).

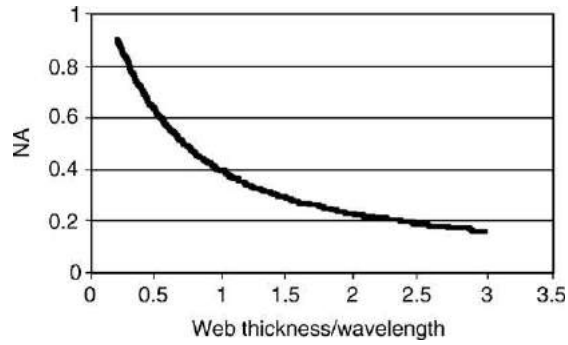


Fig. 11 Predicted NA as a function of the web thickness divided by the wavelength using an infinite slab model.

Air-clad fiber

Air-clad fibers consist of a doped core surrounded by a silica inner cladding. The inner cladding is surrounded by a ring of air, which in turn is surrounded by an outer silica cladding for strength (**Fig. 2**). Thin silica webs connect the inner and outer claddings to hold the inner fiber region in place. The region consisting of air and thin silica webs creates an optical barrier, preventing most of the light from escaping the inner cladding. Such fibers are referred to as double clad fiber.

In traditional double-clad fibers, a material of lower index (typically another layer of glass or a layer of polymer) immediately surrounds the inner silica cladding. A key optical parameter describing the light guiding properties of the inner cladding is the numerical aperture (NA). The NA is defined as the sine of the largest (acceptance) angle of light that will be guided in the inner cladding. The upper NA values of traditional double-clad fibers ($\sim$0.22 and $\sim$0.45 for glass and polymer outer claddings, respectively) are limited by the refractive indices of the available cladding materials.

For the air-clad fiber, the NA is determined by the ability of light to escape from the inner cladding through the silica webs. The NA of the air-clad fiber can be calculated by modeling the web as an infinitely long slab waveguide. As seen in **Fig. 11**, the NA increases as the webs become thinner or the wavelength of light becomes longer. To obtain NA similar to what is achieved by coating a traditional fiber with a low index polymer (NA$\sim$0.45), the web thickness must be about equal to the wavelength of the light. The NA of the air-clad fiber increases dramatically as the web thickness becomes less than half the wavelength of the guided light. Numerical apertures above 0.90 have been experimentally demonstrated.

The large NA values achievable with air-clad fibers are an important advantage for advanced double-clad amplifiers and lasers. A double-clad amplifier or laser (also known as cladding pumped amplifier or laser) contains a double-clad fiber with a rare earth doped core. High-power pump light, launched into and guided within the inner cladding, excites the rare earth ions as the pump light crosses through the core. The excited ions, which emit light with the same wavelength and phase as the signal, amplify the signal. The fiber's efficiency and the rare earth inversion level increase with the amount of pump light absorbed by the core. The light absorbed by the core, in turn, increases with the core-to-inner cladding area ratio and the pump light intensity in the inner cladding.

Traditional fibers have a practical limit to their core-to-cladding ratio. The maximum core diameter is limited by bend loss. And, while ultimately the minimum fiber diameter is limited by draw capabilities, in practice it is limited by the ability to couple pump light into a fiber having a limited inner-cladding NA. For a pump source of given brightness, the product of beam diameter and NA cannot be increased. Such light can be successfully coupled into a larger inner cladding with smaller NA or a smaller inner cladding with larger NA.

Air-clad fibers have other advantages over traditional double-clad fibers. The outer glass cladding in air-clad fibers is optically inactive. This allows the inner cladding diameter to be decoupled from the physical fiber diameter, eliminating draw-related

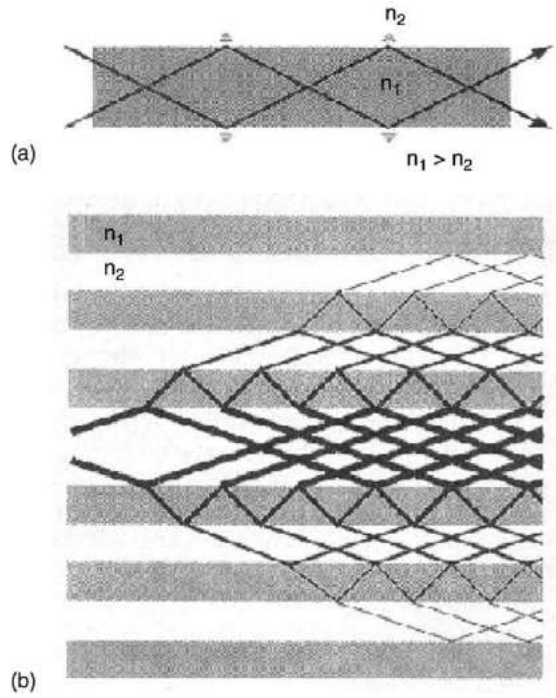


Fig. 12 Schematic of the mechanism of (a) index guided light and (b) band gap guided light. Reprinted with permission from Cregan RF, Mangan BJ, Knight JC, Birks TA, Russell PJ, Roberts PJ and Allan DC (1999) Single-mode photonic band gap guidance of light in air. *Science* 285. Copyright 1999 American Association for the Advancement of Science.

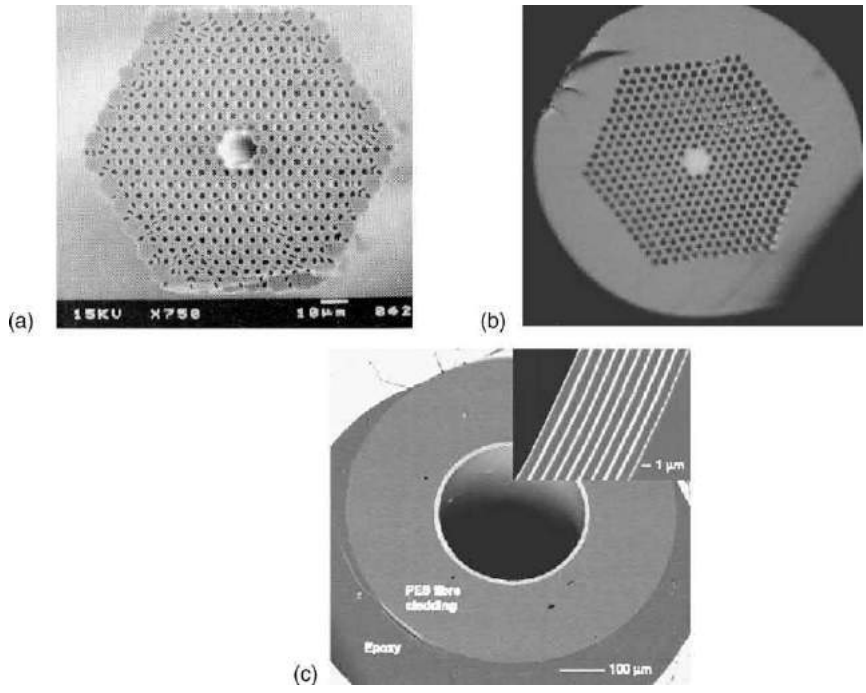


Fig. 13 Photographs and drawing of three types of bandgap fibers: (a) triangular array cladding with an air core (Cregan RF, Mangan BJ, Knight JC, Birks TA, Russell PJ, Roberts PJ and Allan DC (1999) Single-mode photonic band gap guidance of light in air. *Science* vol. 285), (b) tunable with a silica core, (c) concentric rings with an air core. Reprinted with permission from Temelkuran *et al. Nature* 420. Copyright 1999 American Association for the Advancement of Science.

constraints on minimum inner cladding size. Since only the inner cladding is optically active, its size can be optimized (made smaller) while keeping the total fiber diameter at the standard 125 microns. In addition, the pump light does not interact with the polymer coating, which can be important if the polymer properties are affected by the surroundings.

Bandgap Guided Fibers

Bandgap fibers guide light in the core (also referred to as a defect) by confining the light through constructive interference due to Bragg scattering (Fig. 12). Unlike traditional fibers, this mechanism allows light to propagate in a core that has an index lower than the cladding. Bandgap fibers with an air core could theoretically be very low loss, propagate high powers, have a large effective area core, and exhibit very low nonlinearities. In addition, the fibers can be used in novel devices by filling the core with special gasses or liquids for nonlinear processes.

Bandgap fibers can be generalized into three types (Fig. 13). The first type of fiber has an air core surrounded by a cladding that consists of a periodic array of air holes in silica. Fibers of this type have been designed to guide a single mode and hold the greatest promise of guiding light over long distances with very low loss. The second type of bandgap fiber consists of a silica core and a silica cladding with a periodic array of holes which are filled with a high-index, tunable liquid. This design allows the bandgap properties to be tuned and is of interest for use in devices. The third fiber type consists of rings of alternating high- and low-index dielectric material with a center air core. They typically have a large, multimode core which is ideal for sending high powers. In addition, very little light is in the dielectric materials so the fiber can be designed to guide any wavelength with relatively little loss.

The spectrum of a bandgap fiber is quite different than an index guided fiber (Fig. 14). Frequencies that lie within the bandgap cannot propagate in the cladding and are confined to the defect in the lattice (the core). Bandgap fibers with ideal geometry are predicted to have losses over an order of magnitude lower than can be obtained in an ideal traditional fiber. Although the loss of bandgap fibers has not been demonstrated to be less than traditional fiber at telecommunication wavelengths, bandgap fibers have been shown to have much lower loss than traditional fibers at other wavelengths.

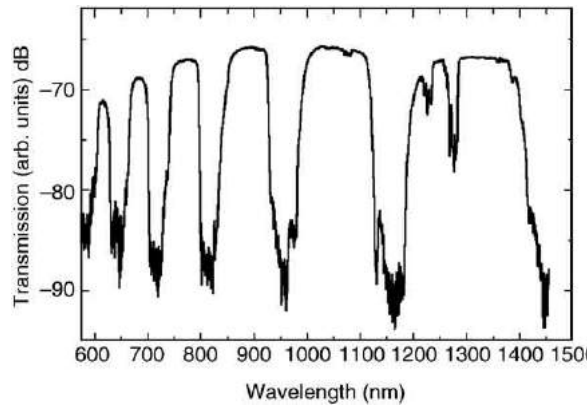


Fig. 14 Spectrum of a typical bandgap fiber.

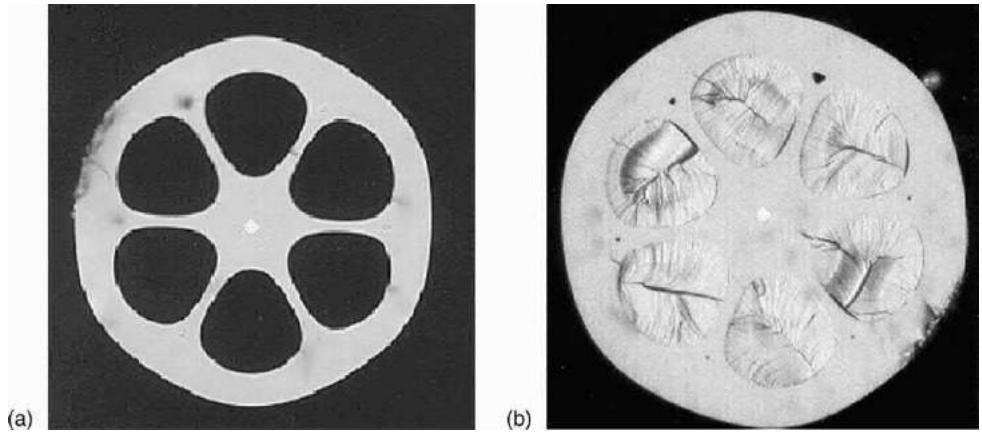


Fig. 15 Holes in the microstructure fibers are filled with controllable materials to affect the fiber properties actively. Photos show (a) original fiber and (b) fiber after polymer is inserted and cured in the microstructure fiber.

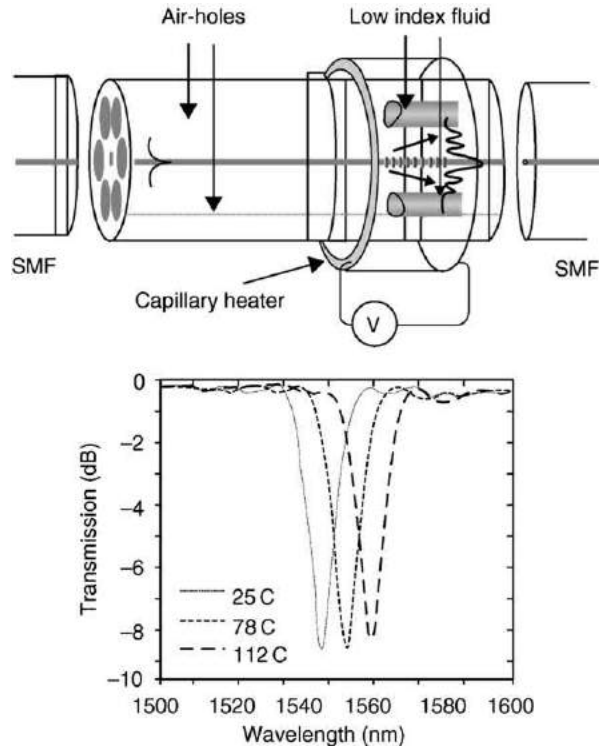


Fig. 16 Schematic of a tunable device in which low index liquid is positioned over the grating. The plot shows the shift in the spectrum as a function of temperature. Reproduced from Kerbage C, Windeler RS, Eggleton BJ, Dolinskoi M, Mach P and Rogers JA (2002) Tunable devices based on dynamic positioning of micro-fluids in microstructure optical fiber. *Optics Communications* 204: 179–184, with permission from Elsevier.

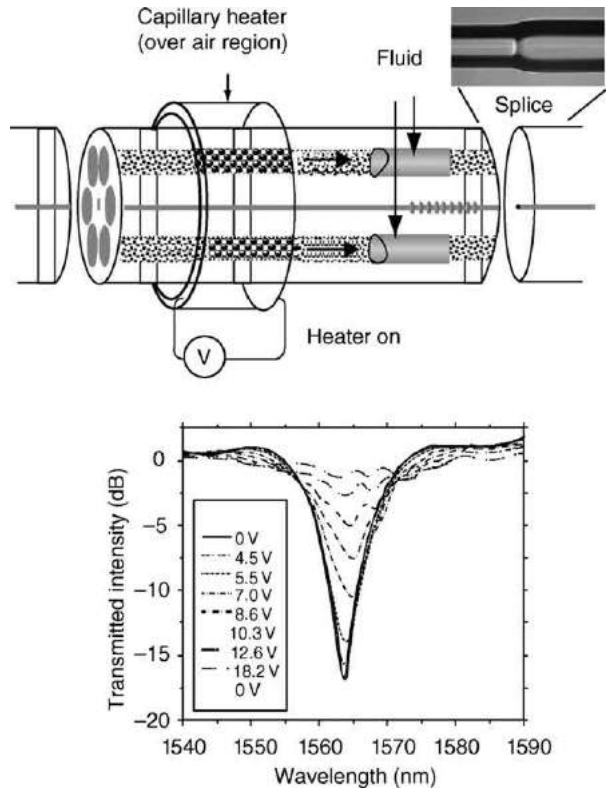


Fig. 17 Schematic of a tunable device in which a high index liquid plug is moved over the grating. The plot shows the intensity of the spectrum as a function of temperature. Reproduced from Kerbage C, Windeler RS, Eggleton BJ, Dolinskoi M, Mach P and Rogers JA (2002) Tunable devices based on dynamic positioning of micro-fluids in microstructure optical fiber *Optics Communications* 204: 179–184, with permission from Elsevier.

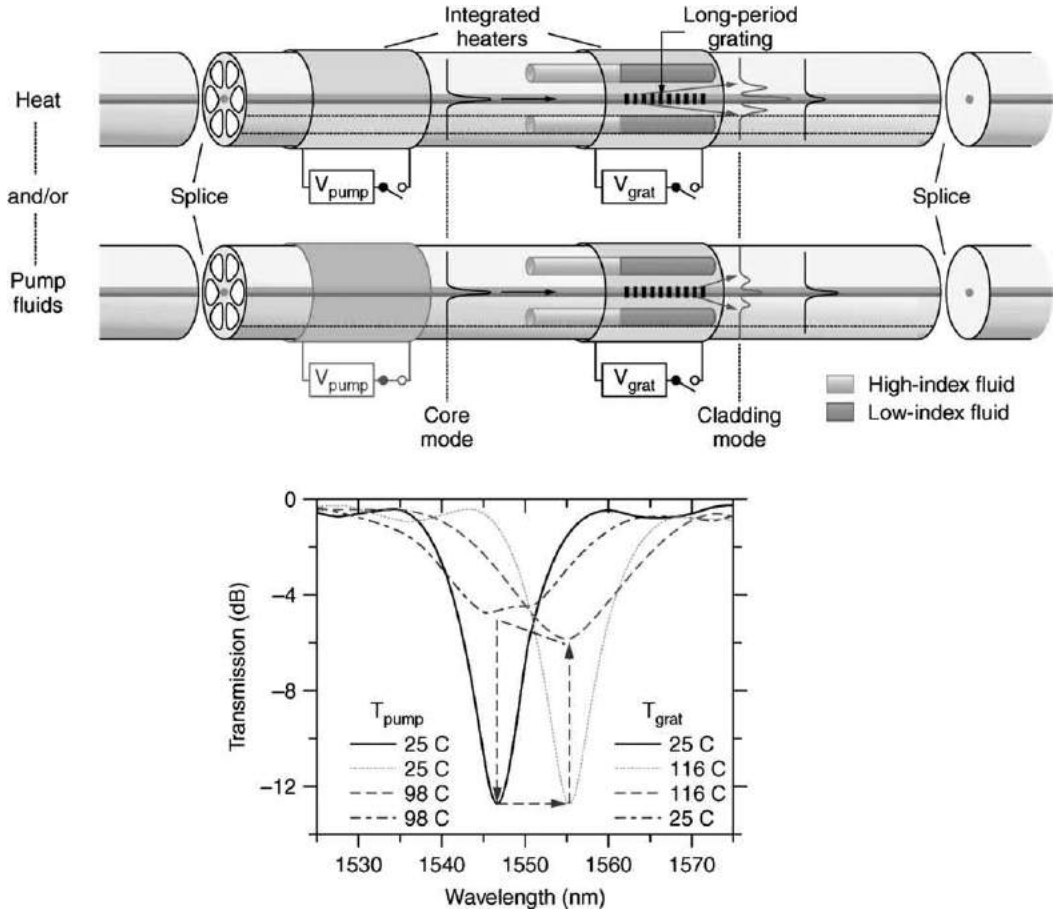


Fig. 18 Schematic of a tunable device using both high and low index liquids with two heaters. The plot shows that the position and strength of the spectrum are adjusted independently. Reproduced with permission from Mach P, Dolinski M, Baldwin KW, Rogers JA, Kerbage C, Windeler RS and Eggleton BJ (2002) Tunable microfluidic optical fiber. *Applied Physics Letters* 80(23). Copyright (2002) by the American Physical Society.

Tunable Microstructure Fiber Devices

Unique tunable devices are made by filling the air holes of microstructure fibers with materials whose properties can be controlled actively (Fig. 15). The materials can be positioned in the core or cladding to affect the corresponding modes. Most commonly, the index of refraction is changed using temperature sensitive liquids or polymers, although materials that respond to electric or magnetic fields enable a quicker response time. These material-filled active fibers can be used to fabricate devices like tunable filters, switches, broadband attenuators, and fibers with tunable birefringence. Below, a few examples of tunable filters are presented in which the device properties depend on the type and position of the liquid inserted in the fiber holes.

A microstructure tunable filter works by controlling the spectral position, shape, or strength of a longitudinal long-period grating written in the fiber. A grating is written in the fiber before filling the holes with tunable liquid. The position of the peak absorption wavelength is controlled by placing a liquid, with an index close to silica and strongly temperature dependent, in air holes in the cladding region over the grating (Fig. 16). The cladding index is tuned by varying the temperature of the liquid, using a small thin film heater located over the liquid. The position of the peak absorption wavelength moves as the cladding index is changed.

The strength or depth of the filter is controlled by inserting a moveable plug of high-index fluid in a sealed air section of a microstructure fiber (Fig. 17). The strength of the grating is determined by the fraction of the grating surrounded by high index fluid. The position of the liquid plug relative to the grating can be finely adjusted with a thin film heater located over the air region adjacent to the material plug. As heat is applied to the air, the air expands and pushes the liquid plug over the grating, decreasing the coupling of the fundamental mode to higher order cladding modes. Fig. 18 shows an example in which the position and strength of a filter are varied independently using two heaters and two adjacent materials.

See also: Fabrication of Optical Fiber

Further Reading

Birks, T.A., Knight, J.C., Russell, P., 1997. Endlessly single mode photonic crystal fiber. Optics Letters 22 (13),
Bjarklev A, Broeng J and Bjarklev AS, Photonic Crystal Fibers, Boston: Kluwer Academic Publishers.
Broeng, J., Mogilevstev, D., Bjarklev, A., Barkou, S.E., 1999. Photonic crystal fibers: A new class of optical waveguides. Optical Fiber Technology vol. 5.
Chandalia, J.K., Eggleton, B.J., Windeler, R.S., Kosinski, S.G., Liu, X., Xu, C., 2001. Adiabatic coupling in tapered air-silica microstructures optical fiber. IEEE Photonics Tech. Letters 13 (1),
Cregan, R.F., Mangan, B.J., Knight, J.C., Birks, T.A., Russell, P.J., Roberts, P.J., Allan, D.C., 1999. Single-mode photonic band gap guidance of light in air. Science 285.
Eggleton, B.J., Westbrook, P.S., White, C.A., Kerbage, C., Windeler, R.S., Burdge, G.L., 2000. Cladding-mode resonance in air-silica microstructure optical fiber. Journal Lightwave Technology 18 (8),
Joannopoulos, J.D., Meade, R.D., Winn, J.N., 1995. Photonic Crystals. New Jersey: Princeton University Press.
Johnson, S.G., Ibanesc, M., Skorobogatiy, M., Weisberg, M., Engeness, O., Soljacic, M., Jacobs, S.A., Joannopoubs, J.D., Fink, Y., 2001. Low-loss asymptotically single-mode propagation in large-core OmniGuide fibers. Optics Express 9 (13),
Kerbage, C., Hale, A., Yablon, A., Windeler, R.S., Eggleton, B.J., 2001. Integrated all fiber variable attenuator based on hybrid microstructure fiber. Applied Physics Letters 79 (19),
Kerbage, C., Windeler, R.S., Eggleton, B.J., Dolinskoi, M., Mach, P., Rogers, J.A., 2002. Tunable devices based on dynamic positioning of micro-fluids in microstructure optical fiber. Optics Communications 204, 179–184.
Knight, J.C., Birks, T.A., Russell, P.J., Atkin, D.M., 1996. All-silica single-mode optical fiber with photonic crystal cladding. Optics Letters 21 (19),
Knight, J.C., Birks, T.A., Russell, P.J., 2001. Holey' silica fibers. In: Markel, V.A., George, T.F. (Eds.), Optics of nanostructured materials. New York: John Wiley & Sons.
Mach, P., Dolinski, M., Baldwin, K.W., Rogers, J.A., Kerbage, C., Windeler, R.S., Eggleton, B.J., 2002. Tunable microfluidic optical fiber. Applied Physics Letters vol. 80.
Ranka, J.K., Windeler, R.S., Stentz, A.J., 2000. Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. Optics Letters 25 (1),
Ranka, J.K., Windeler, R.S., Stentz, A.J., 2000. Optical properties of high-delta air-silica microstructure optical fibers. Optics Letters 25, 796–798.

Omnidirectional Surfaces and Fibers

S Hart, G Benoit, and Y Fink, Massachusetts Institute of Technology, Cambridge, MA, USA

Omnidirectional Reflecting Mirrors

Mirrors are probably the most prevalent of optical devices. Known to the ancients and used by them as objects of worship and beauty, mirrors are currently employed for imaging, solar energy collection, and also in laser cavities. Their intriguing optical properties have captured the imagination of scientists as well as artists and writers.

One can distinguish between two types of mirrors, the age-old metallic, and the more recent multilayer dielectric. Metallic mirrors reflect light over a broad range of frequencies incident from arbitrary angles (i.e., omnidirectional reflectance). However, at infrared and optical frequencies a few percent of the incident power is typically lost due to absorption.

Multilayer dielectric mirrors are used primarily to reflect a narrow range of frequencies incident from a particular angle or particular angular range. Unlike their metallic counterparts, dielectric reflectors can be extremely low loss. The ability to reflect light of arbitrary angle of incidence for all-dielectric structures has been associated with the existence of a complete photonic bandgap, which can exist only in a system with a dielectric function that is periodic along three orthogonal directions. In fact, a sufficient condition for the achievement of omnidirectional reflection in a periodic system with an interface, is the existence of an overlapping bandgap regime in phase space above the light cone of the ambient media.

Consider a system that is constructed of alternating dielectric layers coupled to a homogeneous medium – characterized by n_0 (such as air with $n_0=1$), at the interfaces. Electromagnetic waves are incident upon the multilayer film from the homogeneous medium. While such a system was analyzed extensively in the literature the possibility of omnidirectional reflectivity was not recognized until recently.

The generic system is described by the index of refraction profile in **Fig. 1**, where h_1 and h_2 are the layer thickness, and n_1 and n_2 are the indices of refraction of the respective layers. The incident wave has a wavevector $\vec{k} = k_x\hat{e}_x + k_\gamma\hat{e}_\gamma$, and frequency of $\omega = c|k|$. The wavevector together with the normal to the periodic structure defines a mirror plane of symmetry which allows us to distinguish between two independent electromagnetic modes: transverse electric (TE) modes and transverse magnetic (TM) modes. For the TE mode the electric field is perpendicular to the plane, as is the magnetic field for the TM mode.

General features of the transport properties of the finite structure can be understood when the properties of the infinite structure are elucidated. In a structure with an infinite number of layers, translational symmetry along the direction perpendicular to the layers leads to Bloch wave solutions of the form:

$$u_K(x,y) = E_K(y)\, e^{iKy} e^{ik_x x} \tag{1}$$

where $E_K(y)$ is periodic, with a period of length a, and K is the Bloch wave number. These waves represent solutions to an eigenvalue problem and are completely and uniquely defined by the specification of K, k_x and ω.

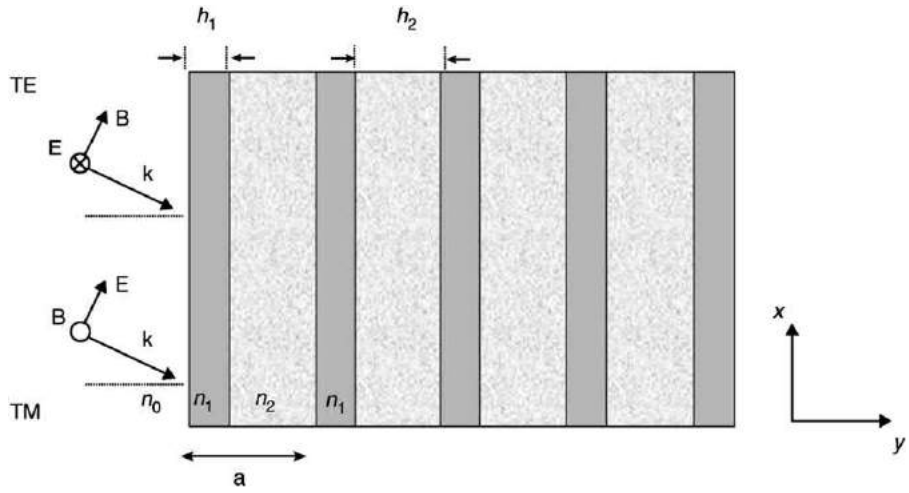


Fig. 1 Schematic of the multilayer system showing the layer parameters (n_α, h_α – index of refraction and thickness of layer α), the incident wavevector $\vec{k}$ and the electromagnetic mode convention.

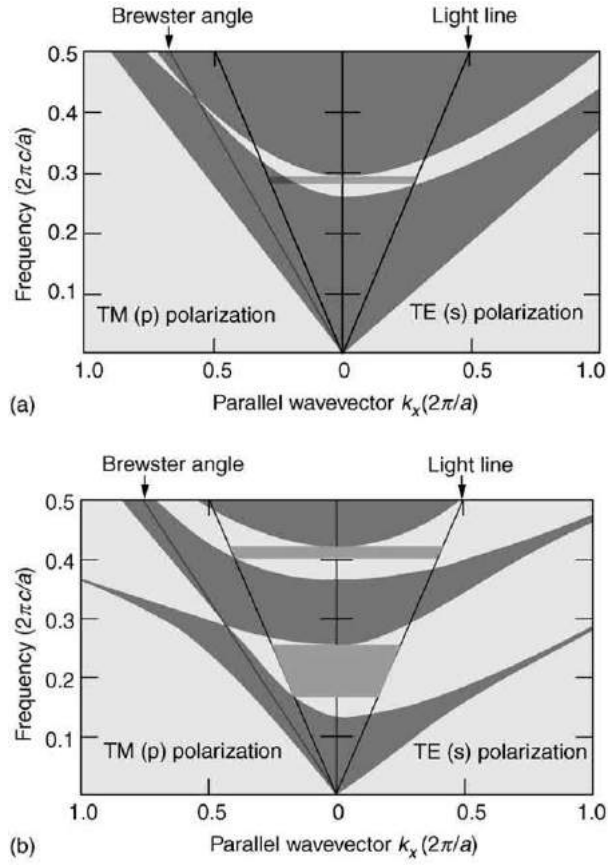


Fig. 2 (a) Projected bandstructure of a multilayer film with the light line and Brewster line, exhibiting a reflectivity range of limited angular acceptance with ($n_1 = 2.2$ and $n_2 = 1.7$ and a thickness ratio of $h_2/h_1 = 2.2/1.7$). (b) Projected bandstructure of a multilayer film together with the light line and Brewster line, showing an omnidirectional reflectance range at the first and second harmonic (propagating states – dark gray, evanescent states – white, omnidirectional reflectance range – light grey). The film parameters are $n_1 = 4.6$ and $n_2 = 1.6$ with a thickness ratio of $h_2/h_1 = 1.6/0.8$. These parameters are similar to the actual polymer–tellurium film parameters measured in the experiment. Reproduced with permission from Fink Y, Winn JN, Fan S, *et al.* (1998) A dielectric omnidirectional reflector. *Science* 282: 1679–1682. Copyright 1998 American Association for the Advancement of Science.

Solutions can be propagating or evanescent, corresponding to real or imaginary Bloch wave numbers, respectively. It is convenient to display the solutions of the infinite structure by projecting the ω (K, k_x) function onto the $\omega - k_x$ plane; **Fig. 2(a)** and **(b)** are examples of such projected structures.

The gray background areas highlight phase space where K is strictly real, that is, regions of propagating states, while the white areas represent regions containing evanescent states. The shape of the projected bandstructures for the multilayer film can be understood intuitively. At $k_x = 0$, the bandgap for waves traveling normal to the layers is recovered. For $k_x > 0$, the bands curve upward in frequency. As $k_x \to \infty$, the modes become largely confined to the slabs with the high index of refraction and do not couple between layers (and are therefore independent of k_x).

In a finite structure, the translational symmetry in the directions parallel to the layers is preserved, hence k_x remains a conserved quantity and can be used to label solutions even though these solutions will no longer be of the Bloch form. The relevance of the band diagram to finite structures is that it allows for the prediction of regions of phase space where waves are evanescently decaying in the multilayer structure.

Since we are primarily interested in waves originating from the homogeneous medium external to the periodic structure, we will focus only on the portion of phase space lying above the light line. Waves originating from the homogeneous medium satisfy the condition $\omega \geq ck_x/n_0$, where n_0 is the refractive index of the homogeneous medium, and therefore they must reside above the light line. States of the homogeneous medium with $k_x = 0$ are normal incident, and those lying on the $\omega = ck_x/n_0$ line with $k_y = 0$ are incident at an angle of 90°.

The states in **Fig. 2(a)**, that are lying in the restricted phase space defined by the light line and that have a (ω, k_x) corresponding to the propagating solutions (gray areas) of the crystal can propagate in both the homogeneous medium and in the structure. These waves will partially or entirely transmit through the film. Those with (ω, k_x) in the evanescent regions (white areas) can propagate in the homogeneous medium but will decay in the crystal – waves corresponding to this portion of phase space will be reflected off the structure.

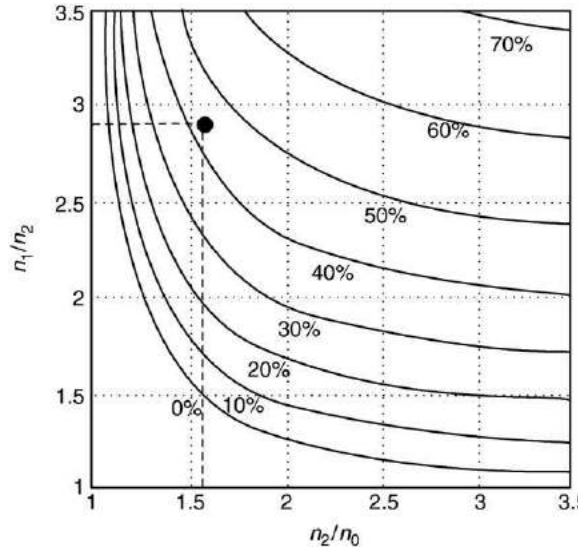


Fig. 3 The range to mid-range ratio $(\omega_h - \omega_l)/\frac{1}{2}(\omega_h + \omega_l)$, for the fundamental frequency range of omnidirectional reflection, plotted as contours. Here, the layers were set to quarter wave thickness and $n_1 > n_2$. The ratio for our materials is approximately 45%. ($n_1/n_2 = 2.875$, $n_2/n_0 = 1.6$) is located at the intersection of the dashed lines (black dot). Reproduced with permission from Fink Y, Winn JN, Fan S, *et al.* (1998) A dielectric omnidirectional reflector. *Science* 282: 1679–1682. Copyright 1998 American Association for the Advancement of Science.

The multilayer system leading to **Fig. 2(a)** represents a structure with a limited reflectivity cone, since for any frequency one can always find a k_x vector for which a wave at that frequency can propagate in the crystal – and hence transmit through the film. The necessary and sufficient criterion for omnidirectional reflectivity at a given frequency is that there exists no transmitting state of the structure inside the light cone – this criterion is satisfied by frequency ranges marked in light gray in **Fig. 2(b)**. In fact, the system leading to **Fig. 2(b)** exhibits two omnidirectional reflectivity ranges.

The omnidirectional range is defined from above by the normal incidence bandedge ω_h ($k_y = \pi/a$, $k_x = 0$) (**Fig. 2(b)**) and below by the intersection of the top of the TM allowed bandedge with the light line $\omega_l(k_y = \pi/a, k_x = \omega_l/c)$ (**Fig. 2(b)**).

A dimensionless parameter used to quantify the extent of the omnidirectional range is the range to mid-range ratio defined as $(\omega_h - \omega_l)/\frac{1}{2}(\omega_h + \omega_l)$.

Fig. 3 is a plot of this ratio as a function of n_2/n_1 and n_1/n_0, where ω_h and ω_l are determined by solutions of **Eq. (2)** with quarter wave layer thickness. The contours in this figure represent various equi-omnidirectional ranges for different material index parameters and could be useful for design purposes.

At normal incidence, there is no distinction between TM and TE modes. At increasingly oblique angles, the gap of the TE mode increases, whereas the gap of the TM mode decreases. In addition, the center of the gap shifts to higher frequencies. Therefore, the criterion for the existence of omnidirectional reflectivity can be restated as the occurrence of a frequency overlap between the gap at normal incidence and the gap of the TM mode at 90°. Analytical expressions for the range to mid-range ratio can be obtained by setting:

$$\omega_h = \frac{2c}{h_2 n_2 + h_1 n_1} \cos^{-1}\left(-\left|\frac{n_1 - n_2}{n_1 + n_2}\right|\right) \omega_l = \frac{2c}{h_2\sqrt{n_2^2 - 1} + h_1\sqrt{n_1^2 - 1}} \times \cos^{-1}\left(\left|\frac{n_1^2\sqrt{n_2^2 - 1} - n_2^2\sqrt{n_1^2 - 1}}{n_1^2\sqrt{n_2^2 - 1} + n_2^2\sqrt{n_1^2 - 1}}\right|\right) \tag{2}$$

Moreover, the maximum range width is attained for thickness values that are not equal to the quarter wave stack though the increase in bandwidth gained by deviating from the quarter wave stack is typically only a few percent.

In general, the TM mode defines the lower frequency edge of the omnidirectional range; an example can be seen in **Fig. 2(b)** for a particular choice of the indices of refraction. This can be proven by showing that:

$$\left.\frac{\partial\omega}{\partial k_y}\right|_{TM} \geq \left.\frac{\partial\omega}{\partial k_y}\right|_{TE} \tag{3}$$

in the region that resides inside the light line. The physical reason for **Eq. (3)** lies in the vectorial nature of the electric field. In the upper portion of the first band the electric field concentrates its energy in the high dielectric regions. Away from normal incidence, the electric field in the TM mode has a component in the direction of periodicity, this component forces a larger portion of the electric field into the low dielectric regions. The group velocity of the TM mode is therefore enhanced. In contrast, the electric field of the TE mode is always perpendicular to the direction of periodicity and can concentrate its energy primarily in the high dielectric region. While the omnidirectional reflection criteria can be used to confine light in various geometries, it is the cylindrical fiber configuration that appears to present significant application opportunities.

Mesostructured Fibers for External Reflection Applications

Polymer fibers are ubiquitous in applications such as textile fabrics, due to their excellent mechanical properties and the availability of low-cost, high-volume processing techniques; however, the control over their optical properties has so far remained relatively limited. Conversely, dielectric mirrors are used to precisely control and manipulate light in high performance optical applications, but the fabrication of these typically fragile mirrors has been mostly restricted to planar geometries and typically involves multiplicity of deposition sequences in a high vacuum thin film deposition system.

Fabrication Approach

The fabrication of extended lengths of omnidirectionally reflecting fibers with large bandgaps and high layer counts poses considerable challenges. To illustrate the nature of this formidable task, one needs to consider the necessity of maintaining the uniformity of sub-100 nm layer thicknesses over kilometer-length scales; creating continuous layers with an aspect ratio of $\sim 10^{10}$ in a single process! To meet this and other challenges associated with the fabrication of mesostructured fibers, we have developed a preform-based fabrication approach (Fig. 4). A scaled-up version of the final fiber, called a preform, is fabricated which shares the geometry and materials of the final fiber but exhibits macroscopic lateral features. The preform is heated up and drawn under tension into the fiber using a simple cylindrical furnace. The macroscopic layers are reduced in the process to microscopic dimensions while maintaining the overall geometry and symmetry of the original preform. Conservation of mass determines the final length of the fiber – typically this length is equal to the lateral reduction factor squared multiplied by the length of the preform. The nature of the process and requirements on the fiber's optical properties lead to the definition of materials selection rules. First, in order to achieve omnidirectional reflectivity, one needs to identify two solid materials exhibiting an index contrast given by the plot in Fig. 3. Second, to enable the codrawing of the two dissimilar materials both will need to have viscosities that are lower than $\sim 10^8$ poise at the drawing temperature. In order to maintain high draw speeds, the majority component needs to be amorphous and the adhesion between these two materials needs to be sufficient to prevent delamination. Finally, their thermal expansion coefficients need to be close or alternatively at least one of the materials needs to be capable of relieving the stress due to CTE mismatch.

Materials Selection Criteria

Pairs of materials that are compatible with the process and fiber property requirements have been identified, including: high glass transition temperature (T_g), thermoplastic polymers such as poly(ether-sulfone) (PES), poly (ether-imide) (PEI) and members of the chalcogenide glass family, such as arsenic triselenide (As_2Se_3) or arsenic trisulfide. Pairs of these materials will have substantially different refractive indices, as shown in Fig. 5, which is a measurement of the real and imaginary indices of refraction of PES and As_2Se_3, obtained using a broadband spectroscopic ellipsometer (SOPRA GES5). They nevertheless exhibit similar thermomechanical properties within a certain thermal processing window.

Adhesion and extensional viscosity in the fluid state are difficult to measure in general, and the measurement of high-temperature surface tension is quite involved. Thus, limited data on these properties are available and it was necessary to empirically identify materials that could be used to draw out mirror-fibers. Various high-index chalcogenide (S, Se, and Te containing) glasses and low-index polymers were identified as potential candidates based on their optical properties and overlapping thermal softening regimes. Adhesion and viscosity matching were tested by thermal evaporation of a chalcogenide glass layer on top of a polymer film or rod and elongation of the coated substrate at elevated temperatures. The choice of a high-temperature polymer, PES, and a simple chalcogenide glass, As_2Se_3, resulted in excellent thermal co-deformation without film cracking or delamination. Approximate matching of extensional viscosity in this manner was also demonstrated using As_2Se_3 and PEI. The properties, processing, and applications of chalcogenide glasses have been explored extensively elsewhere. One advantage in choosing As_2Se_3 for this application is that not only is it a stable glass, but it is a stoichiometric compound that can be readily deposited in thin films through thermal evaporation or sputtering without dissociation. Additionally, As_2Se_3 is transparent to IR

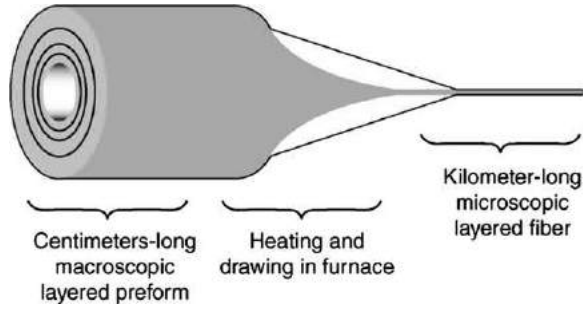


Fig. 4 Conceptual preform based fabrication process for meso-structured fibers.

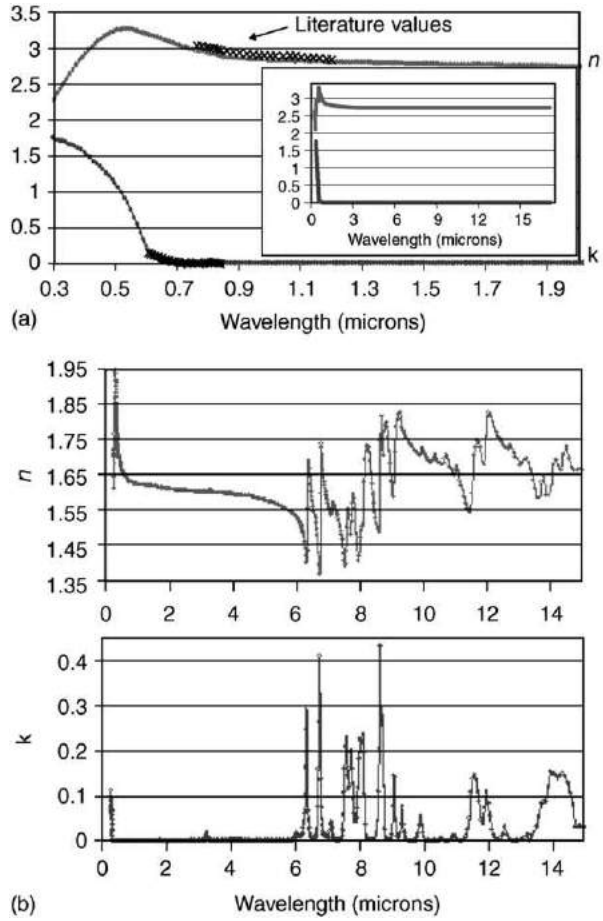


Fig. 5 (a) Real part (n – light gray) and imaginary part (k – dark gray) of the refractive index of annealed As_2Se_3. The black crosses correspond to literature values. (b) Real part (n – upper) and imaginary part (k – lower) of the refractive index of PES.

radiation from approximately 0.8 to 17 μm as shown in the ellipsometric data (Fig. 5(a)), and has a refractive index of ~2.8 in the mid-IR. PES is a high-performance, dimensionally stable thermoplastic with a refractive index of ~1.6 and good transparency to EM waves in a range extending from the visible regime into the mid-IR (~5 μm), as shown in Fig. 5(b).

Preform Construction Process and Fiber Draw

The selected materials were used to construct a multilayer preform rod, which essentially is a macroscale version of the final fiber. In order to fabricate the dielectric mirror fiber preform, an As_2Se_3 film was deposited through thermal evaporation on either side of a free-standing PES film which was then rolled on top of a PES tube substrate, forming a structure having 21 alternating layers of PES and As_2Se_3, as shown in Fig. 6.

The resulting multilayer fiber preform was subsequently drawn down using an optical fiber draw tower into hundreds of meters of multilayer fiber with a precisely controlled submicron layer thickness, creating a photonic bandgap in the mid-IR. Fibers of outer diameters varying from 175–500 μm with a typical standard deviation of 10 μm from target, were drawn from the same preform to demonstrate adjustment of the reflectivity spectra through thermal deformation. The spectral position of the photonic bandgap was controlled by the optical monitoring of the outer diameter (OD) of the fiber during draw, which was later verified by reflectivity measurements on single and multiple fibers of different diameters. Scanning electron micrographs (SEMs) of the cross-section of these fibers are depicted in Fig. 7.

Bandstructure for Multilayer Fibers for External Reflection Applications

In theoretically predicting the spectral response of these fibers, it is helpful to calculate the photonic bandstructure that corresponds to an infinite one-dimensional photonic crystal. This allows for the analysis of propagating and evanescent modes in the structure, corresponding to real or imaginary Bloch wave number solutions.

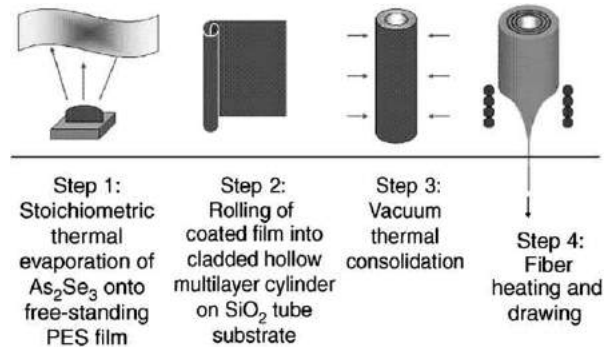


Fig. 6 Multilayer preform fabrication sequence.

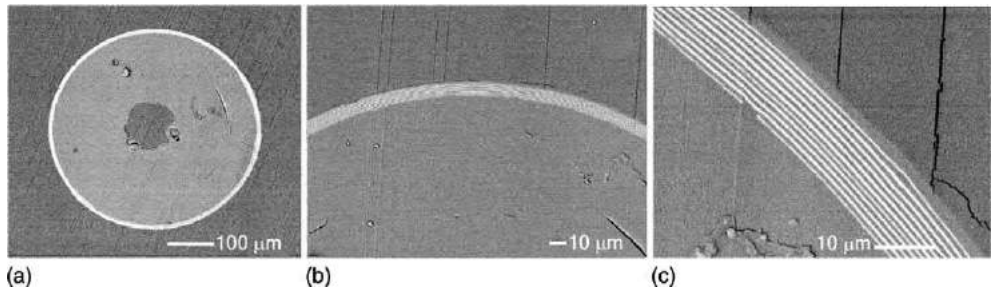


Fig. 7 SEM micrographs of 400 μm OD fiber cross-section. The entire fiber is embedded in epoxy. (a) shows the entire fiber cross-section, with mirror structure surrounding the PES core; (b) demonstrates that the majority of the fiber exterior is free of significant defects and that the mirror structure adheres well to the fiber substrate; and (c) reveals the ordering and adhesion within the alternating layers of As_2Se_3 (bright layers) and PES. Stresses developed during sectioning caused some cracks in the mounting epoxy that are deflected at the fiber interface. Fibers from this batch were used in the reflectivity measurements recorded below in **Fig. 9(a)**. Reproduced with permission from Hart SD, Maskaly GR, Temelkuran B, Prideaux PH, Joannopoulos JD and Fink Y (2002) External reflection from omnidirectional dielectric mirror fibers. *Science* 296: 510–513. Copyright 2002 American Association for the Advancement of Science.

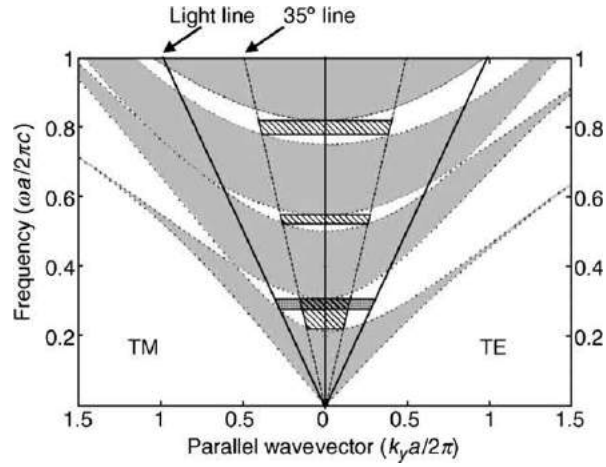


Fig. 8 Photonic band diagram for a one-dimensional photonic crystal having a periodic refractive index alternating between 2.8 and 1.55. Gray regions represent propagating modes within the structure, while white regions represent evanescent modes. Hatched regions represent photonic bandgaps where high reflectivity can be expected for external EM waves over an angular range extending from normal to 35° incidence. The lower dark shaded trapezoid represents a region of external omnidirectional reflection.

The electric or magnetic field vector is parallel to the mirror layer interfaces for the TE and TM polarized modes, respectively. The parallel wavevector (k_y) is the component of the incident electromagnetic (EM) wavevector that is parallel to the layer interfaces. The phase space accessible from an external ambient medium is contained between the light lines (defined by the glancing-angle condition $\omega = ck_y/n_0$), and the modes between the 35° lines correspond to those sampled experimentally. Axes are normalized to the thickness a of one mirror bilayer (a period consisting of one high- and one low-index layer). **Fig. 8** depicts

the photonic band diagram for an infinite structure having similar periodicity and refractive indices to the mirror structures fabricated here. Three photonic bandgaps are present where high reflectivity is expected within the 0–35° angular range, and the fundamental gap contains a region of external omnidirectional reflectivity.

Optical Characterization of 'Mirror Fibers'

Mirror fiber reflectivity was measured from both single fibers and parallel fiber arrays using a Nicolet/SpectraTech NicPlan Infrared Microscope and Fourier Transform Infrared Spectrometer (Magna 860). The microscope objective (SpectraTech 15×, Reflachromat) used to focus on the fibers had a numerical aperture (NA) of 0.58. This results in a detected cone where the angle of reflection with respect to the surface normal of the structure could vary from normal incidence to ~35°, which is determined by the NA of the microscope objective. As a background reference for the reflection measurements, we used gold-coated PES fibers of matching diameters. Dielectric mirror fibers, drawn to 400 μm OD, exhibited a very strong reflection band centered at 3.4 μm wavelength (Fig. 9(a)). Measured reflectivity spectra agree well with planar-mirror transfer matrix method (TMM) simulations, where the reflectivity was averaged across the aforementioned angular range for both polarization modes. Fibers drawn down to 200 μm OD show a similar strong fundamental reflection band centered near 1.7 μm (Fig. 9(b)). This shifting of the primary photonic bandgap clearly illustrates the precise tuning of the reflectivity spectra over wide frequency ranges through thermal deformation processing. Strong optical signatures are measurable from single fibers as small as 200 μm OD. Fiber array measurements, simultaneously sampling reflected light from multiple fibers, agree quite well with single-fiber data (Fig. 9(b)).

These reflectivity results are strongly indicative of uniform layer thickness control, good interlayer adhesion, and low interdiffusion through multiple thermal treatments. This was confirmed by SEM inspection of fiber cross-sections (Fig. 7). The layer thicknesses observed ($a=0.90$ μm for the 400 μm fibers; $a=0.45$ μm for the 200 μm fibers) correspond well to the measured reflectivity spectra. The fibers have a hole in the center, due to the choice of a hollow rod as the preform substrate, which experienced some nonuniform deformation during draw. The rolled-up mirror structure included a double outer layer of PES for mechanical protection, creating a noticeable absorption peak in the reflectivity spectrum at ~3.2 μm (Fig. 9(a)).

A combination of spectral and direct imaging data demonstrates excellent agreement with the photonic band diagram. The measured gap width (range to mid-range ratio) of the fundamental gap for the 400 μm OD fiber is 27%, compared to 29% in the photonic band diagram.

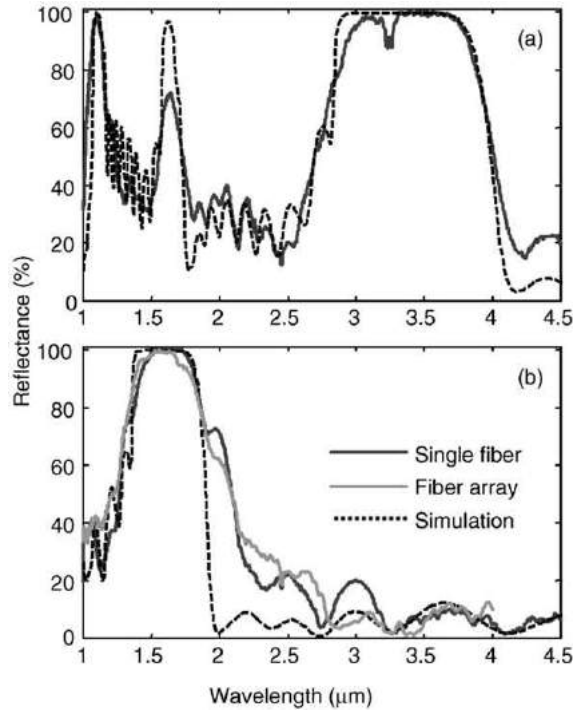


Fig. 9 Measured reflectance spectra for 400 μm OD (a) and 200 μm OD (b) dielectric mirror fibers relative to gold-coated fibers of the same diameter. (a) shows a single-fiber reflectivity measurement, while (b) compares single-fiber reflectivity to that measured from a multifiber array. Simulations were performed using the transfer matrix method.

Tunable 'Fabry–Perot' Fibers

The fabrication of fibers surrounded by or lined with alternating layers of materials with a large disparity in their refractive indices, presents interesting opportunities for passive and active optical devices. While a periodic multilayer structure, such as the one reported above, leads to the formation of photonic bandgaps and an associated range of high reflectivity, it is the incorporation of intentional deviations from periodicity, also called 'defects', which allows for the creation of localized electromagnetic modes in the vicinity of the defect. These structures, sometimes called optical cavities, in turn can provide the basis for a large number of interesting passive and active optical devices such as vertical cavity surface emitting lasers (VCSELs), bi-stable switches, tuneable dispersion compensators, tuneable drop filters, etc. Here we report on the fabrication of a fiber surrounded by a Fabry–Perot cavity structure and demonstrate that by application of axial mechanical stress, the spectral position of the resonant Fabry–Perot mode can be reversibly tuned.

Structure and Optical Properties of the Fabry–Perot Fibers

The fabrication technique described above allows for the accurate placement of optical cavities that can encompass the entire or partial fiber circumference. The fibers discussed above are made of As_2Se_3 and PES and have a low index Fabry–Perot cavity (**Fig. 10**).

This structure was achieved by introducing an extra polymer layer in the middle of the periodic multilayer structure of the preform, thus generating a defect mode in the photonic bandgaps of the drawn fibers. The position of the bandgap center is linearly related to the optical thickness of the layers by the Bragg condition. Through accurate outer diameter control afforded by the laser micrometer mounted on the draw tower we have been able to place gaps (**Fig. 11**) at wavelengths ranging from

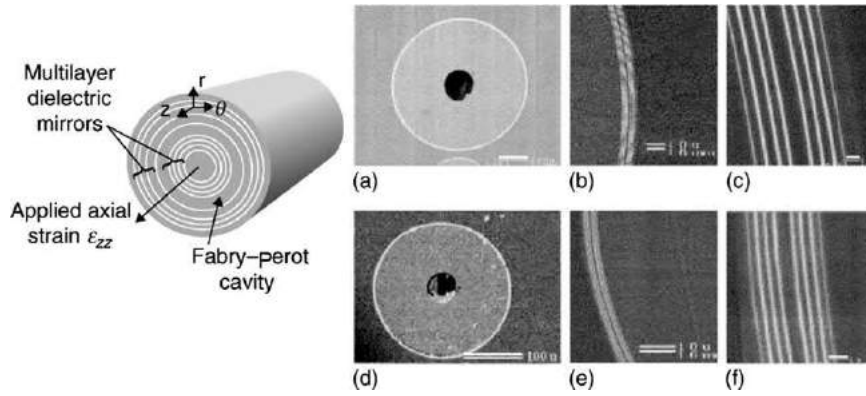


Fig. 10 Schematic of the structure of dielectric mirror fibers made of As_2Se_3 (light gray) and PES (dark gray) with low index Fabry–Perot cavity. The local cylindrical coordinate system is represented as well as the applied axial strain ϵ_{zz}. Typical radii are $\sim$150 μm$\pm$100 μm. Backscattered SEM micrographs of the cross-sections of a 460 micron diameter fiber (a, b, c) and of a 240 micron diameter fiber (d, e, f) embedded in epoxy and microtomed. (a) and (d) show the entire cross-section of the fibers, (b) and (e) demonstrate long-range layer uniformity, and (c) and (f) reveal the ordering and adhesion of the Fabry–Perot cavity structure. Bar scales have been redrawn for clarity. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

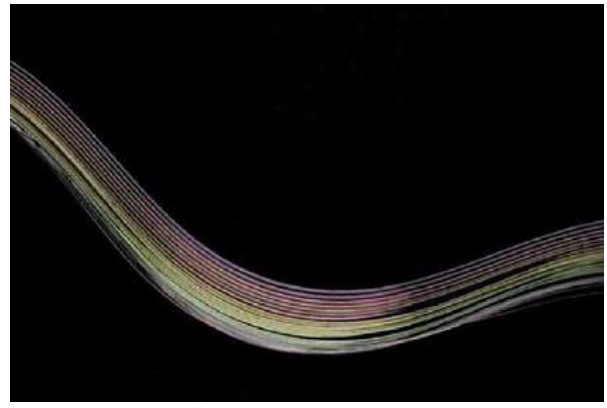


Fig. 11 Array of parallel fibers with outer diameter ranging from $\sim$420 μm (bottom) to $\sim$100 μm (top). The colors are due to the narrow 4th order photonic bandgap in the visible.

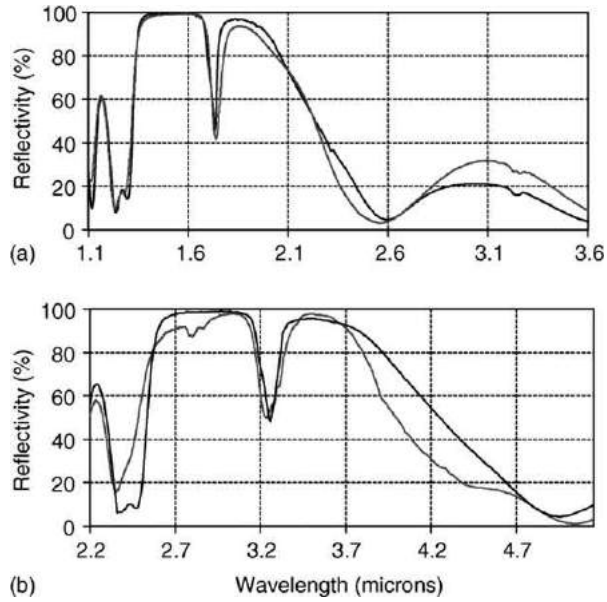


Fig. 12 Computed (black lines) and measured (grey lines) reflectivity spectra for the 240 micron (a) and of the 460 micron diameter fibers (b) with Fabry–Perot resonant modes at 1.74 and 3.2 μm, respectively. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

11 microns to below 1 micron. Such cost-effective tuneable optical filters could lead to applications such as optical switches for wavelength-division-multiplexing (WDM) systems and sensors.

The cross-sectional structures of a 240 micron and a 460 micron diameter fiber were observed by an SEM using a backscattered electron detector (Fig. 10). The structure is composed of a hollow core polymer rod surrounded by six bilayers of As_2Se_3 and PES, separated in the middle by an extra polymer layer, which forms the Fabry–Perot cavity. An extra polymer layer protects the fiber surface. For the 240 micron (460 micron) diameter fiber, the glass layers are ~135 nm (250 nm) thick, except for the first and the last ones which are half as thick due to the fabrication technique; the polymer layers are ~270 nm (540 nm) thick and the defect layer is ~610 nm (1170 nm) thick.

Reflectivity spectra measurements were performed under a microscope using a (Nicolet SpectraTech NicPlan) Infrared Microscope and Fourier Transform Infrared Spectrometer (Magna 860) with a lens numerical aperture (NA) corresponding to 30 degrees of angular spread. They exhibited a single-mode Fabry–Perot resonant mode at 1.74 and 3.2 μm for the 240 and 460 micron diameter fiber, respectively (Fig. 12).

Because of the range of incident angles, the measured quality factor (Q-factor, defined as its spectral position divided by the full width half maximum (FWHM)) was equal to 31. Using the different thicknesses and the real and imaginary part of the refractive index ($n\&k$) of As_2Se_3 and PES carefully measured with a broadband (300 nm to 16 microns) spectroscopic ellipsometer (Sopra GES-5), the reflectivity spectra of these fibers were computed with the TMM by approximating them as a one-dimensional planar stacking. By averaging the calculated spectra over the accessible incident angles and polarizations, we obtained a very good agreement between the simulation and the measurements.

Simulation of the Opto-Mechanical Behavior of the Fabry–Perot Fibers

We focused our analysis on the elastic regime, which ultimately limits the operational range of these fibers. The fiber's Young's modulus (E) can be approximated by modeling this multilayer structure as independent parallel springs under axial (along the z-axis of the fiber) strain assumed to be equal for all the layers, leading to a Young's modulus of 2.64 GPa (using $E_{PES}=2.4$ GPa for bulk PES and $E_{As_2Se_3}=15$ GPa reported for 1.5 μm thick films).

Neglecting the possible strain-induced refractive index variation of the materials, the normalized shift of the Fabry–Perot resonant mode can be related to the applied axial strain by calculating the radial stresses σ_{rr} (resulting from the difference between the Poisson ratios of the materials – $\nu_{PES}=0.45$ and $\nu_{As_2Se_3}=0.289$ – and the adhesion condition between the layers) and displacements u_r in each layer under axial strain. Starting from the equilibrium equations:

$$\frac{\partial \sigma_{zz}}{\partial z}=0 \tag{4}$$

$$\frac{\partial \sigma_{rr}}{\partial r}+\frac{\sigma_{rr}-\sigma_{\theta\theta}}{r}=0 \Leftrightarrow \frac{d^2u_r}{d^2r}+\frac{1}{r}\frac{du_r}{dr}-\frac{u_r}{r^2}=0 \tag{5}$$

and using Hooke's law and Lame's equations, we can derive a general expression for u_r and σ_{rr}, with two unknowns A and B per layer:

$$u_r = \frac{A}{r} + \mathrm{Br} \tag{6}$$

$$\sigma_{\mathrm{rr}} = -\frac{2\mathrm{GA}}{r^2} + 2(\eta + G)B + \lambda\varepsilon_{zz} \tag{7}$$

where $G=E/[2(1+v)]$ is the shear modulus and $\eta=Ev/[(1+v)(1-2v)]$ the Lame modulus. The continuity of the displacement and the radial stress at each interface, plus the boundary conditions (σ_{rr} vanishes at free surfaces) can be expressed as a linear system whose unique solution allows us to relate the radial strain in each layer to the applied axial strain as a linear relation $\varepsilon_{rr}=C\varepsilon_{zz}$, where C can be interpreted as an effective Poisson ratio. Finally, taking into account that the Fabry–Perot resonant mode itself is linearly shifted within the bandgap because of the different effective Poisson ratios of the glass and polymer layers, we obtain a linear relation between the normalized shift of the Fabry–Perot resonant mode and the applied axial strain:

$$\frac{\Delta\lambda}{\lambda} = -0.373\varepsilon_{zz} \tag{8}$$

All the layers (except the outer protective polymer layer) are under tensile radial stress whose maximum (0.22 MPa under 1% axial strain) is located at the interface between the layers and the polymer core where delamination is most likely to occur.

Mechanical Tuning Experiment and Discussion

Measurements were performed on fibers ~30 cm long, which were fixed at one end with epoxy to a load cell (Transducer Techniques MDB-2.5) while the other end was attached with strong tape to a pole mounted on a stepper rotational stage (Newport PR50) (Fig. 13). This end of the fiber was also screwed to the pole to further secure it in place.

The diameter of the pole was equal to 2.1 cm, leading to a normalized shift precision below 0.005%. The uncertainty on the Young's modulus, due to the precision of the load cell, was lower than 30 MPa. All the reflectivity spectra were normalized to a background taken with a flat gold mirror. The measurements were realized as far as possible from the fixed ends of the fiber where edge effects are likely to occur. These edge effects result in a reduction of the length of the fiber that deforms uniformly and consequently increase the real strain far from the edges by a factor of 1.14 and 1.15 for the 240 and the 460 micron diameter fiber, respectively (determined experimentally by measuring the position of two reference points on the fiber) compared to the strain calculated from the rotation of the pole. Moreover, to avoid measuring slight variations in the spectral position of the bandgap resulting from outer diameter variations, typically of the order of 4–5 microns over meters of fibers, the measurements were realized at a fixed reference position on the fiber.

By focusing on the fundamental bandgap of the 240 and the 460 micron diameter fiber, we demonstrated the tuning of the Fabry–Perot resonant mode under increasing axial strain (Fig. 14).

A drop in the reflectivity of 13% was observed at 1.71 μm (dash line) for the 240 micron diameter fiber when increasing the applied axial strain from 0.23% (light gray) to 1.07% (dark gray). The normalized-shift versus strain curves (Fig. 15) appeared to be linear for both fibers up to approximately 0.9% (dash line) with a slope equal to −0.3859 and −0.3843 for the 240 and the 460 micron diameter fiber, respectively, close to the predicted value (−0.373, Eq. (8)).

The normalized shift was equal to −0.347% for 0.9% applied axial strain, which seems to be the limit of the elastic regime and corresponds to small applied loads: 76 and 300 g for the 240 and 460 micron diameter fiber, respectively. Under higher strains, the normalized-shift versus strain curves were no longer linear and the measured loads would start decreasing slightly with time, which could be a sign of delamination between the layers and the polymer core, of plastic deformation and/or of relaxation in the layers. The stress–strain curves also exhibit an elastic regime up to approximately 1% strain with a corresponding Young's modulus equal to 2.59 GPa for the 240 μm OD fiber and to 2.39 GPa for the 460 μm OD fiber. These values are slightly lower than the predicted value, possibly due to relaxation effects in the polymer and the glass layers (for As_2Se_3 and PES, T_g=175 and 220°C, respectively).

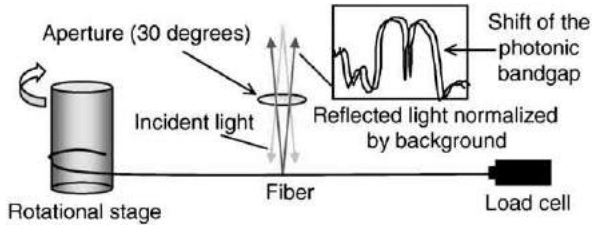


Fig. 13 Experimental setup used for mechanical tuning demonstration. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

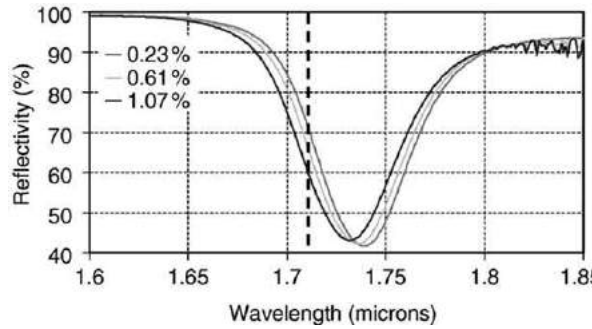


Fig. 14 Reflectivity versus wavelength plot showing the shift of the Fabry–Perot resonant mode of the 240 micron diameter fiber for three increasing values of the applied axial strain. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

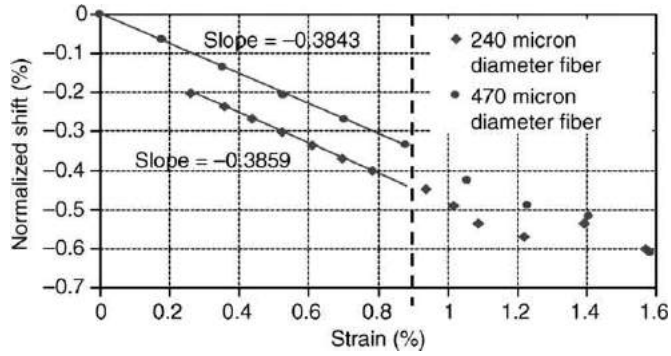


Fig. 15 Normalized shift of the Fabry–Perot resonant mode (defined as $\Delta\lambda/\lambda$) versus applied axial strain for the 240 micron (diamonds) and the 460 micron diameter fiber (dots). The curve obtained for the 240 micron diameter fiber is lower because the experimental 0% strain is likely to correspond to a nonzero positive strain, necessary to keep this thinner fiber straight for the measurement (equivalent to a load less than 10 g). Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

Wavelength-Scalable Hollow Optical Fibers with Large Photonic Bandgaps for CO_2 Laser Transmission

Hollow optical transmission fibers offer the potential to circumvent fundamental limitations associated with conventional index guided fibers and thus have been the subject of active research in recent years. Here we report on the materials selection, design, fabrication, and characterization of extended lengths of hollow optical fiber lined with an interior omnidirectional dielectric mirror. These fibers consist of a hollow air core surrounded by multiple alternating submicron-thick layers of a high-refractive-index glass and a low-index polymer, resulting in large infrared photonic bandgaps. These gaps provide strong confinement of optical energy in the hollow fiber core and lead to light guidance in the fundamental and up to fourth-order gaps. We show that the fiber transmission windows can be scaled over a large wavelength range covering at least 0.75 to 10.6 microns. The utility of our approach is further demonstrated by the design and fabrication of tens of meters of hollow photonic bandgap fibers for 10.6 micron radiation transmission. We demonstrate transmission of carbon dioxide (CO_2) laser light with high power-density through more than 4 meters of hollow fiber and measure the losses to be less than 1.0 dB/m at 10.6 microns. This establishes suppression of fiber waveguide losses by orders of magnitude compared to the intrinsic fiber material losses.

Silica optical fibers have been extremely successful in telecommunications applications, and other types of solid-core fibers have been explored at wavelengths where silica is not transparent. However, all fibers that rely on light propagation principally through a solid material have certain fundamental limitations stemming from nonlinear effects, light absorption by electrons or phonons, material dispersion, and Rayleigh scattering that limit maximum optical transmission power and increase attenuation losses. These limitations have, in turn, motivated the study of a fundamentally different light-guiding methodology: the use of hollow waveguides having highly reflecting walls. Light propagation through air in a hollow fiber eliminates or greatly reduces the problems of nonlinearities, thermal lensing, and end-reflections, facilitating high-power laser guidance and other applications which may be impossible using conventional fibers. Hollow metallic or metallo-dielectric waveguides have been studied fairly extensively and found useful practical application, but their performance has been bounded by the notable losses occurring in metallic reflections at visible and infrared (IR) wavelengths, as well as by the limited length and mechanical flexibility of the fabricated waveguides. Hollow all-dielectric fibers, relying on specular or attenuated total reflection, have also been explored, but

high transmission losses have prevented their broad application. More recently, all-dielectric fibers, consisting of a periodic array of air holes in silica, have been used to guide light through air using narrow photonic bandgaps. Solid-core, index-guiding versions of these silica photonic crystal fibers have also been explored for interesting and important applications, such as very large core single-mode fibers, nonlinear enhancement and broadband supercontinuum generation, polarization maintenance, and dispersion management. However, the air-guiding capabilities of such waveguides thus far remain inferior to transmission through solid silica, due to various factors such as the difficulties in fabricating long, uniform fibers which must have a high volume fraction of air and many air–hole periods, as well as by the large electromagnetic (EM) penetration depths associated with the small photonic bandgaps achievable in these air–silica structures. In our fiber, the hollow core is surrounded by a solid high-refractive-index-contrast multilayer structure leading to large photonic bandgaps and omnidirectional reflectivity. The pertinent theoretical background and recent analyses indicate that such fibers may be able to achieve ultralow losses and other unique transmission properties. The large photonic bandgaps result in very short EM penetration depths within the layer structure, significantly reducing radiation and absorption losses while increasing robustness. Omnidirectional reflectivity is expected to reduce intermode coupling losses.

To achieve high index contrast in the layered portion of the fiber, we combined a chalcogenide glass with a refractive index of $\sim$2.8 As_2Se_3, and a high-performance polymer with a refractive index of $\sim$1.55 PES. We recently demonstrated that these materials could be thermally co-drawn into precisely layered structures without cracking or delamination, even under large temperature excursions. The same polymer was used as a cladding material, resulting in fibers composed of $\sim$98% polymer by volume (not including the hollow core) and thus combine high optical performance with polymeric processability and mechanical flexibility. We fabricated a variety of fibers by depositing a 5–10 micron thick As_2Se_3 layer through thermal evaporation onto a 25–50 micron thick PES film and the subsequent 'rolling' of that coated film into a hollow multilayer tube called a fiber preform. This hollow macroscopic preform was consolidated by heating under vacuum and cladded with a thick outer layer of PES; the layered preform was then placed in an optical fiber draw tower and drawn down into tens or hundreds of meters of fiber having well-controlled submicron layer thicknesses. The nominal positions of the photonic bandgaps were determined by laser monitoring of the fiber OD during the draw process. Typical standard deviations in the fiber OD were $\sim$1% of the OD. The resulting fibers were designed to have large hollow cores, useful in high-energy transmission.

SEM analysis (**Fig. 16**) reveals that the drawn fibers maintain proportionate layer thickness ratios and that the PES and As_2Se_3 films adhere well during rigorous thermal cycling and elongation. Within the multilayer structure shown in **Fig. 1**, the PES layers (gray) have a thickness of 900 nm, and the As_2Se_3 layers (bright) are 270 nm thick (except for the first and last As_2Se_3 layers, which are 135 nm). Broadband fiber transmission spectra were measured with a Fourier transform infrared (FTIR) spectrometer (Nicolet Magna 860), using a parabolic mirror to couple light into the fiber and an external detector. The results of these measurements are shown in the lower panel of **Fig. 2** for fibers having two different layer structures. For each spectrum, light is guided at the fundamental and high-order photonic bandgaps. Also shown in the upper panel of **Fig. 2**, is the corresponding photonic band diagram for an infinite periodic multilayer structure calculated using the experimental parameters of our fiber (layer thicknesses and indices). Good agreement is found between the positions of the measured transmission peaks and the calculated bandgaps, corroborated with the SEM-measured layer thicknesses, verifying that transmission is dominated by the photonic bandgap mechanism. In order to demonstrate 'wavelength scalability' (i.e., the control of transmission through the fiber's structural parameters) another fiber was produced, having the same cross-section but with thinner layers. We compared the transmission spectra for the original 3.55 micron bandgap fibers to the fiber with the scaled-down layer thicknesses.

Fig. 17 shows the shifting of the transmission bands, corresponding to fundamental and high-order photonic bandgaps, from one fiber to the next. The two fibers analyzed in **Fig. 17** were fabricated from the same fiber preform using different draw-down

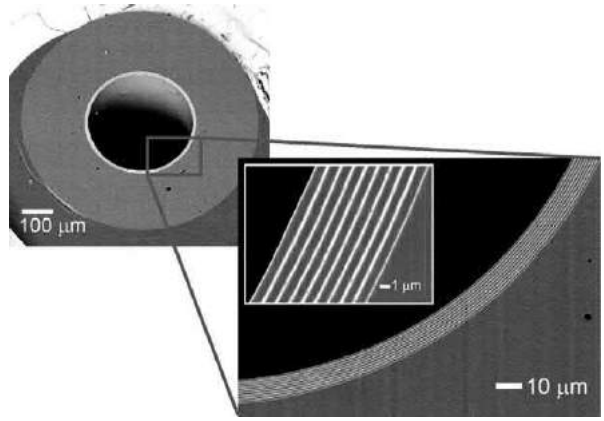


Fig. 16 Cross-sectional SEM micrographs at various magnifications of hollow cylindrical multilayer fiber mounted in epoxy. The hollow core appears black, the PES layers and cladding gray, and the As_2Se_3 layers bright white. This fiber has a fundamental photonic bandgap at a wavelength of $\sim$3.55 microns. Reproduced with permission from Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

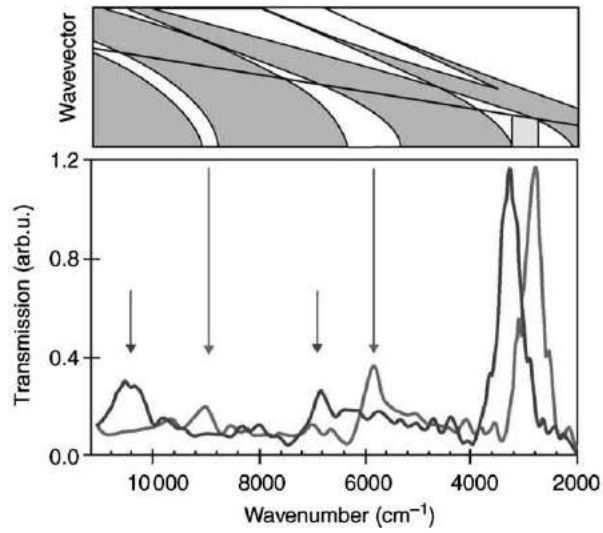


Fig. 17 Upper panel: Calculated photonic bandstructure associated with the dielectric mirror lining of the hollow fiber. Modes propagating through air and reflected by the fiber walls lie in the bandgaps (white) and within the light cone defined by the glancing-angle condition (black line). The gray regions represent modes radiating through the mirror. The fundamental bandgap has the widest range-to-mid-range ratio, with a region of omnidirectional reflectivity highlighted in black. Lower panel: Comparison of transmission spectra for two different hollow fibers of ~30 cm length having similar structure but scaled photonic crystal (multilayer) period dimensions. The spectrum in gray is from a fiber with a fundamental photonic bandgap at 3.55 microns; the black spectrum is from a fiber where the corresponding bandgap is at 3.1 microns. High-order bandgaps (indicated by arrows) are periodically spaced in frequency. Reproduced with permission from Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

ratios (fibers with a fundamental bandgap centered near 3.55 microns have an OD of 670 microns; those with a gap at 3.1 microns have an OD of 600 microns). The high-order bandgaps are periodically spaced in frequency, as expected for such a photonic crystal structure.

The wavelength scalability of our fibers was further demonstrated in the fabrication of hollow fibers designed for the transmission of 10.6 micron EM radiation. This not only shows that these structures can be made to guide light at extremely disparate wavelengths, but that specific useful bandgap wavelengths can be accurately targeted during fabrication and fiber drawing. Powerful and efficient CO_2 lasers are available that emit at 10.6 microns and are used in such applications as laser surgery and materials processing, but waveguides operating at this wavelength have remained limited in length or loss levels. Using the fabrication techniques outlined above, we produced fibers having hollow core diameters of 700–750 microns and ODs of 1300–1400 microns with a fundamental photonic bandgap spanning the 10–11 micron wavelength regime, centered near 10.6 microns. **Fig. 5** depicts a typical FTIR transmission spectrum for these fibers, measured using ~30 cm long straight fibers.

In order to quantify the transmission losses in these 10.6 micron bandgap hollow fibers, fiber cutback measurements were performed. This involved the comparison of transmitted intensity through ~4 meters of straight fiber with the intensity of transmission through the same section of fiber cut to shorter lengths (**Fig. 18** inset). This test was performed on multiple sections of fiber, and the results found to be nearly identical for the different sections tested. The measurements were performed using a 25 watt CO_2 laser (GEM-25, Coherent-DEOS) and high power detectors (Newport 818T-10). The fiber was held straight, fixed at both ends as well as at multiple points in the middle to prevent variations in the input coupling and propagation conditions during fiber cutting. The laser beam was sent through focusing lenses as well as 500 micron diameter pinhole apertures and the input end face of the fiber was coated with a metal film to prevent accidental laser damage from misalignment. The transmission losses in the fundamental bandgap at 10.6 microns were measured to be 0.95 dB/m, as shown in the inset of **Fig. 18**, with an estimated measurement uncertainty of 0.15 dB/m. These loss measurements are comparable to some of the best reported loss values for other types of waveguides operating at 10.6 microns. A bending analysis for fibers with a bandgap centered at 10.6 microns revealed bending losses below 1.5 dB for 90 degree bends with bending radii from 4–10 cm. We expect that these loss levels could be lowered even further by increasing the number of layers, through optimization of the layer thickness ratios and by creating a cylindrically symmetric multilayer fiber with no inner seam (present here because of the 'rolling' fabrication method). In addition, using a polymer with lower intrinsic losses should greatly improve the transmission characteristics.

One reasonable figure of merit for optical transmission losses through hollow all-dielectric photonic bandgap fibers is to compare the hollow fiber losses to the intrinsic losses of the materials used to make the fiber. As_2Se_3 has been explored as an IR-transmitting material, yet the losses at 10.6 microns reported in the literature are ~10 dB/m for highly purified material, and more typically are greater than 10 dB/m for commercially available materials such as those used in our fabrication. Based on FTIR transmission and spectroscopic ellipsometer measurements that we have performed on PES, the optical losses associated with propagation through solid PES should be greater than 40 000 dB/m at 10.6 microns. This demonstrates that guiding light through

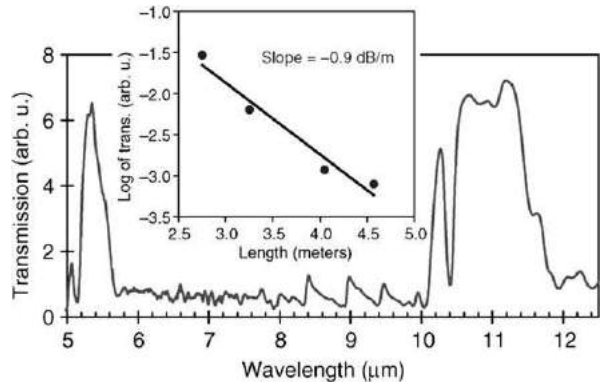


Fig. 18 Typical transmission spectrum of hollow fibers designed to transmit CO_2 laser light. The fundamental photonic bandgap is centered near a wavelength of 10.6 microns and the second-order gap is at ~5 microns. Inset: Log of transmitted power (arbitrary units) versus length of fiber (meters). The slope of this graph is the loss in dB/m. The measured fiber has a hollow core diameter of 700 microns. Reproduced with permission from Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

air in our hollow bandgap fibers leads to waveguide losses that are orders of magnitude lower than the intrinsic fiber material losses, which has been one of the primary goals of hollow photonic bandgap fiber research. These comparatively low losses are made possible by the very short penetration depths of EM waves in the high refractive index contrast photonic crystal structure, allowing these materials to be used at wavelengths that may have been thought improbable. Another long-standing motivation of infrared fiber research has been the transmission of high-power laser light. As a qualitative demonstration of the potential of these fibers for such applications, both straight and smoothly bent fibers of lengths varying from 0.3–2.5 meters were used to transmit enough CO_2 laser energy to burn holes through paper. The maximum laser power density coupled into our fibers in these trials was approximately 300 W/cm^2, more than sufficient to burn a homogeneous polymer material, including PES. No damage to the fibers was observed when the laser beam was properly coupled into the hollow fiber core. These results indicate the feasibility of using hollow multilayer photonic bandgap fibers as a low-loss wavelength-scalable transmission medium for high-power laser light.

Further Reading

Abeles, F., 1950. Investigations on the propagation of sinusoidal electromagnetic waves in stratified media. Application to thin films. Annales de Physique 5, 706.

Benoit G and Fink Y, Spectroscopic Ellipsometry Database, http://mit-pbg.mit.edu/Pages/Ellipsometry.html.

Born, M., Wolf, E. (Eds.), 1980. Principles of Optics, 6th edn. New York:: Pergamon Press, p. 67.

Engeness, T.D., Ibanescu, M., Johnson, S.G., *et al.*, 2003. Dispersion tailoring and compensation by modal interactions in OmniGuide fibers. Optics Express 11, 1175–1198, http://www.opticsexpress.org/abstract.cfm?URI=OPEX-11-10-1175.

Fink, Y., Winn, J.N., Fan, S., *et al.*, 1998. A dielectric omnidirectional reflector. Science 282, 1679–1682.

Fink, Y., Ripin, D.J., Fan, S., Chen, C., Joannopoulos, J.D., Thomas, E.L., 1999. Guiding optical light in air using an all-dielectric structure. Journal of Lightwave Technology 17, 2039–2041.

Hart, S.D., Maskaly, G.R., Temelkuran, B., Prideaux, P.H., Joannopoulos, J.D., Fink, Y., 2002. External reflection from omnidirectional dielectric mirror fibers. Science 296, 510–513.

Ibanescu, M., Johnson, S.G., Soljacic, M., *et al.*, 2003. Analysis of mode structure in hollow dielectric waveguide fibers. Physics Review E 67, 1–8. 046608.

Johnson, S.G., Ibanescu, M., Skorobogatiy, M., *et al.*, 2001. Low-loss asymptotically single-mode propagation in large-core OmniGuide fibers. Optics Express 9, 748–779, http://www.opticsexpress.org/abstract.cfm?URI=OPEX-9-13-748.

Kuriki, K., Shapira, O., Hart, S.D., *et al.*, 2004. Hollow multilayer photonic bandgap fibers for NIR applications. Optics Express 12, 8.

Soljacic, M., Ibanescu, M., Johnson, S.G., Joannopoulos, J.D., Fink, Y., 2003. Optical bistability in axially modulated OmniGuide fibers. Optics Letters 28, 516–518.

Temelkuran, B., Hart, S.D., Benoit, G., Joannopoulos, J.D., Fink, Y., 2002. Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. Nature 420, 650–653.

Yeh, P., Yariv, A., Hong, C.-H., 1977. Electromagnetic propagation in periodic stratified media. 1. General theory. Journal of the Optical Society of America 67 (4), 423.

Yeh, P., Yariv, A., Marom, E., 1196. Theory of Bragg fiber. Journal of the Optical Society 68, 1196–1201.

Guided Wave Optics

Alan Mickelson, University of Colorado at Boulder, Boulder, CO, United States

Introduction

Already in 1842, Colladan had published his observation of light being guided within a curving jet of water (Colladan, 1842a,b). The demonstration led to some scientific study as well as numerous artistic and carnival displays (Colladan, 1842b; Hecht, 2004). From this 19th century curiosity, guided wave optics has grown to be the physical level technology that transfers most of the world's information. The waveguide of the worldwide telecommunications network is the optical fiber. The optical fiber itself is but a passive waveguide and guided wave optics is the technology which includes all of the passive and active components which are necessary to generate light, impress (electrical) information on light to produce information bearing optical signals, regenerate these optical signals during transmission, and convert optical signals back to (electrical) information streams at the transmission system output. In this article, some introduction to this rather encompassing topic of guided wave optics will be given.

This article will be separated into six sections. In the next section, discussion will be given to fiber optics, that is, the properties of these optical waveguides which allow light to travel in distinctly non-rectilinear paths over terrestrial distances. In Section Active Fiber Compatible Components of the present article will then turn to the components which can be used along with the fiber optical waveguides to create transmission systems. These components include light sources, modulators and detectors as well as optical amplifiers. In Section Telecommunications Technology of the article, we will discuss the telecommunications network which has arisen due to the availability of fiber optics and fiber optic compatible components. The article would hardly be complete without some discussion of the ever growing field of optical data communications that is Section Data Communications of the exposition. The article closes with a Summary section.

Fiber Optics

Fiber optics refers to a technology in which light (actually infrared, visible or ultraviolet radiation) is transmitted through the transparent core of a small (250 μm diameter – a human hair is circa 75 μm diameter) thread of composite material. The composite material consists of a core concentric with a cladding of lower optical density (index of refraction) than the core and a coating, generally a polymer, that is applied during manufacture for environmental protection (see **Fig. 1**). The composite material is referred to as an optical fiber. Typical dimensions for a telecommunications fiber (which most fiber is) are core diameter of 10 μm for a single mode fiber and 50 μm for a multimode fiber, cladding diameter 125 μm and coating of 250 μm.

The fiber coating here is included as a part of the fiber, unlike in most textbooks interested in mathematical analysis alone, because uncoated fiber has a short lifetime. Rarely does one see an uncoated fiber unless one has just mechanically or chemically stripped off the coating in anticipation of further processing of the bare fiber, that is, splicing or other attachment to a device. Fiber that is clear and coiled on a spool is coated fiber. When fiber is jacketed it appears much as conventional insulated wire although the resulting wire is lighter and more flexible than copper (or other metal core) wire. A fiber cable contains many jacketed fibers

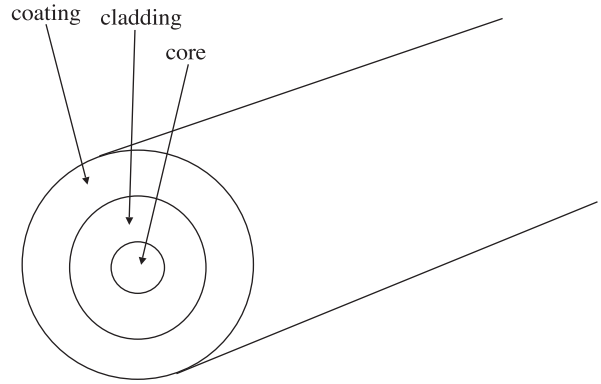


Fig. 1 Schematic depiction of the concentric layers that make up a fiber optic thread. The fiber consists of a core in which the light is guided, a cladding that confines the light to the core as well as adding mechanical support and a coating that protects the fiber from the environment. Silica fiber, the choice of material for telecommunications fiber, is especially susceptible to contamination by humidity. Typical dimension for a telecommunications fiber are a coating diameter of 250 μm, a cladding diameter of 125 μm and a core diameter of 10 μm for single mode fiber and 50 μm for multimode fiber. At the wavelength of 1.55 μm, the index of refection of a pure silica cladding is 1.45 and of the germano-silicate core circa 1.46 for a typical numerical aperture of 0.024.

(840 is a usual maximum number for the largest of sea cables) with some support material for rigidity. Optical fiber for telecommunications is fabricated by a process of gas phase chemical deposition of fused silica doped with various trace chemicals. Because of the distances involved in the world telephone network, it is safe to say that practically all optical fiber in existence is telecommunications fiber. Fiber can be made from a number of different material systems and in a number of different configurations for use with various types of light sources in specialty applications. In what follows, we will limit discussion to the basic properties of the light guided by the fiber and leave fabrication and manufacture for the those readers who will to research on their own from the numerous sources available on these topics.

There are two complimentary mathematical descriptions of the propagation of light in an optical waveguide. Ray theory is an approximate description that is quite accurate for properties of multimode fibers. Wave theory is a more rigorous description that is necessary to elucidate the properties of single mode propagation but is needlessly complex for the description of multimode fiber where there are typically 1000 modes. We will briefly consider both descriptions of the propagating light in order to understand coupling of light to and from fiber, attenuation of light in fiber and the dispersion of light pulses when propagating along a fiber axis.

In the ray description, light is considered to be made up of a bundle of rays. In a uniform homogeneous medium, each ray is an arrow that exhibits rectilinear propagation from its source to its next interface with a dissimilar material. These rays satisfy the law of reflection (incident angle equals reflected angle) and Snell's law of refraction (bending is proportional to the optical density of the material) at interfaces between materials with dissimilar optical properties. That is, at an interface, a fraction of the light is reflected backwards at an angle equal to the incident angle and a portion of the light is transmitted in a direction which is directed more toward the normal to a plane interface when the index of refraction increases across the boundary and is directed more away from the normal when the index decreases. Using these laws at the input enface to the fiber, a portion of the energy guided by each ray is reflected back into space due to the change in refractive index at the guide surface. The refracted ray enters the fiber to exhibit a continued path.

In a step index optical fiber where the index of refraction is uniformly higher in the fiber core than in a surrounding cladding, a ray will propagate along a straight path until encountering the core cladding interface. Unguided rays (see **Fig. 2**) will be only partially reflected at the core cladding interface. The quantity of light remaining in the fiber will rapidly die out while propagating along the axis of the fiber due to this loss of intensity (energy) at each successive surface reflection. Rapidly is obviously a relative term. Typical distances between successive encounters with the boundary are 1 mm and total propagation distances may be many kilometers. The loss of even a tenth of 1% per reflection will result in the wave being decreased in intensity by about a factor of ten in 1 m (1000 reflections).

Because of the law of refraction, the unguided rays refracted at the interface are directed ever more closely to the direction of the core cladding boundary as the for the rays that are ever more directed along the axis. There is then an incident angle at which the direction of the refracted ray is along the interface. If the incident angle is more axially directed than this cut-off angle, there is no refracted ray in the cladding. These guided rays (see **Fig. 3**) are totally internally reflected back into the fiber core to again be totally internally reflected at the next core cladding interface. Although none of the guided rays suffer loss at the interface, for each

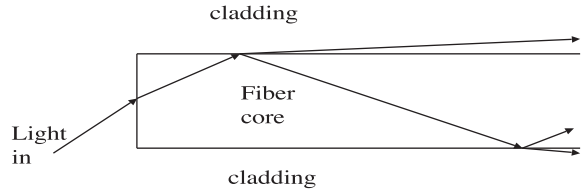


Fig. 2 A graphic indicating a ray path for an unguided (unbound) light ray. In the ray picture of light, a source is represented as an emitted bundle of rays. A consequence of Snell's law of refraction is that there is an incident angle for which the transmitted ray travels along the interface when the ray is incident from a higher to a lower optical density (index of refraction). When the ray is incident at an angle closer to the normal than this critical angle, a ray is transmitted with an amplitude given by a Fresnel coefficient. As light is shared with the cladding at each refection point, the light ray is unbound.

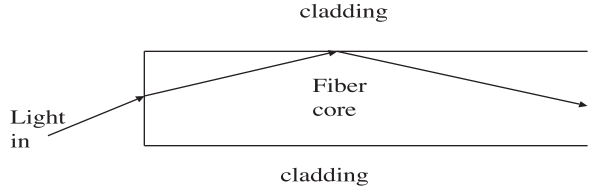


Fig. 3 A graphic indicating a ray path for a guided (bound) light ray. In the ray picture of light, a source is represented as an emitted bundle of rays. A consequence of Snell's law of refraction is that there is an incident angle for which the transmitted ray travels along the interface when the ray is incident from a higher to a lower optical density (index of refraction). For a ray incident more nearly tangential to the interface than the critical angle, there is no transmitted ray. This is to say that rays incident in this range of angles are bound to the core of the fiber.

incident angle, the guided ray will travel a different distance along its path in traversing a given distance along the fiber axis. This effect is referred to as modal dispersion. A bit of information is a single pulse of light. A single pulse of light if launched into a number of modes will then spread as it propagates down a step index, multimode fiber. As the quantity of information transmitted is proportional to the number of pulses transmitted per unit time, dispersion limits the quantity of information that can be transmitted over a given distance along the fiber axis.

In a graded index fiber, the refractive index within the fiber core varies continuously from a maximum somewhere within the core to a minimum at the core cladding interface. The guided ray paths within such fibers are curved whereas the unguided simply exit the fiber when they reach the cladding. That is, the guided rays in graded index medium are characterized by the fact that once in the fiber they never encounter the cladding. Unlike in step index fiber, there is no clear relation between the angle at which ray enters the fiber and the distance that a guided ray actually traverses along its path in order to traverse a distance along the fiber axis. The dispersion of a graded index multimode fiber is dependent on the grading of the index. Modal dispersion is generally expressed as the number of cycles of modulation that can be transmitted over a given length with only a 50% decrease in cycle amplitude. The unit of a cycle is a Hz (hertz). A 1 Hz bandwidth is roughly equivalent an information rate of 1 bps (bit per second). As bps is a small unit, thousands (Kbps), millions (Mbps), billions (Gbps) and trillions (Tbps) are generally used. The bandwidth of a step index telecommunications fiber is roughly 100 MHz km and a graded index fiber 1 GHz km.

Guided rays suffer no excess loss per propagation cycle, that is, a pair of reflections in a step index or complete curvilinear path in a graded index fiber. That is not to say that the propagation is loss free. There is some residual absorption and scattering of the light that are dependent on the wavelength of the light. Telecommunications wavelengths are clustered into three wavelength windows, one about 0.85 μm, the second at 1.3 μm and the third at 1.55 μm. These intrinsic losses are result in the wave being attenuated by a factor of 10 in about 3 km at 0.85 μm, 20 km at 1.3 μm and 50 km at 1.55 μm. These losses are much lower than the leakage losses associated non guidance discussed above.

The ray description of fiber propagation is quite simple. Rays incident on the input surface at less than a critical angle (defined by the fiber numerical aperture (NA) of the fiber) are guided. Other rays do not couple. At the fiber output, the converse is true. Rays exit from the enface at all angles up to the cut-off angle defined by the NA (which happens to be the sine of the cut-off angle) of the fiber. What the ray picture does not take into account is the wave nature of light. In the wave picture, each ray is carrying a clock that remembers how long it has been propagating along its ray path. When two rays come together, they can either add or subtract depending on the reading on their respective clocks and the wavelength of the source. This is the process known as interference. The ray picture gives a poor description of the lowest order mode on an optical fiber, the one that follows the axis. When the fiber has a small enough core and NA (small enough acceptance angle), only this fundamental mode can propagate. This defines a single mode fiber. As there is only one ray path, all light coupled into the fiber propagates at the same velocity. Hence, we say that there is no modal dispersion. There is still dispersion as difference wavelengths will travel at different velocities.

Multimode fibers have larger cores and higher numerical apertures than single mode fibers. They are easier to couple to. As their description (the ray description) did not require interference to be taken into account, coupling does not require the source to radiate at only one wavelength, that is, to be monochromatic. Monochromatic, single wavelength, is the lower end limit of what we call the spectral (band) width of a source. One can couple light from even a white light source (broad spectral band) into a sufficiently multimode fiber. As each more, though, propagates at a different velocity, the ease of coupling and alignment comes at the cost of information rate. The information content of signal is inversely proportional to the temporal width of the pulses (how many pulses can be jammed into a given time interval) in the signal. The bandwidth of the signal we can send over a multimode fiber is limited by the number of modes. In contrast, a signal mode fiber requires small dimensions and tight tolerances to couple. Single mode coupling is also phase dependent so a narrow band source such as a single mode laser is necessary in order to couple light efficiently. Pulses, though, disperse much more slowly with propagation distance in the single mode fiber than in the multimode fiber. Much more information an be transmitted over the same distance although at a cost of higher alignment tolerances and use of more coherent sources. As stated above, a graded index multimode fiber can support an information bandwidth of roughly 1 GHz-km, that is, one can transmit 1 GHz 1 km, half a GHz over 2 km, etc. The bandwidth of a single mode fiber is defined by the source wavelength and spectral width as well as the length of propagation. The dispersion can also be managed by alternating the type of fiber in which the signal propagates along the propagation path. At present (2017), single mode fibers can transmit bandwidths of 25 GHz over 40 km spans between repeaters. Further improvement is possible.

Active Fiber Compatible Components

The previous discussion of propagation in fiber optic waveguides would be rather sterile were there not a number of available fiber compatible components that can generate and detect light streams. In order to construct fiber optic systems, components to impress information streams on light and to read out the impressed information at a receiver are also necessary. In this section, we will discuss some of the theory of operation of such active optical components with a goal of understanding possibilities and limitations of guided wave technology.

In a passive component such as an optical fiber, the optical signal is considered as given. The power carried by the signal can only be attenuated. The attenuation may be through imperfect coupling or loss mechanisms. Signal fidelity can be altered through dispersion but the impressed information is considered as given. New information cannot be created. Active components are different. Active components require that the optical fields interact with a medium such that energy and/or information can be

transferred from the active medium to the light field or from the light field to the active medium. Light can be amplified. In an optical amplifier, energy is transferred from an atomic medium to a propagating wave. Light can be generated where there was no light. That is, amplification (more correctly said, oscillation) can take place even without an input signal. Information can be impressed on a light stream. That is, in a modulator, a signal is applied to the atomic medium while a constant optical traverses it. The result is the impression of an electrical information stream on an optical carrier. This operation of mixing signals is inherently nonlinear. There is no direct inverse operation. However, a light stream can be converted back to an electrical stream. A detector can works as a source but in reverse. Fortunately, optical detectors with short response times can be used to follow the variations of a modulated optical signal and thereby recover a modulating electrical information stream while converting the optical to electrical power.

Sources compatible with fiber optic systems are quite generally non-thermal sources. Fiber compatible sources operate on quantum mechanical principles. The electrons in an atom can only exist in certain energy states, that is, in certain atomic configurations. An isolated atom is then characterized by a set of energy levels, that is, energies that are absorbed or emitted when the atom changes configuration. Energy must be applied to reconfigure the atom to a higher energy state. Energy is released when an atom relaxes to a lower energy state. The energy can take the form of a single particle of light of a particular wavelength (frequency). Such a particle of light is called a photon. It is usual to consider a single pair of levels of an atom. A pair of levels is referred to as a transition. For example, the rare earth ions of a rare earth doped optical amplifier act independently of one another. The medium acts as a sum of atomic media. A single transition of the rare earth ion interacts with a stream of light, that is, of photons of light with a given wavelength. By preparing the atoms such that they are in the upper state of their transition, one can extract energy from the medium to the light.

In a semiconductor, atoms act collectively. That is, inner electrons are bound to specific nuclei whereas the outer electrons are shared between all of the nuclei. The shared electrons can either all be weakly bound to the nuclei, or free to roam in such a manner that charge neutrality is preserved. The level spacing of interest in the semiconductor is that of the band gap. The band gap is the spacing between the least energetic freely propagating (conduction) election and the most energetic bound (valence) electron. In the semiconductor, then, the transition of interest is the valence to conduction band transition. The wavelength of operation of such an active medium is determined by the energy spacing between the upper and lower levels of this quantum mechanical transition. As illustrated in **Fig. 4**, injecting current into a semiconductor diode introduces electrons into the conduction band from the negative electrode. Valence band electrodes in the junction that recombine with these conduction band electrons to give off photons at the wavelength of the transition. A source, be it light emitting diode (LED) or laser diode (see **Fig. 4**) will emit light at a wavelength which corresponds to an energy slightly above the minimum gap energy. Detectors are in essence the inverse of sources (see **Fig. 5**). However, there are important differences.

A detector can detect almost any energy greater than the band edge. That is, a laser or LED will emit light in a band of wavelengths around a center that corresponds to an energy values (photon energy is equal to a constant (that happens to Planck's constant times the speed of light) divided by the wavelength of the photon) slightly greater than the band gap. A detector is not limited on the high energy end. If the detector sees red light, it will also see blue light. A more important difference, though, has to do with the direction of the transition. In a detector, the final electron state is a conduction electron. Conduction electrons are meant to be swept out of a junction as current. In a source, the final state is a photon that is due to recombination of a conduction and valence electron. It is not so easy to radiatively combine electrons and holes as the valence electrons are called. To make a long story short (the concerned reader will be able to find numerous sources on semiconductor lasers), only direct band gap materials can be used for sources. Silicon and germanium, the materials of most electronics, are indirect gap. Gaaa and InGaAsP are direct gap and are the materials of choice for semiconductor lasers.

The historical development of the fiber system sources serves to explain what was previously only stated. Telecommunications employs essentially three windows of wavelengths, one centered about 0.85 μm, one about 1.3 μm and a third about 1.55 μm. Why the first window is referred to as first can be explained by the development by the ternary source material GaAlAs before the

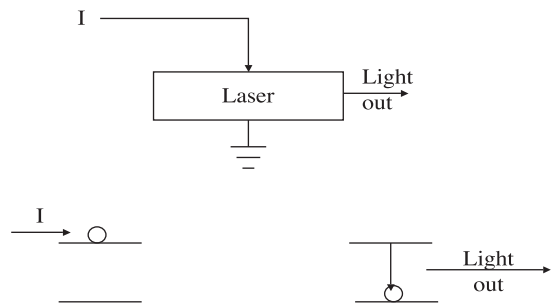


Fig. 4 Schematic depiction of the quantum mechanical action of an electrically pumped, non-thermal light source, that is, a light emitting diode (LED) or laser. Current is injected into a semiconductor diode structure that results in injected electrons entering the conduction band of the light emitting region. The conduction band is an excited band at an energy level higher than that of the ground state valence band. To relax to the valence band, the electron must give up energy in the form of a photon, a particle of light whose frequency corresponds to the transition energy that is related to the transition frequency by energy equals Planck's constant times the frequency.

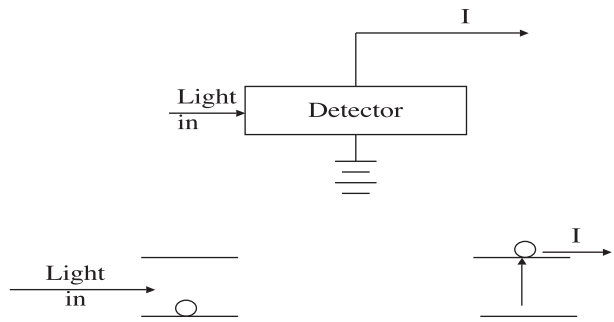


Fig. 5 Here, the detection process in a semiconductor medium is illustrated. The process is inverse to the emission process in some respects but not others. A photon is incident on the detecting region of a semiconductor. If the photon energy is higher in energy than the band gap then there is a probability that it will be absorbed by a valence band electron that will be elevated in energy to the conduction band. Electrical bias of the active region of the detector will then sweep the conduction band carrier out of the junction area. Whereas in a source, recombination requires that both an electron and hole are present before emission, any valence band electron can be promoted to the almost empty conduction band and then swept out of the junction by a bias field.

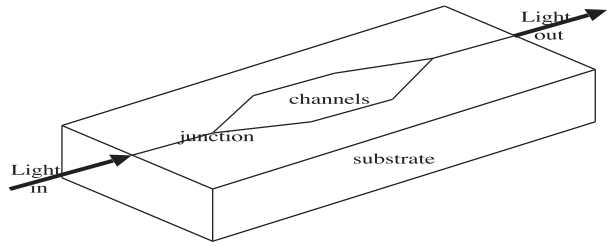


Fig. 6 Schematic depiction of the arrangement of waveguide regions in a Mach-Zhender modulator. Light that enters the modulator through the single mode waveguide to the left is split, and then propagated through straight parallel channels before being recombined. If the light in the two arms is in phase on recombination, the output power will equal the input power. If the light is out of phase, the light will radiate from the down coupling region. If the channels are electro optic and a time-varying signal is applied between them, the relative phase induced by the signal will be transduced into a time varying intensity at the output junction.

quaternary materials of the InGaAsP family. Pure GaAs has a gap that corresponds to 0.85 μm whereas increasing the Al fraction shifts the operation to shorter wavelengths. Doped fused silica is the material used for telecommunications grade optical fiber. The loss of fused silica fibers decreases with increasing wavelength until 1.6 μm with a single exception of a broad absorption peak around 1.1 μm. The loss abruptly increases orders of magnitude above 1.6 μm. The window around 0.85 μm is where the wavelength of GaAlAs lasers overlap a minimum loss window of fused silica. InGaAsP is a family of materials whose resonant wavelength can be tuned to lie from circa 1. 0 μm to 1.6 μm. The dispersion minimum of fused silica lies about 1.3 μm and the minimum loss of fused silica lies at 1.55 μm.

The last piece of the active device puzzle concerns modulation. Semiconductor sources that operate by current injection offer the option of direct modulation, that is, one can add an information stream in the form of an input current stream. A problem with such modulation is that it disturbs the spectral characteristics of the source. Much has been done to minimize this distortion. There are a myriad of different semiconductor sources with a myriad of acronyms to describe that were mainly introduced to minimize spectral distortion under modulation. An important one is DFB or distributed feedback laser. In the most exacting of applications, one uses an external modulator such as the Mach-Zhender structure (MZ) seen in **Fig. 6**. No attempt has been made here to specify the technology or even the electrode structure. The material systems and drive systems are too numerous to describe in this article. The operating principle, though, is simple. It is also the principle shared by most free space interferometers. In an external modulator system, a stabilized laser output is input into the left hand input of the MZ. The light is split as equally as possible into two streams. In the channel region, these streams propagate in parallel but at a distance to each other such that they do not interact. If nothing happens in this region and the path lengths are equal, the light will recombine. If an information source is used to affect the relative phases of the waves in the two arms, though, the light will not recombine. The phase modulation will be converted to intensity modulation.

How to impress phase modulation of the arms of the MZ? The highest speed method is to use the electro optic effect in an electro optic medium. By placing electrodes over the two channels and applying the information stream between these channels, one can alter the index of refraction and hence optical path in one arm relative to the others. The electro optic effect in a number of materials (lithium niobate, silicon, GaAs, InGaAsP, etc.) is electronic in origin and, therefore, can respond to an electrical signal at rates as high as optical frequencies. In fact, light can be used to modulate light. More commonly, the modulation frequency, though, is limited by the frequencies that can be electrically generated and applied to electrodes. We leave this topic to the

concerned reader and leave the optic to say that, for example, silicon photonics MZ modulators that modulate to 50 Gbps are presently available at a number of foundries.

Telecommunications Technology

The physical transmission layer of the world wide telecommunications network has come to be dominated by fiber optic technology. The long haul network that transmits the majority of the world's bits is comprised of wavelength division multiplexed (WDM) single mode fiber optic links operating in the 1.55 μm telecommunications window. In the following, we will first discuss how this transformation to optical communications took place and then go on to discuss some of the specifics of the technology. The highest capacity channels are present (2017) transmit 100 Gbps per channel by using four level phase modulation at 25 Gbps on each of the polarization states of the fiber at one of the International Telecommunications Union (ITU) defined wavelengths within the so-called C- band of optical communications spectrum. The C-band spans 35 nm from 1530 to 1565 nm in the infrared. At 1550 nm, a 1 nm of wavelength bandwidth is equivalent to 125 GHz of frequency bandwidth (using that wavelength equals the speed of light divided by frequency). It is, therefore, possible to place 40 of these 100 Gbps channels (each one occupying circa 20 GHz on 50 GHz spacing) in the C-band in order to transmit a composite rate of 8 Tbps per optical fiber. Bandwidths transmitted by transoceanic cables are now approaching composite bandwidths of 1 Pbps (10^{15} bits/s).

Already by the middle of the 1960s, it was clear that changes were going to have to take place in order that the exponential growth of the telephone system in the United States as well as in Europe could continue. The telephone system then, pretty much as now, consisted of a hierarchy of tree structures connected by progressively longer lines. A local office is used to connect a number of lines emanating in a tree structure to local users. The local offices are connected by trunk lines. The lines from there on up the hierarchy are long distance ones which are termed long lines and can be regional and longer. The most pressing problem in the late 1960s was congestion in the so-called trunk lines which was most severe in urban areas. In many cities, these trunk lines were two kilometers in length, a length that was fixed by the underpayment ducts in which the cables were places. The problem was that there was no more space in the ducts that housed these lines in any number of cities including New York City. The solution to duct congestion had traditionally been to use progressively more time division multiplexing (TDM) to increase the traffic that could be carried by each of the lines already buried in the conduit.

The human voice produces sounds in the frequency range of 20 Hz to as high as 20 kHz. The upper range (above circa 10 kHz), though, is necessary only for opera, dog whistles and a some other more esoteric purposes. Usual speech is still understandable if frequencies only as high as 4 kHz are included. By digitally sampling a 4 kHz band limited conversion twice per period and quantizing each amplitude into 8 bits (where $2^8 = 64$ levels), a digital copy of that conversation can be stored as 64 Kbps. If one then TDM's, 24 of these conversations, one obtains a composite rate of circa 1.53 Mbps, a standard known as T1. T1 TDM was already becoming common is the 1960s with the dawning of the digital communications era. If 28 of these T1 channels are then TDM'ed together, one obtains a T3 channel of roughly 45 Mbps. By 1975, the plan for the next generation of trunk lines was to use T3 (also called DS1). The digital equipment was ready but the twisted pair lines that were already installed in the ducts were problematic. The twisted pair lines with a bandwidth of circa 10 MHz km would not even carry the T2 rate of 4 T1 (6 MHz) over two kilometers. That is, dispersion of 10 MHz km over two kilometers (bandwidth of 5 MHz) would smear out the edges of the time varying bit streams reducing the modulation power transmitted. In 1975, the only viable lasers were GaAlAs operating in the first (0.85 μm window) with silicon detectors. The best step index, multimode fiber (evidently for use at 0.85 μm) in 1975 was rated at 100 MHz km and would transmit the edges of a 45 Mbps (T3) signal. The problem was thought to lie in the sources and detectors, but not so.

A 1975 demonstration in Atlanta (Hecht, 2004) of a multimode fiber optic system operating in the 0.85 μm window was successful. Subsequent progress was meteoric. Already by 1980, advances in single mode laser and single mode fiber technology operating at the 1.3 μm wavelength had made the inclusion of fiber into the long lines viable as well. For roughly the decade from 1985 onward, single mode fiber systems dominated long line replacement for terrestrial as well as transoceanic links. The erbium doped fiber amplifier which operated in the 1.55 μm wavelength third telecommunication window had proven itself to be viable for extending repeater periods already by around 1990. Development of the InGaAsP quaternary semiconductor system allowed for reliable lasers to be manufactured for this window as development of strained lattice laser technology in the InGaAs system allowed for efficient pump lasers for the fiber amplifiers. As the optical amplifiers could amplify across the wavelength band that could be occupied by many aggregated channels of TDM signals, the move of the long line systems to the third telecommunications window was accompanied by the adoption of wavelength division multiplexing or WDM.

WDM, unlike TDM, can be carried out using only optics. That is, dispersive optical elements (e.g., diffraction gratings) can split channels into channels of spectral widths as small as 0.2 nm (25 GHz). The International Telecommunications Union defined a standardized wavelength grid down to 25 GHz spacing in the earl 1990s. Soon after, 10 Gbps channels on 50 GHz (0.4 nm) spacing were allowing single fibers to carry up to 32 channels. The 10 GHz proved to be a challenge that electronics was not able to surmount until the 2010s. At present, 25 Gbps modulation rates and ethernet cards are becoming ubiquitous along with the 100 Gbps rates per channel with which the discussion of this section began. But internet demand continues to rise at a rate of 22% per year so the story of the telecommunications network is nowhere near complete.

Data Communications

As is evident from the exposition of the last section on guided wave optics in telecommunications, fiber optics dominates all applications that require a bandwidth length product greater than some limit. The limit is somewhat application dependent (due to cost) but lies in the range of a 100 MHz km for telecommunications. Although fiber to the home (FTTH) has been discussed for many years and even deployed in a number of trials in a number of countries, digital telephony with a requirement of 64 kbps per conversation does not lead to rates requiring a fiber solution for single homes. The rates required for data communication, however, continue to increase. Clock rates of computer chips have stagnated near 4 GHz since circa 2007, but stream serialization has resulted most recently in 25 Gbps ethernet cards for individual servers within data networks.

Fig. 7 illustrates the structure of a present day warehouse scale computer (WSC). A WSC is a type of data center, that is, an arrangement of processors and memory that can be used to both store data as well as carry out operations on that day bank. WSC's are the elements of what is commonly called the cloud. Various companies use worldwide networks of WSC's to ply their trade. Google uses WSC's for inquiry response, Microsoft rents WSC capacity as a source of income, Facebook uses WSC's to store and retrieve member data whereas Amazon uses WSC's as both a cloud provider and as a sales house.

At the lowest level in the diagram of Fig. 7 are the individual servers, each server consisting or processing and storage. The servers, each of which possessing an ethernet card output, are arranged into 6 foot high racks of perhaps 50 servers per rack. The racks are arranged into clusters and the clusters are arranged to fill a warehouse. The racks are limited to contain roughly 1000 cores (20 cores per server is common). The limit is economic. Cores dissipate circa 10 W and cooling a rack that dissipates more than 10 kW is expensive. The WSC is generally limited to roughly 10,000 racks or 100 MW of dissipation. The limit here the requirement of 250 MW of power (100 MW of dissipation requires 100 MW of cooling and 50 MW of input servers) as power stations larger than 250 MW become exponentially more expensive.

The racks of a data center require interconnection lengths of circa 10 m. A cluster may require lengths up to 100 m. Gigabit ethernet became available already in circa 2000. 1 Gbps ethernet cards already were requiring 100 MHz km bandwidth length products and, indeed, it was at this time that the rack top to rack top interconnections of WSC's first used went optical. As multimode grade index fiber have achieved bandwidths of 1 GHz km, the original data center links consisted of multimode graded in tex fiber with vertical cavity surface emitting laser (VCSEL) transmitters in the 0.85 μm windows and silicon detectors. In the time since 2000, ethernet cards for servers first increased in maximum rate to 10 Gbs and most recently to 25 Gbps. The 10 Gbps rates are still primarily multimode VSCEL driven links. With the 25 Gbps, links are being to migrate to single mode fiber (SMF) in the second and third windows with InGaAsP sources and detectors. The move to SMF is also seeding efforts to increase rates to 100 Gbps though WDM.

At present, about 70% of all internet traffic passes through a data center. This percentage in increasing as more businesses are moving to the cloud to eliminate the need to maintain enterprise data centers. The WSC's as of 2015 consume roughly 1.8% of all US electrical power (Shehabi *et al.*, 2016) and internet traffic is increasing at a rate of circa 22% per year. The energy use is so pervasive that new efforts to increase throughput are now inevitably coupled with efforts to improve energy efficiency. Such efforts are bound to include increasing levels of TDM and WDM that in turn will make optics viable at shorter interconnection length scales. The optical revolution continues.

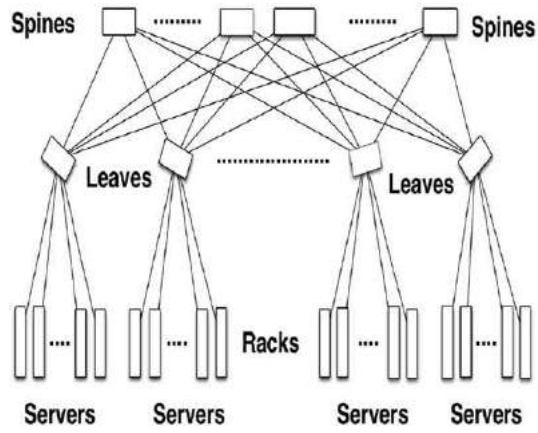


Fig. 7 A sketch that illustrates the interconnection structure of a warehouse scale computer (WSC). In the WSC, racks of servers are interconnected with each other through rack top switches. Each server has an ethernet card that is connected with a copper cable to the rack top switch. The rack top switches are then connected to higher level switches with optical fibers that have transceivers on either end of the fiber connection. The transceivers contain optical sources, modulators that are driven by the information stream from the computer to generate the optical signal that is to be transmitted to the distant location.

Summary

Although optical guidance was both understood and applied in the 19th century, the development of the low loss (less than 100 dB/km) optical fiber (Kapron *et al.*, 1970) and the room temperature semiconductor laser in 1970 (Alferov *et al.*, 1970; Hayashi and Panish, 1970) nucleated a new field of transmission technology, that of fiber optics. The first system demonstration of the technology already occurred in 1975 (Hecht, 2004). Optical fibers come in two primary varieties, multimode and single mode. Multimode fiber is easier to attach to other components than single mode but suffers from much higher dispersion. That is, the information bandwidth that a multimode fiber can propagate over any length is limited and can not easily be extended. Single mode fiber (SMF) not only has much larger bandwidth length product but exhibits dispersion of a type that can be compensated. Hence, SMF is the choice for telecommunications. And telecommunications is the field that employs more fiber than any there field because of the distances involved.

Fiver compatible components include light emitting diodes, semiconductor lasers, optical amplifiers, optical detectors and external modulators, primarily Mach-Zhender (MZ) modulators. Again, to the present, telecommunications has been the most high volume of optical component application areas. Telecommunication components operate in three windows, one around 0.85 μm, 1.3 μm and 1.55 μm (Agrawal, 2010). In the first window, the lasers are GaAs, detectors are silicon and most of the fiber is multimode. In the second and third windows, the laser and detectors are primarily InGaAsP and much of the fiber is single mode. Optical amplifiers operate in the third window and are generally rare earth doped and optically pumped. Date rates in this third channel can now (2017) be as high as 100 Gbps per channel with 32 channels of WDM in the C-band. These numbers are likely to continue to increase.

Data communications, especially within warehouse scale computers (WSC's), is switching to optics. The original data link employed data rates that could be supported over the required lengths by multimode fiber in the first window. The required rates for rack to rack interconnections have already increased to the point where second and third window components with WDM and SMF are necessary. It is expected that data rates will continue to increase and that optics will therefore become viable for shorter and shorter length links.

See also: Fabrication of Optical Fiber

References

Agrawal, G.P., 2010. Fiber Optic Communication Systems, fourth ed. Wiley.
Alferov, Z.I., Andreev, M.V., Portnoi, E.L., Trukan, M.K., 1970. Alas-gaas heterojunction injection lasers with a low room-temperature threshold. Soviet Physics Semiconductors 3, 1107–1110.
Colladan, J.-D., 1842a. Sur les reflexions d'un rayon de lumiere a l'interieur d'une veine liquide parabolique. Comptes Rendus 15, 800–882.
Colladan, J.-D., 1842b. La fontaine colladon. La Nature 15, 525–526.
Hayashi, I., Panish, M.B., 1970. Gaas-gaalas heterostructure injection lasers which exhibit low thresholds at room temperature. Journal of Applied Physics 41, 150–163.
Hecht, J., 2004. City of Light: The Story of Fiber Optics, Expanded, revised ed. Oxford University Press.
Kapron, F.P., Keck, D.B., Maurer, R.D., 1970. Radiation losses in glass optical waveguides. Applied Physics Letters 17, 423–425.
Shehabi, A., Smith, S., Dale, S., *et al.*, 2016. United states data center energy usage report. Technical Report LBNL-1005775, Ernest Orlando Lawrence Berkeley National Laboratory, June.

Optical Fiber Gratings

Paul S Westbrook and Tristan Kremp, OFS Fitel, LLC, Somerset, NJ, United States

Nomenclature

ASE Amplified spontaneous emission
CMT Coupled mode theory
CPA Chirped pulse amplification
CW Continuous wave
DFB Distributed feedback
FBG Fiber Bragg grating
HR High reflector
IR Infrared
LPG Long period grating
LP Linearly polarized
MPI Multipath interference
OC Output coupler
TAP Turn around point
UV Ultraviolet
WDM Wavelength division multiplexing

Fiber gratings are periodic index variations inscribed into the core of an optical fiber and are important devices for manipulating fiber guided light. Since their first fabrication in the late 1970s, fiber grating use and manufacture has increased significantly and they are now employed commercially as in-fiber optical filters and reflectors in telecommunications systems, fiber lasers, and sensors. We review optical fibers and the fundamentals of fiber grating characteristics as well as fabrication. Finally we consider several of the major industrial and scientific applications of fiber gratings. Fiber gratings have been reviewed in two monographs (Kashyap, 2010; Othonos and Kalli, 1999) and various review articles (Hill *et al.*, 1997).

Fiber Modes

A fiber grating couples light between the modes of an optical fiber, and therefore a knowledge of these modes is essential to understanding the characteristics of fiber gratings (Kogelnik, 1990; Marcuse, 1974; Adams, 1981). In what follows, we emphasize single mode fibers used near 1550 nm because these are very important in most practical applications that use fiber gratings. A single mode fiber guides light in a raised index core region, typically germanium doped silica, surrounded by a silica cladding layer and protective polymer coating. The core diameter of a few microns is sufficiently small that only a single transverse mode is guided through total internal reflection at the core-cladding boundary. If the index is only slightly higher in the core than in the cladding, then the electric field (E-field) of the fundamental mode propagating in the core may be described with a characteristic transverse E-field profile and a propagation constant k:

$$E = E(\rho,\theta)e^{-i(\omega t - kz)}, \qquad k = 2\pi n_{eff}/\lambda \tag{1}$$

Here, ρ and θ are the radial and azimuthal positions, z is the longitudinal position along the fiber, ω is the radial frequency, λ is the vacuum wavelength, and we assume that the E-field is linearly polarized (LP), i.e., it points in the same direction everywhere. As the second equation indicates, the fiber mode propagates in the fiber as though it has an effective refractive index n_{eff}.

The silica cladding acts to isolate the core mode from the outside, hence the core mode penetrates very little into the cladding. However, if the outer cladding surface allows total internal reflection, the cladding may guide light as well. Such "cladding modes" are also important in understanding fiber gratings, since the grating couples the core mode with the cladding modes. As shown in **Fig. 1**, the cladding modes extend outside the core into the cladding region, and their effective index is therefore always lower than that of the core mode. We also note that in the case when the outer coating of the fiber does not support total internal reflection, "leaky" cladding modes result. In the limit that there is no reflection at the cladding boundary, a continuum of "radiation modes" results. A grating couples power from the core mode to the cladding/radiation modes, which exist regardless of the presence of the grating. The grating is typically only a small quasi-periodic perturbation on the index profile of the waveguide and the corresponding mode fields.

Fiber Grating Theory

Fiber gratings are longitudinally periodic variations in the refractive index (or, more generally, the electric permittivity) of the core and/or cladding of an optical fiber. The scattering from any grating may be understood simply by considering the scattering off of each successive period and adding these contributions while taking the phase of the E-field into account. **Fig. 2** illustrates this analysis in the case of a fiber Bragg reflector. When the reflections from each period of the grating add constructively, a strong back reflection results. This is known as Bragg reflection. The condition for Bragg reflection is given by:

$$\lambda_{\text{Bragg}} = 2n_{\text{eff}}\Lambda_{\text{grating}} \tag{2}$$

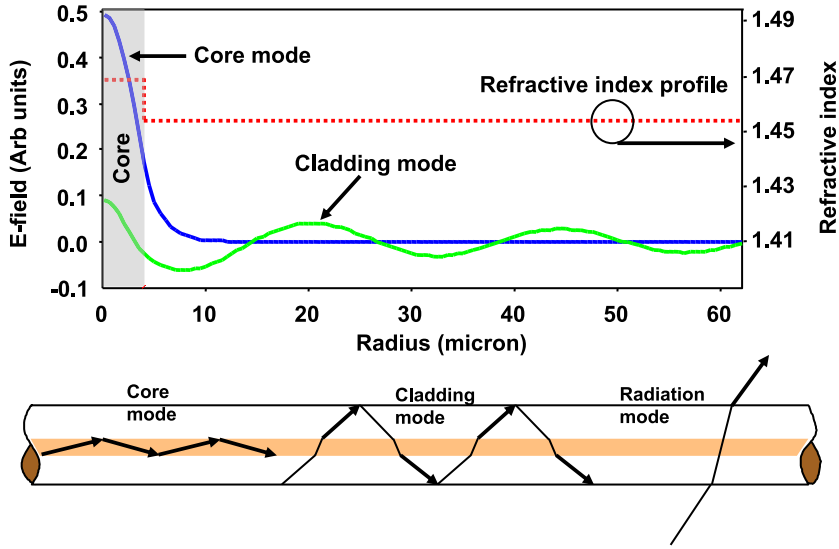


Fig. 1 Fiber modes. Refractive index profile of a step index, single mode fiber (dashed line). Radial profiles of mode E-fields of the fundamental (LP01) core mode, guided within the raised index region defined by the core, and one of the many higher order cladding modes guided by the outer, air-silica interface, assuming an uncoated fiber. Also shown is a fiber indicating the ray paths for core, cladding and radiation modes.

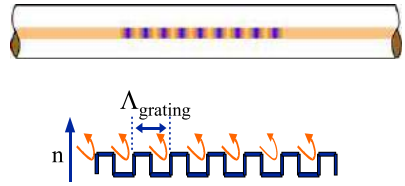


Fig. 2 Bragg reflection. Each period of the grating reflects a small fraction of the incident light. When the period of the grating is adjusted so that these reflections add coherently, a strong Bragg reflection results and a narrow dip is observed in the transmission spectrum of the fiber (see resonance near 1557 nm in **Fig. 3(a)**). Bragg reflection can also occur into cladding modes as shown in the resonances below 1548 nm in **Fig. 3(a)**.

Here, λ_{Bragg} is the vacuum wavelength of the light (corresponding to a radial frequency $\omega_{\text{Bragg}}=2\pi c_0/\lambda_{\text{Bragg}}$), c_0 is the speed of light in vacuum, n_{eff} is the effective index of the mode, and Λ_{grating} is the period of the grating refractive index modulation. This Bragg reflection is evident in a narrow dip in the transmission spectrum of the grating, see **Fig. 3(a)**. The ability of a fiber grating to selectively reflect the core light at a particular wavelength is one of its most important practical properties since it can be used as a narrow-band filter or reflector. As a fiber component, the fiber grating has the advantage of being compact and providing in-line wavelength filtering.

The condition for Bragg reflection may be expressed more generally as a phase matching condition. Phase matching refers to the condition when the phases of all the grating-scattered waves are the same and there is constructive interference. More specifically, this condition expresses the relationship between the spatial frequencies of the incident light, the scattered light, and the grating modulation, necessary for constructive interference in the scattered light. For a grating, the phase matching condition may be expressed in terms of the grating wave vector, $K_{\text{grating}}=2\pi/\Lambda_{\text{grating}}$ and of the two mode propagation constants, k_{incident} and $k_{\text{scattered}}$ (defined in Eq. (1))

$$k_{\text{incident}} - K_{\text{grating}} = k_{\text{scattered}} \tag{3}$$

Using phase matching, we can understand the various transmission spectra that result from fiber gratings. **Fig. 3** shows the transmission characteristic observed for core guided light in a single mode fiber containing three different types of gratings: Bragg reflector, radiation mode coupler, and co-propagating mode coupler. Each grating has a different type of phase matching condition. **Fig. 3** also depicts the phase matching condition in Eq. (3) graphically. The vertical line contains a mark representing the propagation constants, k, of modes of a single mode fiber (both positive and negative for the two propagation directions) and a solid black region representing the radiation mode continuum. Bragg reflection into backward propagating core and cladding modes occurs for short grating periods ($K_{\text{grating}}\sim 2k_{\text{core}}$). For a Bragg resonance near 1.5 μm, this period is roughly 500 nm. Such gratings are known as Fiber Bragg Gratings (FBGs). A typical transmission spectrum with a strong Bragg reflection and several cladding mode resonances is shown in **Fig. 3(a)**. Since the effective indices of cladding modes are lower than the effective index of the core mode, the Bragg wavelength $\lambda_{\text{Bragg}}=(n_{\text{eff,incident}}+n_{\text{eff,scattered}})\Lambda_{\text{grating}}$, which we obtain from Eq. (3) as a generalization of Eq. (2), is shorter for core-cladding mode reflection in comparison to core-core Bragg reflection for any given grating period

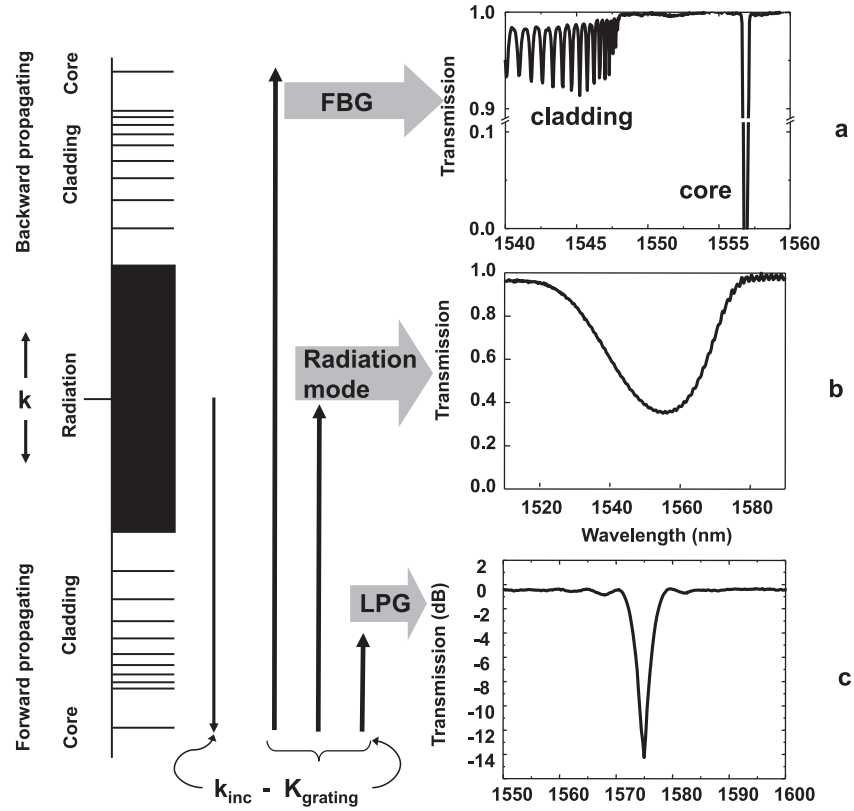


Fig. 3 Phase matching. Fiber grating resonances occur when the phase matching condition is satisfied: $k_{inc} - K_{grating} = k_{scat}$. Propagation constants ($k = 2\pi n_{eff}/\lambda$) of the fiber modes are indicated as horizontal lines on the vertical line. The black region indicates the regime of unguided radiation modes. The fiber grating is represented by its longitudinal spatial frequency (wave vector) $K_{grating} = 2\pi/\Lambda_{grating}$, where $\Lambda_{grating}$ is the period of the grating. Assuming an incident forward going core mode, the grating wave vector then phase matches to (a) counter propagating core and cladding modes (Fiber Bragg Grating, FBG), (b) continuum of radiation modes, or (c) a co-propagating cladding mode (Long Period Grating, LPG).

$\Lambda_{grating}$. When the grating period is very long ($\Lambda_{grating} \sim 1000$ times that of an FBG, typically hundreds of microns), phase matching to co-propagating cladding modes occurs. Such gratings are known as Long Period Gratings (LPGs). A typical LPG spectrum is shown in **Fig. 3(c)**. The LPG resonance corresponds to coupling between the core mode and a single co-propagating cladding mode. For wave vectors in between these two regimes, coupling to the continuum of radiation modes is possible. An example of such a grating spectrum is shown in **Fig. 3(b)**. We note that fiber cladding modes can be either well guided (resulting in the sharp grating resonances of **Fig. 3(a)**) or poorly guided (resulting in a broad continuum as in **Fig. 3(b)**), depending on the fiber surface. The radiation mode approximation is valid in the limit that the cladding mode is very poorly guided (i.e., has high loss) within the grating region.

A fundamental difference between FBGs and LPGs is the direction of the scattered mode, i.e., the sign of $k_{scattered}$ in Eq. (3). In the case of LPGs, $k_{scattered}$ has the same sign as $k_{incident}$, corresponding to a co-propagating mode. Thus, the grating wave vector $K_{grating} = 2\pi/\Lambda_{grating} = k_{incident} - k_{scattered}$ is orders of magnitude smaller than in the case of FBGs, where the counterpropagating $k_{scattered}$ has the opposite sign of $k_{incident}$. As we discuss below, this effect also makes LPGs more sensitive to modal dispersion.

While phase matching yields the *resonance wavelengths* of a fiber grating using only the effective indices, the *amplitudes* of these resonances depend on the length and profile of the gratings as well. These amplitudes must be derived from Maxwell's equations or a sufficiently accurate approximation. The most common approximation used in grating calculations is known as Coupled Mode Theory (CMT) (Kashyap, 2010; Othonos and Kalli, 1999; Kogelnik, 1990). In CMT, the longitudinal refractive index variation is assumed to be a small perturbation to the transverse refractive index distribution. Hence, in the framework of a first-order perturbation analysis, only the propagation constants of the modes change, but not the transverse structure of the E-fields of the modes (Eq. (1)). CMT then assumes that the E-field amplitude of each mode varies slowly (compared to a wavelength) along the grating. A pair of coupled equations relating the two mode amplitudes as a function of position along the grating can then be derived, and these allow for computation of the grating reflection and transmission. The coupling between two modes is determined by an overlap integral κ_{12} between the product of the E-fields of the two modes (E_1 and E_2) and the periodic refractive index variation that defines the grating:

$$\kappa_{12} \sim \iint_{Area} E_1(\rho,\theta) E_2^*(\rho,\theta) \delta n(\rho,\theta) \mathrm{d}A, \tag{4}$$

where $\delta n(\rho,\theta)$ is the transverse part of the uniform grating refractive index modulation

$$\delta n(\rho,\theta,z) = \delta n(\rho,\theta)\cos\left(K_{\text{grating}}z\right) \tag{5}$$

For sufficiently weak gratings, this overlap integral allows for a comparison of the strength of resonances for different modes without the need to obtain a full CMT solution. In the simple case of Bragg reflection of the core mode in a standard single mode fiber (**Fig. 1**), the mode coupling coefficient κ_{12} becomes:

$$\kappa = \frac{\pi\eta\delta n}{\lambda} \tag{6}$$

where $\eta \approx 0.8$ is the overlap integral of the core mode E-field with itself within the core region that has the grating index modulation.

To understand fiber grating spectra, we first discuss a uniform grating of length L and period Λ. By definition, such a grating has uniform index modulation along its length, which is defined in Eq. (5) and illustrated in **Fig. 4(a)** and **(b)**. Depending on the length and refractive index modulation of the grating, it may be classified as either weak or strong. The division between these two regimes may be understood by considering the light scattered by each successive period of the grating. The fraction of light

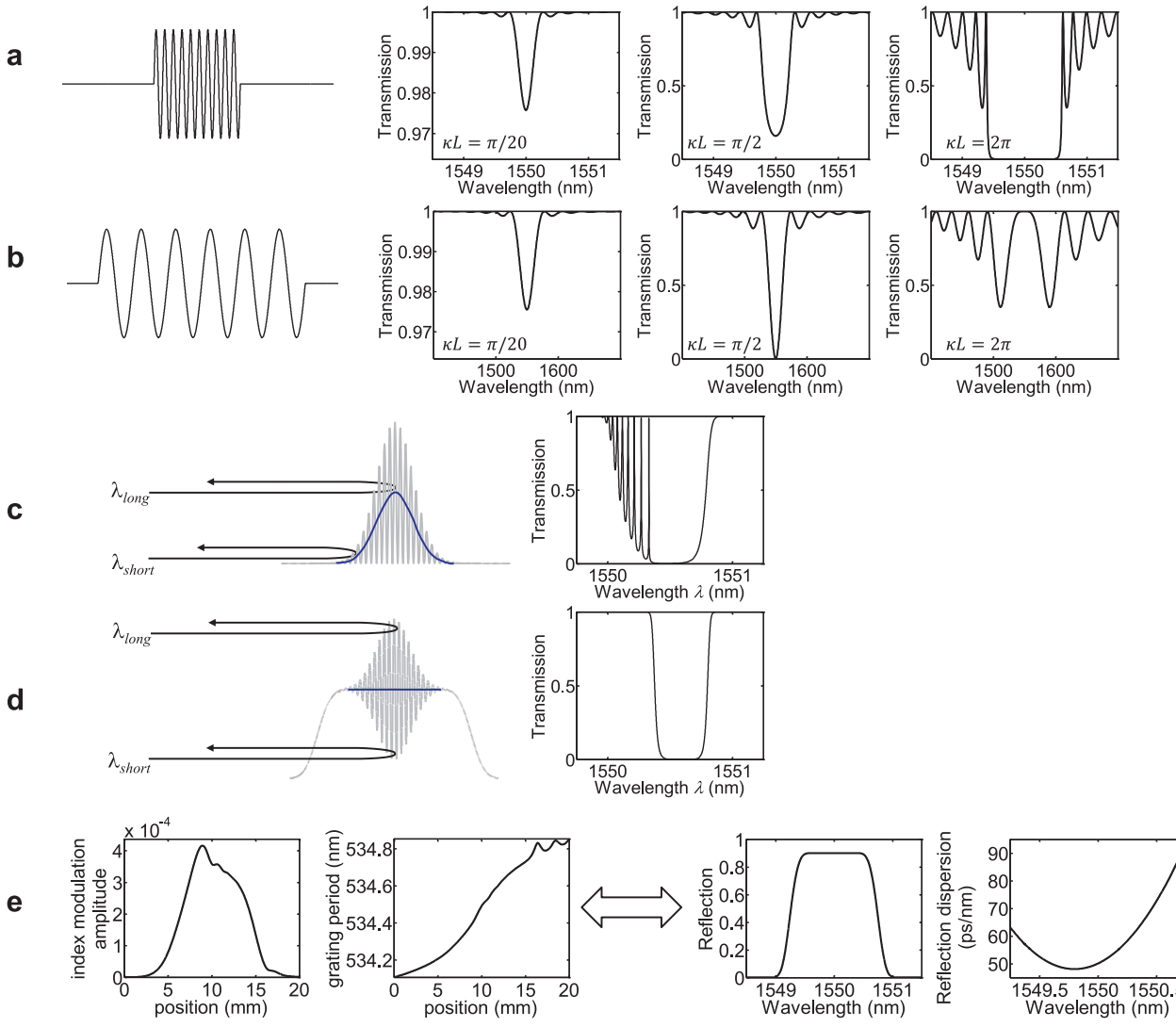


Fig. 4 Fiber grating spectra. (a) Uniform Bragg grating of length $L=3$ mm with period $\Lambda=534.5$ nm and (b) uniform long period grating of length $L=3$ cm with period $\Lambda=534.5$ μm, both with increasing levels of refractive index modulation, given by κL. (c) An "AC apodized" Gaussian grating profile of length $L=3$ cm and strength $\kappa L=8\pi$ showing Fabry-Perot like structure within the grating resonance. (d) A "100% DC apodized" Gaussian grating profile of length $L=3$ cm and strength $\kappa L=8\pi$ showing a symmetric transmission spectrum. (e) Complex grating profile and spectrum designed using inverse scattering. The two plots on the right show the desired flat reflection amplitude and nonuniform chromatic dispersion of reflected light. The two plots on the left show the computed profile of the grating index modulation and grating period required to obtain the desired spectrum.

scattered from each period can be approximated by $\delta n/2n$ using the Snell's Law reflectance formula. The boundary between a strong and a weak grating occurs when the grating is either long enough or strong enough that the sum of the scattering from all periods (L/Λ) is of the order 1: $(L/\Lambda)(\delta n/2n) \approx 1$. This division may also be expressed in terms of κ and L: $\kappa L \approx 1$ (here we are neglecting a factor $\pi/2 \approx 1$).

In the weak limit, the incident light propagates through the grating with very little scattering. In this case, both the LPG and FBG resonances have transmission spectra that approximate the Fourier transform of the grating index profile (See discussion below on inverse scattering. Also, for LPGs, this may not hold if modal dispersion is large; see the discussion below.). In the case of a uniform grating profile, this results in a 1-sinc^2-shaped transmission spectrum with a spectral width $\delta\lambda/\lambda \sim \Lambda\ /L = 1/N_{\text{periods}}$ as shown in **Fig. 4(a)** for a weak FBG and in **Fig. 4(b)** for a similarly weak LPG. The characteristic side lobes of such a perfectly uniform grating result from the sudden "turn on" of the grating modulation. In the FBG, they can be understood as resonances of the effective Fabry-Perot cavity of length L that is defined by the sharp grating boundaries. Asymptotically, the spectral distance between adjacent side lobes is therefore $c_0/(2n_{\text{eff}}L)$ in frequency and $\lambda^2_{\text{Bragg}}/(2n_{\text{eff}}L)$ in wavelength. In the case of an LPG, the term $2n_{\text{eff}}$, which is the sum of the effective indices of the forward and backward propagating core mode, needs to be replaced by the difference $|n_{\text{eff,incident}} - n_{\text{eff,scattered}}|$ of the effective indices of the co-propagating modes.

In the strong limit, a substantial fraction of light is scattered out of the incident mode, and the FBG and LPG transmission spectra become different. As shown in **Fig. 4(a)**, the FBG spectrum increases to a width determined only by δn: $\delta\lambda/\lambda = \delta n/n$. The incident light penetrates only a short distance ($\sim \lambda/\delta n$) into the grating before being completely reflected. The strong LPG in **Fig. 4(b)**, on the other hand, becomes "overcoupled" as the second mode is reconverted into the incident mode at the center of the spectrum. The result is a higher transmission for the incident core mode (provided that the scattered mode propagates without loss). If we continuously increase κL, the transmission at the exact center of the spectrum of a uniform LPG becomes periodically stronger and weaker, with the period π: At $\kappa L = \pi/2$, $3\pi/2$, $5\pi/2$ etc., the transmission is zero, and at $\kappa L = 0$, π, 2π, etc., the transmission is maximized, see also **Fig. 4(b)**. In contrast, the transmission in the center of a uniform FBG spectrum decreases monotonically with respect to κL, see **Fig. 4(a)**.

In the case of a uniform Bragg grating of length L, the peak reflectivity may computed analytically (Kashyap, 2010; Othonos and Kalli, 1999): $R_{peak} = \tanh^2 \kappa L$, where κ is the coupling constant defined in Eq. (6). This formula is approximately correct for many non-uniform Bragg gratings (e.g., the Gaussian apodized gratings of **Fig. 4(c)** and **(d)**) as well, and, with an appropriate choice of the effective length L, provides a useful estimate of the value of κ and the index modulation of the grating.

While uniform gratings are easy to model and manufacture, most practical fiber gratings have a nonuniform profile. In a nonuniform grating profile, the refractive index modulation amplitude, $\delta n_{\text{AC}}(z)$, and the average effective index value, $\delta n_{\text{DC}}(z)$, as well as the local grating wave vector $K_{grating}(z)$ and phase $\varphi(z)$ may vary along the fiber grating length:

$$\delta n(z) = \delta n_{DC}(z) + \delta n_{AC}(z)\cos\left(K_{grating}(z)z + \varphi(z)\right) \tag{7}$$

By controlling these parameters, a greatly improved filter may be fabricated. **Fig. 4(c)** shows the index profile and the grating transmission spectrum of a strong fiber Bragg grating with an "AC apodized" Gaussian profile. Such a profile results when a grating is inscribed with a spot exposure using a Gaussian beam. Note that due to the smooth increase in modulation amplitude, the sidebands of the uniform grating have been eliminated. However, because the DC index (and hence Bragg condition) is not constant within the grating, a sharp resonant structure is observed within the grating resonance. This structure can be understood as Fabry-Perot resonances in the cavity formed by the nonuniform Bragg condition within the grating. In contrast, **Fig. 4(d)** shows a grating with a "100% DC apodized" Gaussian profile. This grating has a constant DC refractive index in the region where the AC refractive index δn_{AC} is nonzero, and therefore the resonance spectrum is smooth and symmetric. Most useful bandpass filters require a profile close to 100% apodization. Although not shown in **Fig. 4**, a smoothly varying index profile also eliminates the sidelobes in LPG resonances. In practice, LPGs are more often fabricated with a uniform profile because of their greater sensitivity to variations in the phase matching condition within the grating.

Other important types of nonuniform fiber gratings include chirped gratings, phase shifted and superstructure fiber gratings. In a chirped fiber grating, the grating period $K_{grating}(z)$ slowly changes along the length of the grating. This spatial chirp of the grating period induces a proportional spatial chirp of the local grating wavelength and is analogous to the temporal chirp of an increasing or decreasing audio signal. As discussed below, chirped fiber gratings may be used as pulse compressors and chromatic dispersion compensators. In a phase shifted grating, a single discrete phase shift, for example, $\Delta\varphi = \pi$, is introduced in a uniform profile. Such a phase shifted grating can be used as a distributed feedback laser cavity (DFB). Superstructure gratings have a spatially modulated profile and can exhibit the same spectrum at regular frequency intervals determined by the period of the spatial modulation. The properties of nonuniform gratings may be computed using CMT and can also be understood intuitively using "band-diagrams" and an effective index model of the fiber grating response (Poladian, 1993).

While uniform and chirped fiber gratings are useful for most applications, it is also possible to design gratings with complex filter responses using a process called "inverse scattering", see, for example, Buryak *et al.* (2009). This process is usually applied to the design of FBGs, while having only limited use in designing LPGs. In such an algorithm, a given desired phase and amplitude response is used in one of several design methods to obtain the desired filter response. For instance, for some pulse compression applications, it is necessary to have both first and second order chromatic dispersion, while still maintaining a uniform reflection over the spectral bandwidth of the pulse. These parameters can be used to design a specific grating profile that will give the desired chromatic dispersion. **Fig. 4(e)** shows the grating profile and spectrum for such a grating.

Inverse scattering can be computationally intensive due the effect of multipath interference (MPI) within the grating. However, for sufficiently weak or short gratings, MPI effects can be neglected and the grating spatial profile and optical spectrum are more easily related. In this single scattering approximation, the coupled mode equations result in a simple Fourier transform relation between the (in general complex-valued) mode coupling coefficient κ and the dimensionless complex E-field reflection spectrum

$$r(\lambda) = \int_0^L \kappa(z) e^{i4\pi n_{group} z/\lambda} dz \tag{8}$$

where n_{group} is the group index of the propagating mode. This Fourier transform relationship between grating profile and spectrum can be seen in the weak ($\kappa L = \pi/20$) LPG and FBG spectra of **Fig. 4(a)** and **(b)**.

Grating spectra showing radiation mode coupling are unlike FBGs or LPGs because there is little or no conversion of the radiation modes back to the incident core mode. Since the radiation mode spectrum is continuous, the spectrum of a grating that couples to radiation modes also smoothes out into a broad continuum. The grating transmission spectrum depends less on the refractive index profile of the grating than on the overlap of the core mode and grating with the radiation modes (Kashyap, 2010). Radiation mode coupling is greatly affected by tilting the grating planes with respect to the axis of the fiber. Such tilted gratings may be designed as loss filters and optical fiber taps. Grating tilt has also been applied to decrease core mode back reflection (Kashyap, 2010; Othonos and Kalli, 1999).

While the resonance of an FBG is determined by the sum of the effective indices of two modes, an LPG resonance depends on the relatively tiny difference of two effective indices according to Eq. (3). This makes the resonance condition of an LPG very sensitive to any sort of perturbation along the waveguide. For instance, LPG resonances may exhibit a strong dependence on the modal dispersion (or variation of the propagation constant, k, with wavelength). This wavelength dependence can change the phase matching condition enough to dominate the wavelength dependence arising from CMT, making the resonance more or less narrow than predicted by a similar CMT with a linear phase matching curve (Kashyap 2010; Othonos and Kalli, 1999). In an extreme case, very broadband LPG resonances (>100 nm at 1550 nm) may be produced between the core mode and a specially designed higher order mode whose propagation constant varies such that it is resonant with the core mode over a large bandwidth. **Fig. 5** shows the spectrum and phase matching curves for both a standard LPG with a near linear phase matching curve, and a broadband LPG whose phase matching curve exhibits a turn around point (TAP). An LPG with a period and resonance wavelength near the TAP will exhibit a very broadband resonance.

Fiber Grating Fabrication

Photosensitivity

Fiber gratings can be fabricated with various high power light sources, including pulsed and continuous wave (CW) ultra violet (UV) sources, femtosecond infrared (IR) lasers and CO_2 lasers. The key early discovery leading to the study and practical application of fiber gratings was the phenomenon of UV-induced photosensitivity in the conventional germanosilicate optical fibers used in telecommunications systems (Hill *et al.*, 1978). UV irradiation of the germanosilicate core region can produce refractive index changes in the range 0.01 to 0.0001. Such refractive index changes are large enough to produce useful filters and reflectors. The refractive index change is most efficiently produced by irradiation close to 242 nm or beyond 200 nm, where

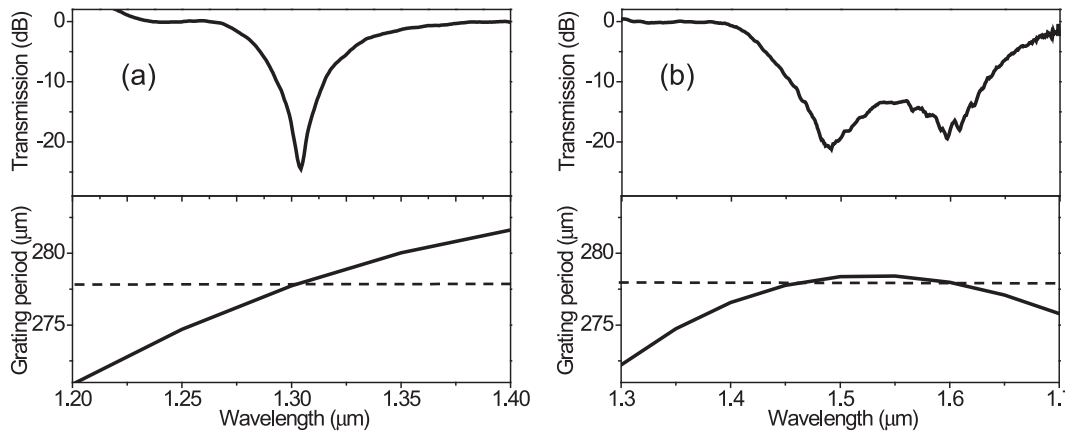


Fig. 5 Long Period Grating (LPG) phase matching curves. LPG fundamental mode transmission spectrum (top) and phase matching curve (bottom) for (a) standard, nearly linear phase matching, and (b) "turn around point" phase matching, resulting in a very broad transmission resonance. The solid line on the phase matching curves gives the resonance wavelength (x-axis) for a grating with a period indicated on the y axis. Dashed line indicates the grating period.

germanosilicate glass has strong absorptions. Numerous other dopants have been shown to be photosensitive at wavelengths ranging from 484 nm to 155 nm, including P, Ce, Pb, Sn, N, Sn-Ge, Sb, and Ge-B. Other dopants, such as Al and F, show very little photosensitivity to UV radiation. With other laser sources such as IR femtosecond and CO_2 lasers, the index modification results from thermal and multiphoton processes which can be used to inscribe gratings in a much larger range of materials, including pure silica. A drawback of this type of photosensitivity, though, is increased loss and more stringent alignment tolerances to yield the required intensity and placement of the grating index perturbations. For CO_2 lasers operating near 10.6 μm, there is a limit to the minimum grating period that can be inscribed. As a result, such lasers are used only for long period gratings.

A second major discovery enabling practical manufacture of fiber gratings was the effect of "hydrogen loading" (Lemaire *et al.*, 1993). Using this procedure, the fiber is placed in a pressurized hydrogen atmosphere until molecular hydrogen has diffused into the core region to a level of a few mole percent (a level similar to that of the photosensitive dopants). UV irradiation of the resulting fiber produces index changes an order of magnitude larger than in the unloaded fiber. Hydrogen loading allows gratings to be easily imprinted in conventional single mode fibers, whose Ge content is too low to allow for strong gratings. The increase in photosensitivity is true for most photosensitive dopants. Index changes in excess of 10^{-2} have been produced with hydrogen loading. In order to avoid increased OH absorption in the telecommunications bands (1520 nm–1630 nm), deuterium (D_2) is normally used in place of hydrogen.

At least two mechanisms are generally agreed to contribute to the UV induced index change. Firstly, as is evident from **Fig. 6**, the UV absorption spectrum is modified during UV exposure. This bleaching corresponds to a change in refractive index in the infrared which may be computed using the Kramers-Kronig relations. In particular, the absorption band at 242 nm resulting from germanium oxygen deficient centers is bleached, indicating that these defects are modified during exposure. The infrared index change results both from this bleaching and from the change in the UV absorption edge below 200 nm. Secondly, absorption of UV causes the germanosilicate matrix to contract. This compaction and the resulting stress distribution also change the refractive index.

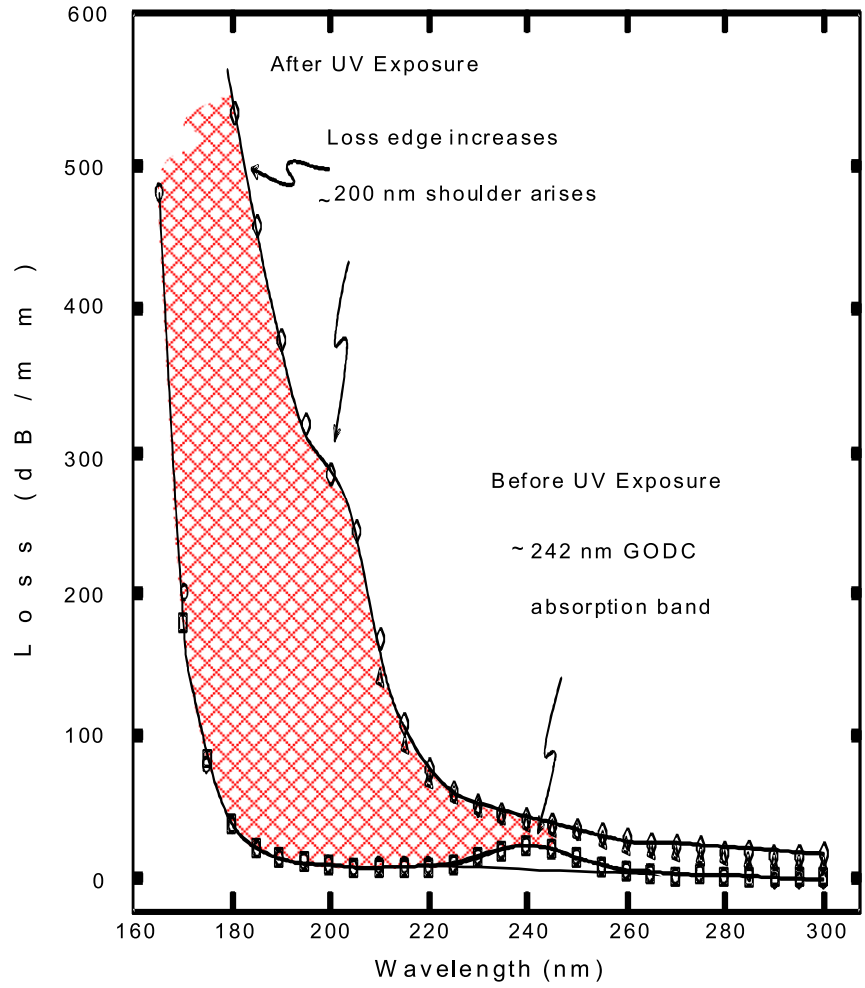


Fig. 6 UV absorption spectrum of Ge doped silica before and after irradiation with 242 nm light, showing bleaching of the Ge-oxygen deficient defect feature at 242 nm as well as a shift in the UV absorption edge.

Both mechanisms are enhanced by the presence of hydrogen in the silica lattice. The hydrogen reacts with oxygen, thus increasing the number of UV absorbing germanium oxygen deficient centers. As a result, in hydrogen loaded fibers, more of the Ge atoms participate in the index change, explaining the large increase in photosensitivity. The number of oxygen deficient defects may also be increased during fiber manufacture by collapsing the fiber preform in a low oxygen atmosphere. Such fibers show increased photosensitivity without the need for hydrogen loading. Co-doping a Ge doped fiber with boron has also been shown to increase the level of photosensitivity of a fiber for a given level of index change in the core.

An important aspect of UV induced index changes is that they require thermal annealing to ensure long term stability (Hill *et al.*, 1997). Sufficient exposure to elevated temperatures can completely erase the refractive index change in some Ge-doped gratings, for example, by several hours of heating to greater than 500°C. Thermal stability has been described by an activation energy model which assumes a distribution of energies for the defects giving rise to the refractive index change. Therefore, at a given anneal temperature, a fixed fraction of the index change will quickly decay, leaving only stable defects that easily survive at lower temperatures. By annealing at several temperatures for a fixed time, a relationship between temperature and time required for a given refractive index decay may then be derived and used in accelerated aging tests of fiber gratings. Annealing conditions vary depending on the application. One typical annealing condition for 20 year reliability is 120°C for 24 h (Kashyap, 2010; Othonos and Kalli, 1999). Gratings written using IR femtosecond radiation can show greater stability that UV written gratings due to the more significant modification that the IR multiphoton process produces in the fiber.

Grating Inscription

In general, fiber gratings are fabricated by exposure of the fiber through the side of the fiber. In the most common method, a UV interference pattern is projected through the side of the fiber onto the core region. The first demonstration of this technique employed a free space interferometer to produce the UV fringes (Meltz *et al.*, 1989). A significant improvement on this technique employs a zero order nulled phase mask to generate the interference pattern. The phase mask is typically a quartz plate with grooves patterned and etched into it using standard semiconductor industry techniques. Such a phase mask splits an incident beam into multiple orders, and these form the interference pattern that is projected onto the core of the fiber. The method is shown in **Fig. 7**. The advantage of the phase mask technique is that the fringes are very stable, since the UV beams propagate only a short distance. Another important technique for writing fiber gratings is the point-by-point or direct writing method. In a point-by-point technique, the UV interference pattern is modulated and translated along a fiber in such a way that a long grating may be formed. Although complex and sensitive, such methods can yield gratings as long as a few meters and also allow for the fabrication of complex grating profiles. Point-by-point methods have been used to fabricate broadband chirped dispersion compensating gratings and Bragg reflectors with very low dispersion.

Both phase mask and point-by-point methods can be used with IR femtosecond writing beams as well (Mihailov *et al.*, 2004). In addition, femtosecond lasers can be used to inscribe the grating line by line with a focused beam that creates a single index modification. Precision translation is required to ensure precise periodicity and a well defined optical spectrum.

LPGs may also be fabricated with UV or IR radiation. An amplitude mask is typically used, though point-by-point methods are also possible. Requirements for beam coherence is greatly reduced because of the long period (typically a few hundred microns). Other techniques may also be used to fabricate LPGs. The fiber may be periodically deformed by heating or with periodic micro-bends. Dynamic LPGs may also be produced by propagating acoustic waves along the fiber (Kashyap, 2010; Othonos and Kalli, 1999).

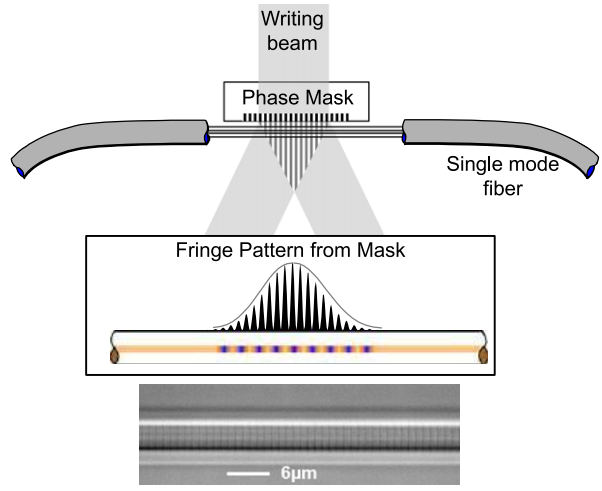


Fig. 7 Phase mask fabrication of a fiber grating. Diffraction by the phase mask grooves yields two beams that form an interference pattern projected into the fiber to write the grating. Bottom: High resolution image of the grating refractive index modulation.

Applications

Fiber gratings are key enabling technologies in a number of important applications. In this section we review these, highlighting the unique functionalities that can be provided by fiber gratings.

Fiber Sensors

Fiber gratings are sensitive to both strain and temperature and may therefore be used as sensors. Both LPG and FBG resonances change with strain and temperature. Since LPGs show greater but more complicated sensitivity, FBGs are typically used in such applications. Strain sensitivity arises from the fact that gratings are extended in length. Because fibers are very thin (diameter $\sim$125 μm), significant extension is possible. A smaller strain sensitivity arises from the stress-optic effect. Typical sensitivity at 1550 nm is $\sim$1.2 pm/1 $\mu\varepsilon$ in FBGs. Temperature sensitivity arises from temperature dependence of the refractive index and gives rise to a change of order 0.01 nm/°C of temperature change of the FBG. FBG strain sensors have been applied to map stress distributions and can be employed in a distributed manner with many Bragg reflectors in one fiber being examined with one detector. **Fig. 8** shows a typical distributed Bragg grating sensor application that uses wavelength division multiplexing (WDM). Every position along the array has a sensor with a different Bragg wavelength, allowing for many sensing points that are mapped onto a spectrum of reflection peaks. Shifts in these peaks correspond to variations in strain at the corresponding sensing point.

These sensitivities may also be used to make fiber grating filters tunable. Temperature changes typically give 1–2 nm of tunability in an FBG resonance at 1550 nm. Strain tuning is more difficult to package commercially but can give up 4 nm in extension and more than 10 nm in compression. Thermal tuning is typically slower than strain tuning. Passive thermal stabilization of grating resonances may also be achieved by designing packages in which strain and temperature variations cancel each other.

It is also possible to inscribe gratings in fiber with more than one core. Such multicore fiber gratings can be used in applications such as shape sensing (Westbrook *et al.*, 2017). Light scattering from the center core and several offset cores is collected in an interrogator and used to reconstruct the shape of the fiber. Twisted offset cores provide sensitivity to the bend and twist direction of the fiber. **Fig. 9** shows a twisted multicore fiber grating. Gratings may be inscribed in such fibers using phase mask inscription and UV irradiation similar to that used in single core gratings.

Fiber Grating Stabilized Diode Lasers and Fiber Lasers

One of the first important commercial applications of fiber gratings was in stabilization of fiber coupled diode lasers. Normally, these lasers operate on several cavity modes making the output unstable. If feedback is provided in the form of a narrow band reflection, then the laser output will be only in this narrow bandwidth. The fiber grating is implemented in the fiber pigtail of the diode laser and typically provides a few percent reflectivity over a bandwidth of $\sim$1 nm. The application is depicted schematically in **Fig. 10(a)**.

More generally, fiber gratings have been used to realize fiber laser cavities. Typically, a laser cavity may be defined by inscribing a high reflector (HR) and an output coupler (OC) at matched wavelengths. The resulting cavity may then be pumped at a wavelength outside the reflection spectra of the narrow HR/OC pair. **Fig. 10(b)** illustrates an example of a fiber laser formed by an HR/OC pair. **Fig. 10(c)** shows an example spectrum of the broadband HR grating and the much weaker OC grating. Another important application is the cascaded Raman resonator fiber laser, in which several FBGs recycle Raman pump light to produce high power, fiber-coupled light at any design wavelength. As described earlier, it is also possible to inscribe distributed feedback (DFB) cavities using fiber Bragg gratings. Such gratings typically have a uniform profile and a single phase shift at the center of the

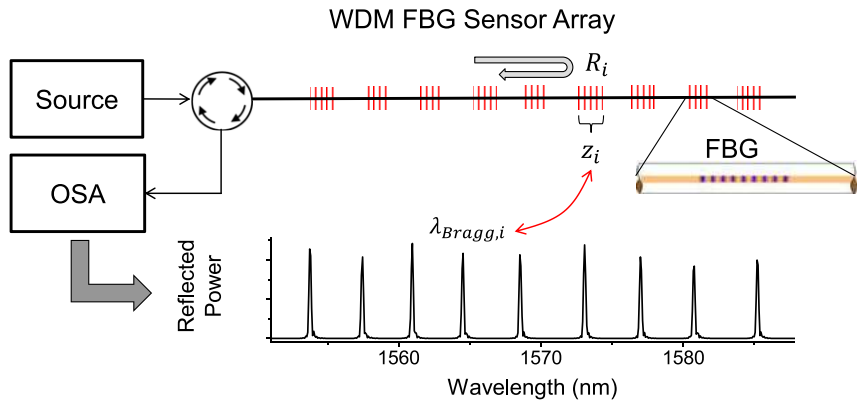


Fig. 8 Wavelength division multiplexed (WDM) fiber Bragg grating (FBG) sensor array. Signals reflected from sensor grating at z_i are measured with the Optical Spectrum Analyzer (OSA) and mapped to Bragg wavelength $\lambda_{\text{Bragg},i}$, whose resonance wavelength shift yields strain and temperature information from position z_i.

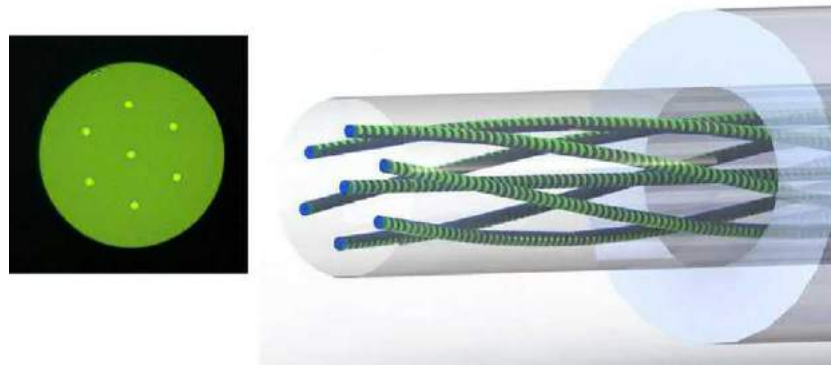


Fig. 9 Twisted multicore fiber with continuous grating sensor array used for fiber bend and shape sensing.

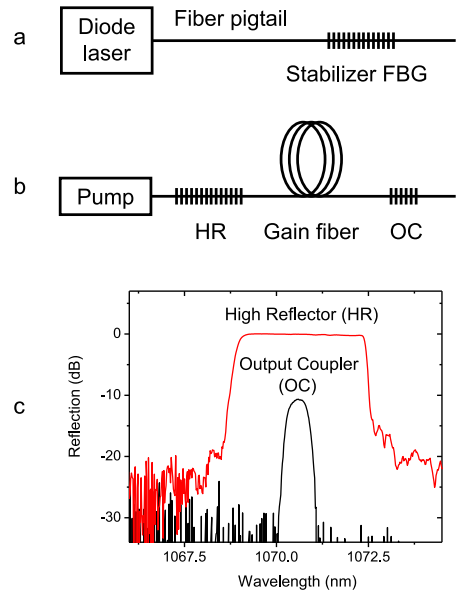


Fig. 10 FBGs in laser applications. (a) Fiber Bragg Grating stabilized diode laser. Fiber grating reflector in the laser pigtail provides small level of feedback to stabilize laser output power and wavelength. (b) Fiber laser cavity defined by a high reflector (HR) and output coupler (OC) grating. (c) Typical spectra for the HR and OC.

grating to define an efficient, low threshold cavity. Fiber DFB lasers can provide excellent performance at extremely narrow linewidths below 1 kHz.

Optical Filtering

Fiber gratings offer unprecedented in-fiber, low loss, filtering of guided wave signals. Filter bandwidths from 10 pm to more than 100 nm are possible. A primary strength of fiber gratings is the diversity of filter shapes that is achievable through modification of the grating profile. One example is a gain flattening filter. Such filters have seen important application in telecommunications systems in order to smooth the wavelength dependence of fiber amplifiers (such as Er or Raman amplifiers). Both LPGs and FBGs may be employed as gain flattening filters. The desired loss spectrum is translated into a corresponding grating profile. LPGs have the advantage of low return loss and are useful for filters with bandwidths larger than a few nm. FBGs may also be used in both tilted and untilted form and are often accompanied by an optical isolator to reduce back reflections. Phase shifted FBGs can be configured with a circulator to provide a very narrow band filter for some applications. Fig. 11 shows an example of a gain equalizing loss filter using LPGs.

Chromatic Dispersion Control

Chromatic dispersion arises from the variation in propagation velocity with wavelength. A temporally sharp pulse is thus dispersed and broadens out in the time domain. Fiber Bragg gratings can provide strong chromatic dispersion when operated in reflection

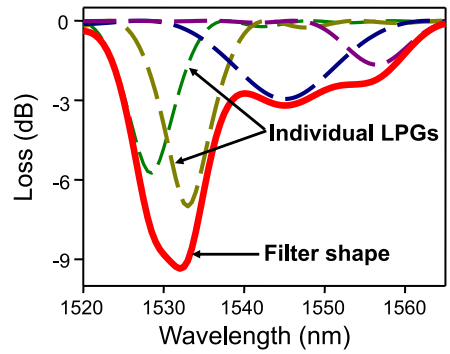


Fig. 11 Long Period Grating gain equalizing filter. Several individual gratings (dashed line) are combined to produce the overall shape (solid line).

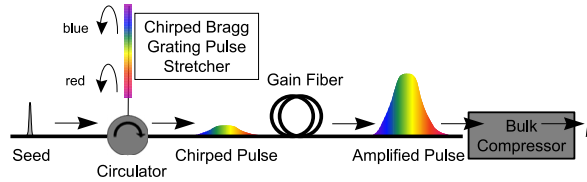


Fig. 12 Chirped pulse amplifier (CPA) using a chirped fiber Bragg grating to stretch the input pulse before amplification.

with a 3-port circulator. In this case, the period is made to vary along the length of the grating. This means that different wavelengths in a pulse reflect at different positions in the grating. This gives rise to wavelength dependent group delay and therefore chromatic dispersion.

Such chirped gratings can be used to compress and disperse pulses. Pulse compression can improve pulsed telecommunications signals that have been broadened by the dispersion of an optical fiber. On the other hand, in high power pulse amplification, such gratings can be used to greatly broaden a pulse, thereby reducing its peak power and allowing for amplification without excess nonlinearities. Such an amplifier is called a chirped pulse amplification system (CPA). **Fig. 12** shows the use of a chirped fiber grating pulse stretcher in a CPA system.

Nonlinear Applications of Fiber Gratings

While most applications of gratings involve linear propagation of light, when the optical power in the fiber is high enough, nonlinear propagation effects are also observed. While such nonlinearities can limit the performance of fiber grating devices, they can also lead to potentially useful effects, some of which we discuss here. Optical nonlinearities arise from the intrinsic nonlinearity of the material system in which the grating is inscribed (Agrawal, 2001). In the case of optical fibers, the dominant nonlinearity is the Kerr nonlinearity, which is a third order effect that manifests as a change in the refractive index of the material that depends on the peak intensity of the optical field: $n = n_0 + n_2 I$. This nonlinearity has been studied in a number effects, including all-optical switching, bistability, multistability, soliton generation, pulse compression and wavelength conversion (Slusher and Eggleton, 2013). Critical to these phenomena is the shift in the grating reflectivity as the input power is increased. Sufficiently high power can result in increased transmission through a strong Bragg grating due to the shift of the Bragg resonance. For certain high power pulses, these propagation effects may result in the formation of a Bragg soliton, i.e., a pulse that maintains its shape during propagation, within the photonic bandgap formed by the fiber grating. The enormous dispersion provided by the fiber Bragg grating is balanced by the fiber nonlinearity in a manner similar to solitons observed in standard single mode fiber where the dispersion is determined by material properties and waveguide design. Since the dispersion of a Bragg grating can be orders of magnitude larger than that of the bare fiber, Bragg solitons can be observed in gratings of only a few centimeters in length. Such Bragg solitons can show a considerable decrease in group velocity, allowing the possibility of an optical delay line.

Another set of nonlinear effects can occur if continuum generation occurs within a fiber Bragg grating. Continuum generation of very intense pulses can occur in a fiber due to the interplay of the full third order nonlinear interaction with the effect of fiber dispersion. For certain input pulsed sources, the resulting continuum is comprised of a comb of individual frequency lines, and such optical frequency combs can be used in precision spectroscopy. When such a process occurs in the presence of a fiber Bragg grating, the formation of the continuum is greatly modified in the spectral region near the grating (Westbrook *et al.*, 2004). The spectral density near the Bragg resonance can increase by more than an order of magnitude due to the interplay of the grating dispersion and the continuum formation, in effect producing a more efficient phase matching of the nonlinear processes in that

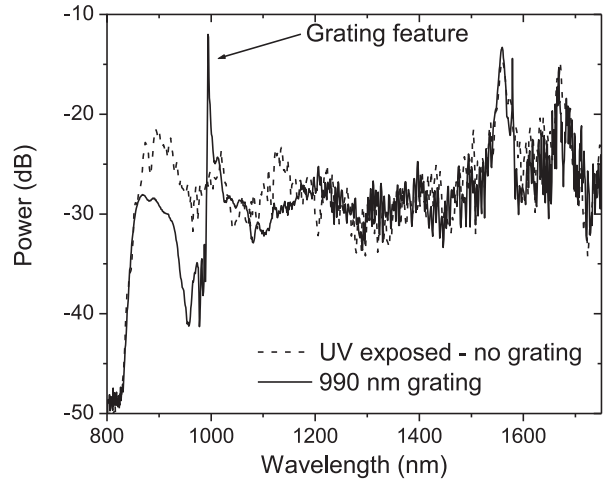


Fig. 13 Grating enhanced continuum. Continuum generation in a fiber with and without a fiber Bragg grating at 990 nm, showing a large enhancement peak in the continuum near the fiber grating resonance.

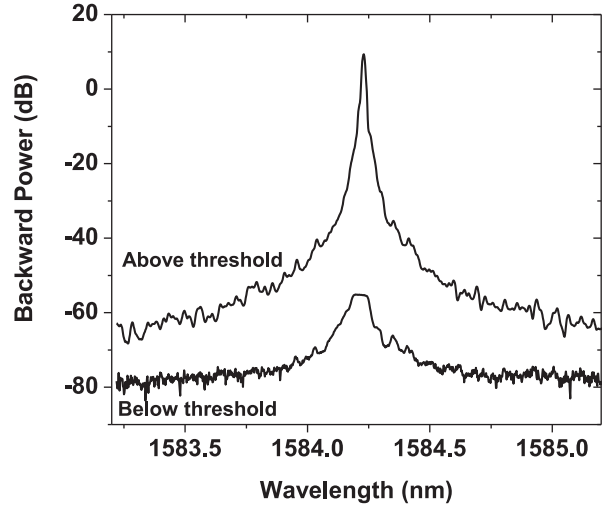


Fig. 14 Raman fiber distributed feedback (DFB) laser. Optical spectrum back reflected from a phase shifted fiber grating at 1584.2 nm pumped by a Raman pump near 1480 nm, showing a Bragg reflection spectrum at low pump power and a narrow linewidth lasing line when the Raman gain exceeds the lasing threshold within the Bragg grating DFB cavity.

spectral region. **Fig. 13** shows an optical continuum with and without a grating present in the fiber. On the long wavelength side of the grating, the spectrum is enhanced by more than a factor of 10. Such grating enhancement peaks have been shown to increase the signal-to-noise ratio by more than an order of magnitude in applications that use optical frequency combs for precision measurements.

It is also well known that fibers exhibit Raman gain, another manifestation of the third order nonlinearity of glass materials. An optical pump at any wavelength will result in Raman gain for a range of wavelengths longer than the pump wavelength. Such gain can be used to produce Raman fiber amplifiers and Raman fiber lasers. Raman lasers are very noisy and exhibit large linewidths, however, it is possible to exploit Raman gain within a fiber distributed feedback (DFB) cavity to achieve narrow linewidth lasing. The DFB cavity is formed by a FBG with a phase shift in the center. **Fig. 14** shows an example of such a Raman DFB fiber laser (Westbrook *et al.*, 2011). The lower trace on the plot shows the light back reflected from the DFB fiber grating at low pump power. Raman amplified spontaneous emission (ASE) generated in the fiber back reflects off of the grating spectrum and reveals a typical flat-topped reflection spectrum. The upper trace shows lasing from the cavity when the Raman pump power is increased beyond the lasing threshold. Raman gain and feedback from the FBG DFB cavity cause light intensity to build up, resulting in a strong and very narrow back reflected signal centered at the Bragg wavelength. Such lasers can generate light over a range of frequencies defined by the pump and DFB Bragg wavelengths and exhibit linewidths orders of magnitude narrower than a typical Raman fiber laser.

See also: Diffraction Gratings. Nonlinear Optics

References

Adams, M.J., 1981. An Introduction to Optical Waveguides, vol. 14. New York, NY: Wiley.
Agrawal, G., 2001. Applications of Nonlinear Fiber Optics. Academic press.
Buryak, A., Bland-Hawthorn, J., Steblina, V., 2009. Comparison of inverse scattering algorithms for designing ultrabroadband Bragg gratings. Optics Express 17 (3), 1995–2004.
Hill, K.O., Fujii, Y., Johnson, D.C., Kawasaki, B.S., 1978. Photosensitivity in optical fiber waveguides: Application to reflection filter fabrication. Applied Physics Letters 32 (10), 647–649.
Hill, K.O., Russell, P.S.J., Meltz, G., Vengsarkar, A.M. (Eds.), 1997. IEEE Journal of Lightwave Technology: Special Issue on Fiber Gratings, Photosensitivity, and Poling 15 (8), 1261–1511.
Kashyap, R., 2010. Fiber Bragg Gratings, second ed. New York, NY: Academic.
Kogelnik, H., 1990. Theory of optical waveguides. In: Tamir, T. (Ed.), Guided-Wave Optoelectronics. Berlin: Springer-Verlag.
Lemaire, P.J., Atkins, R.M., Mizrahi, V., Reed, W.A., 1993. High pressure H/sub 2/loading as a technique for achieving ultrahigh UV photosensitivity and thermal sensitivity in GeO/sub 2/doped optical fibres. Electronics Letters 29 (13), 1191–1193.
Marcuse, D., 1974. Theory of Dielectric Optical Waveguides. New York, NY: Academic.
Meltz, G., Morey, W., Glenn, W.H., 1989. Formation of Bragg gratings in optical fibers by a transverse holographic method. Optics Letters 14 (15), 823–825.
Mihailov, S.J., Smelser, C.W., Grobnic, D., *et al.*, 2004. Bragg gratings written in all-SiO_2 and Ge-doped core fibers with 800-nm femtosecond radiation and a phase mask. Journal of Lightwave Technology 22 (1), 94.
Othonos, A., Kalli, K., 1999. Fiber Bragg Gratings: Fundamentals and Applications in Telecommuncations and Sensing. Norwood, MA: Artech House.
Poladian, L., 1993. Graphical and WKB analysis of nonuniform Bragg gratings. Physical Review E 48 (6), 4758.
Slusher, R.E., Eggleton, B.J. (Eds.), 2013. Nonlinear Photonic Crystals, vol. 10. Springer Science & Business Media.
Westbrook, P.S., Abedin, K.S., Nicholson, J.W., Kremp, T., Porque, J., 2011. Raman fiber distributed feedback lasers. Optics Letters 36 (15), 2895–2897.
Westbrook, P.S., Kremp, T., Feder, K.S., *et al.*, 2017. Continuous multicore optical fiber grating arrays for distributed sensing applications. Journal of Lightwave Technology 35 (6), 1248–1252.
Westbrook, P.S., Nicholson, J.W., Feder, K.S., Li, Y., Brown, T., 2004. Supercontinuum generation in a fiber grating. Applied Physics Letters 85 (20), 4600–4602.

Optical Amplifiers: SOAs

Michael J Connelly, University of Limerick, Limerick, Ireland

Introduction

The rapid growth in the development of optical communication systems requires small, inexpensive and easy to integrate Semiconductor Optical Amplifiers (SOAs) for basic applications such as power boosters, in-line amplifiers and receiver pre-amplifiers. SOAs can also be used to carry out optical signal processing functions such as wavelength conversion, modulation, demultiplexing, switching, logic and regeneration (Connelly, 2002; Dutta and Wang, 2013).

Optical Fiber Amplifiers (OFAs), which cannot be integrated, are often the choice for power, in-line and preamplifier applications where small-size is usually not a critical factor. Their advantages include wide bandwidth (10s–100s of nm), high gain, high saturation output power $P_{o,sat}$, defined as the output power at which the amplifier gain is half the unsaturated gain, low Noise Figure (*NF*) and low polarization sensitivity. Because OFA gain dynamics are very slow, they do not impart significant distortion to amplified optical data signals irrespective of the modulation format; however this precludes their use in optical signal processing applications.

The gain mechanism in SOAs is based on achieving a population inversion between the conduction and valence bands of the amplifier active region material driven by an electrical current. Net amplification results when stimulated emission induced by the amplified lightwave exceeds losses due to stimulated absorption and other material or structural losses. By appropriate choice of the active material SOAs can be designed to operate in the wavelength region of choice, usually one of the optical communications bands in the 1.3 μm and 1.55 μm regions. SOAs have gain, *NF*, $P_{o,sat}$ and bandwidth values comparable to OFAs, as well as low polarization sensitivity. Compared to OFAs, the principle advantages of SOAs are their small size and compatibility with Photonic Integrated Circuits (PICs). SOAs have much faster dynamics than OFAs, which can result in signal distortion; but in association with nonlinearities can be exploited to realize all-optical signal processing functions. Advanced SOAs based on Quantum-Dot (QD) structures have demonstrated processing speeds in the THz range.

This article reviews SOA principles and structures, some simple SOA models that provide insights into SOAs physics and performance, basic network applications and the important all-optical signal processing applications of wavelength conversion, optical logic, data regeneration, optical remodulation and microwave phase shifting.

SOA Principles

The principle of operation of an SOA is shown in **Fig. 1**, where Antireflection Coatings (ARs) are used to suppress end facet reflections that would result in the SOA acting as a Fabry-Perot cavity leading to ripples in the gain spectrum and the possibility of oscillation at high gains. Such SOAs are often referred to as traveling-wave SOAs. The input lightwave is amplified as it propagates through the electrically pumped active waveguide. Although SOA waveguides are designed to be single-mode they support two orthogonal polarization modes, the Transverse Electric (TE) and Transverse Magnetic (TM) modes. Confinement of the signal lightwave to the active region is quantified by the optical confinement factor Γ, defined as the fraction of the transverse (to the propagation direction) optical intensity overlapping with the active region. Γ is usually polarization dependent, so the TE and TM confinement factors Γ_{TE} and Γ_{TM} respectively are unequal except in SOAs having a square cross-section active waveguide; however such SOAs are difficult to fabricate. SOAs usually have rectangular cross-section waveguides so $\Gamma_{TE} \neq \Gamma_{TM}$; a consequence of which is that if the active material gain is polarization independent the SOA gain will be polarization dependent.

A consequence of optical amplification is the addition of broadband Amplified Spontaneous Emission (ASE) noise due to the spontaneous recombination of the active material charge carriers – conduction band electrons and valence band holes. In the linear operating region (low input signal power), the gain and ASE do not depend on the input signal power. In most practical applications a narrowband optical bandpass filter of bandwidth B_o, which passes the amplified signal, is placed after the SOA to reduce the ASE. If the filter is assumed to be an ideal rectangular filter centered at the signal photon energy E_s, the noise power

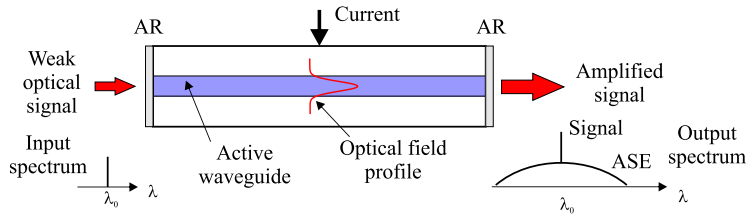


Fig. 1 Basic concept of an SOA.

 doi:10.1016/B978-0-12-803581-8.09462-5

spectral density, in a single-polarization, is given by

$$\rho_{ASE} = n_{sp}E_s(G-1) \tag{1}$$

where G is the gain and n_{sp} is the population inversion factor. A common metric for quantifying the noise properties of an amplifier is the *NF*, defined as the ratio in dB of the input to output Signal-to-Noise Ratios (SNRs) as measured in terms of the signal and noise levels of the current from an ideal photodetector having responsivity R (A/W) and electrical bandwidth B_e; in linear terms the ratio is called the noise factor F. If the input signal power is P_s, its associated electrical current is RP_s. Assuming the input signal is quantum noise (often referred to as shot noise) limited, its current variance is $\sigma_{in}^2 = 2eRP_sB_e$ (A^2) and so the input SNR $SNR_{in} = P_s/(2E_sB_e)$. The output noise current consists of several uncorrelated components: the quantum noise of the amplified signal, the quantum noise of the ASE, the signal-spontaneous beat noise and the spontaneous-spontaneous beat noise, having respective current variances (Olsson, 1989).

$$\sigma_s^2 = 2eRGP_sB_e \tag{2}$$

$$\sigma_{sp}^2 = 2eR\rho_{ASE}B_oB_e \tag{3}$$

$$\sigma_{s-sp}^2 = 4R^2GP_s\rho_{ASE}B_e \tag{4}$$

$$\sigma_{sp-sp}^2 = 2R^2\rho_{ASE}^2B_e(2B_o - B_e) \tag{5}$$

The total current variance $\sigma_{out}^2 = \sigma_s^2 + \sigma_{sp}^2 + \sigma_{s-sp}^2 + \sigma_{sp-sp}^2$. The output SNR $SNR_{out} = (RGP_s)^2/\sigma_{out}^2$. If B_o is small enough the signal dependent noise components dominate, in which case

$$NF = 10\log_{10}\left(\frac{2\rho_{ASE}}{GE_s} + \frac{1}{G}\right) \tag{6}$$

When, $G \gg 1$, the signal-spontaneous beat noise dominates, so

$$NF = 10\log_{10}(2n_{sp}) \tag{7}$$

The maximum possible value of n_{sp} is 1, so the minimum achievable *NF* is 3 dB.

SOA Structures

The key parameters required for practical SOAs are:

- Low end reflectivities (typically $< 10^{-4}$).
- Low polarization sensitivity (< 0.5 dB)
- Wide optical bandwidth (10s nm).
- High gain at low currents.
- High $P_{o,sat}$.
- Low fiber-to-chip coupling losses.
- Fast gain recovery time to reduce amplified signal distortion.

The aim of most SOA structural designs, choice of active region material and packaging is to realize some or all of the above objectives. SOAs can also be incorporated into complex configurations using PICs or discrete optics to perform optical signal processing functions, for which it may be desirable for the SOA to possess enhanced nonlinearities. In this section a bulk material SOA is described, which illustrates practical SOA operation, as well as SOAs based on semiconductor nanostructures that have several important advantages over bulk devices.

Bulk SOA

A typical SOA, fabricated from bulk material, operating in the 1550 nm region, is shown in **Fig. 2**. The active region is sandwiched between two Separate Confinement Heterostructure (SCH) layers having refractive indices less than the active region refractive index, which provides waveguiding, i.e., helps prevent the propagating lightwave from spreading into the surrounding lossy regions. The p-n junctions formed by the p- and n-type InP layers act as current blocks and provide good confinement of the injected carriers in the active region. Tensile stain is introduced between the active region and SCH layers via lattice mismatching. This increases the ratio of the TM to TE material gain coefficients to compensate for the higher Γ_{TE} caused by the waveguide asymmetry thereby reducing polarization sensitivity. The tapered regions act as mode-expanders that couple light from the active waveguide to an underlying passive waveguide to simplify coupling of the amplified light to lensed optical fibers. Low end reflectivities ($< 10^{-5}$) are obtained by combining buried windows and a 7° waveguide tilt angle with respect to the AR coated end facets.

Typical tensile-strained SOA static characteristics are shown in **Fig. 3**. The small-signal gain spectrum and ASE spectrum (**Fig. 3 (a)** and **(b)**) at a given current have similar shapes near the peak wavelength but are significantly different as the wavelength deviation from the peak increases. At high bias currents the 3 dB gain bandwidth is approximately 65 nm. At high currents the gain

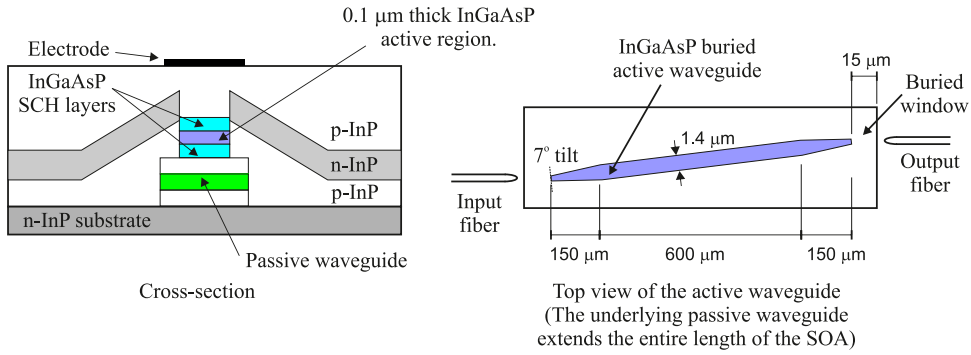


Fig. 2 Tensile-strained bulk SOA with some typical dimensions.

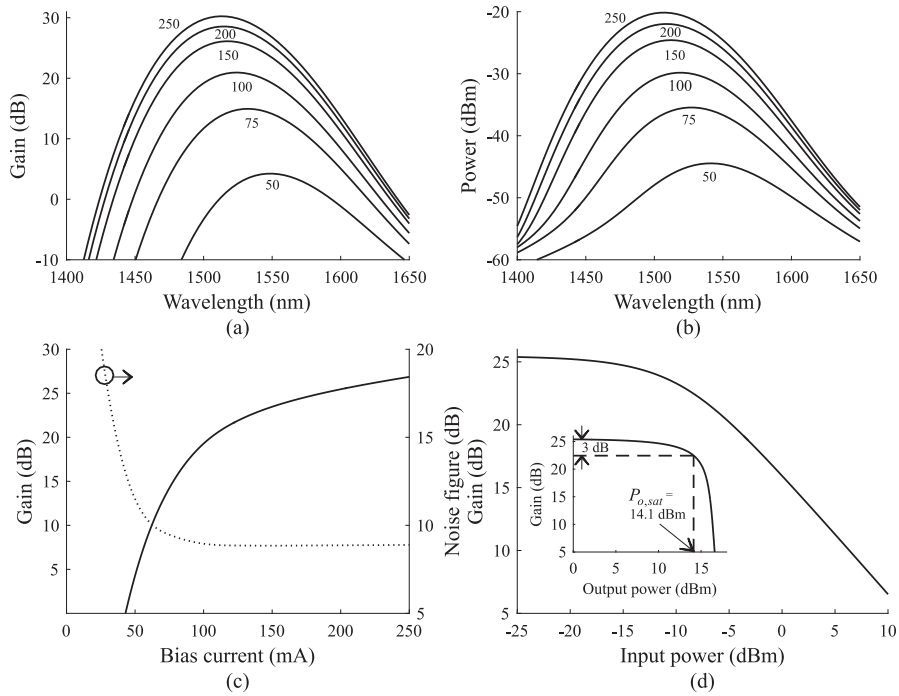


Fig. 3 Typical SOA experimental static characteristics. (a) Small-signal gain vs. wavelength and (b) output ASE spectrum with 0.07 nm resolution. The parameter is the bias current in mA. (c) Small-signal gain and NF vs. bias current at 1550 nm. (d) Gain vs. input signal power at 1550 nm and a bias current of 200 mA. The inset shows the gain vs. output signal power and the value of $P_{o,sat}$.

saturates because the ASE depletes the conduction band electrons, which are then not available to impart gain to the signal. Saturation is also evident in the small-signal gain versus bias current characteristic shown in **Fig. 3(c)**. The *NF* decreases as the gain increases as shown in **Fig. 3(c)**, which includes a degradation of 2 dB due to the output fiber coupling loss. The SOA gain dependency on input power is shown in **Fig. 3(d)**; the inset shows the gain as a function of output power from which $P_{o,sat}$ can be determined.

Semiconductor Nanostructure SOAs

In a bulk SOA, the motion of the carriers is not restricted. Placing semiconductor layers or structures with a reduced dimensionality in the active region results in a restriction of the carrier motion and has a very significant impact on SOA behavior. The motion of a carrier is restricted if the material dimension is of the order of the carrier de Broglie wavelength – of the order of 7–70 nm in typical semiconductors. Reducing the material dimension in one (Quantum-Well, QW), two (Quantum-Wire) and three dimensions (quantum-dot) results in carrier confinement in 1D, 2D and 3D respectively. Reduced dimensionality semiconductors are referred to as semiconductor nanostructures and are usually embedded in a bulk semiconductor of larger bandgap energy. A Quantum Dash (QDash), or short wire, is an elongated nanostructure, having a cross-section similar to a QD but with a length of typically 100 s of nm (Lelarge *et al.*, 2007).

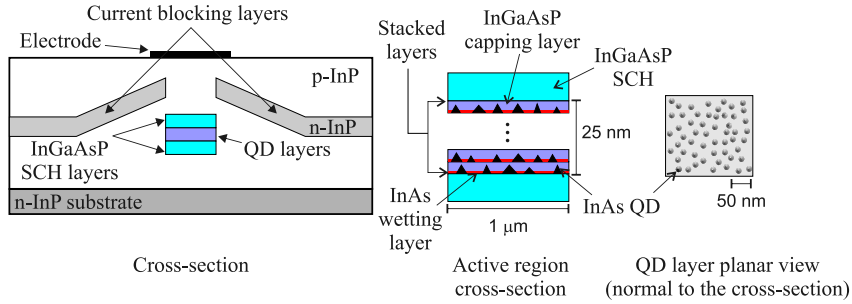


Fig. 4 QD SOA structure with some typical dimensions (not to scale). Typical dot densities are $(2-10)\times 10^{24}$ mm^{-2}.

The active region of a single QW SOA is very similar to that for a bulk SOA except that the thickness is much lower. This permits many QWs to be stacked to form a Multiple-QW (MQW) active region, to mitigate the effect of the reduced single QW confinement factor and thereby increase the gain coefficient. MQW SOAs have a number of advantages compared to bulk SOAs. Because of the small active region volume compared to bulk SOAs, a reduced injection current is sufficient to create a large carrier density resulting in a broader gain spectrum and shorter carrier lifetime τ_c (time for conduction band electrons to recombine with valence band holes). Lower τ_c values result in reduced gain recovery times, which is of particular importance in reducing pattern effects when the SOA is used to amplify high-speed data signals. The loss coefficient in MQW active regions is significantly smaller than that in bulk devices, which leads to an improvement in noise performance. The intrinsic polarization sensitivity of QW structures can be reduced by using strained QWs. Strain-induced band structure modifications also result in a reduction in loss mechanisms such as Auger recombination and intervalence band absorption (Chuang, 2009). The Linewidth Enhancement Factor (LEF), a proportionality factor relating phase changes to gain changes is also reduced, which results in amplified pulses experiencing less spectral broadening, leading to superior high-speed performance compared to bulk and unstrained QW SOAs. Bulk and QW SOAs have typical gain recovery times in the range of 100 s picoseconds so, when used to amplify data signals, significant pattern effects can be present at symbol rates above 10s Gbaud/s. Pattern effects also induce Pattern Dependent Phase Distortion (PDPD) due to Self-Phase Modulation (SPM). SPM in an SOA arises when a high intensity signal depletes the carrier density, which leads to a modification in the refractive index and thereby the instantaneous phase of the propagating lightwave. PDPD can be a more serious performance degradation factor than pattern effects in coherent optical communication systems employing advanced multi-level modulation formats such as Phase Shift Keying (PSK) and especially Quadrature Amplitude Modulation (QAM). The amount of SPM is directly related to the LEF.

A Quantum-Dot (QD) is a semiconductor nanostructure having dimensions that typically range from 2 to 10 nm, which confines the motion of injected carriers to all three spatial directions. In QD SOAs (Akiyama *et al.*, 2007) the device structure for implementing electrical pumping and waveguiding, surrounding the QD material active region is very similar to that for bulk and QW SOAs as shown in **Fig. 4**. The active region consists of stacked layers of QD material fabricated using the Stranski-Krastanov (SK) growth technique that results in self-assembled QDs. Initially, a lattice matched strained InP 2D film (wetting layer) is formed. The accumulated strain is relieved by the random formation of 3D InP islands. The growth of an InGaAsP capping layer completes the QD formation. If the capping layer has a wider bandgap than the substrate a QW structure around the dots is formed, which extends the emission wavelength of the dots thereby leading to a larger gain bandwidth. Self-assembled SK grown QDs are inherently polarization dependent having only TE mode gain. Low polarization sensitivity QD SOAs have been realized by using tensile-strained capping layers or through the use of optimized QD shapes. Compared to bulk and MQW SOAs, QD SOAs have shown to have significantly improved gain bandwidth (100s nm), $P_{o,sat}$ (>25 dBm) and NF(~ 5 dB). The former two properties are especially important when the SOA is used to simultaneously amplify many channels as is the case in Wavelength Division Multiplexed (WDM) systems. Of particular significance is the short τ_c typically a few picoseconds, which is a factor of ten lower than that for bulk and QW SOAs, so QD SOAs are capable of amplifying ultrafast data signals with little or no pattern effects. The physical mechanism leading to short τ_c is that when the QD electrons are depleted by an incoming lightwave, they can be refilled very quickly by the electron population in the material surrounding the dots. Because of these advantages and also enhanced nonlinear effects, QD SOAs are very useful for efficient high-speed all-optical signal processing applications. QD SOA LEFs can be much less than for bulk and MQW SOAs and so are particularly suited for amplifying very high baud rate advanced modulation format data signals.

QDash SOAs are of interest as an alternative to QD SOAs, since they possess some dot-like properties, especially wide gain bandwidth and ultrafast gain recovery time (similar to QD SOAs), and can more easily be made to operate in the 1.55 μm region.

SOA Models

Mathematical models of SOAs are useful in understanding static, dynamic and nonlinear behavior of SOA; in particular the dependency on device materials and structure (Connelly, 2002; Piprek, 2017). In this section simple steady-state, time domain and pulse amplification models are used to describe fundamental SOA physics and characteristics.

Steady-State Model

To determine the factors that influence gain, a simple traveling-wave based model, omitting ASE, can be used (Connelly, 2002). The active material gain g_m per unit length at the signal photon energy E_s is taken to be a linear function $g_m = a(n - n_t)$ of the carrier density n at distance z (from the SOA input) and time t, where the differential gain coefficient $a = \partial g_m/\partial n$ at E_s and n_t is the transparency carrier density. n is determined from the rate equation

$$\frac{dn}{dt} = \frac{\eta I}{eV} - \frac{n}{\tau_c} - \frac{\Gamma g_m P}{AE_s} \tag{8}$$

where I is the bias current, the current injection efficiency η is the fraction of the injected current entering the active region. The active region cross-section area and volume are $A = dW$ and $V = AL$ respectively, where L, d, W and are the SOA length and the active waveguide thickness and width respectively. τ_c models interband effects such as radiative and non-radiative spontaneous recombination processes. In practice τ_c reduces as n increases. The signal power P in the SOA is determined from the traveling-wave equation

$$\frac{dP}{dz} = (\Gamma g_m - \alpha)P \tag{9}$$

where α is the loss coefficient. In the steady-state $dn/dt = 0$, so from Eq. (8),

$$n = \frac{P_{sat}}{P + P_{sat}}\left(\frac{\tau_c \eta I}{eV} + n_t \frac{P}{P_{sat}}\right) \tag{10}$$

The saturation power P_{sat} is defined as

$$P_{sat} = \frac{AE_s}{\Gamma a \tau_c} \tag{11}$$

Defining the normalized power $p(z) = P(z)/P_{sat}$, unsaturated carrier density $n_0 = \tau_c \eta I/(eV)$ and unsaturated gain coefficient $g_0 = \Gamma a(n_0 - n_t)$, Eq. (10) can be written as

$$n(z) = \frac{(n_0 + n_t\, p)}{1 + p} \tag{12}$$

Inserting Eq. (12) into Eq. (9) gives

$$\frac{dp}{dz} = \left(\frac{g_0}{1 + p} - \alpha\right)p \tag{13}$$

The unsaturated gain $G_0 = \exp[(g_0 - \alpha)L]$, which corresponds to a particular value of bias current. Eq. (13) can be solved numerically, using, for example, the Runge-Kutta method. The boundary condition is $p(0) = P_{in}/P_{sat}$, where P_{in} is the input signal power. The amplifier gain G is the ratio of the output power $P_{out} = p(L)P_{sat}$ to the input power. The calculated gain versus output power is shown in **Fig. 5(a)** for various values of G_0 with $\exp(\alpha L) = 10$. The carrier density is calculated using Eq. (12). As the input signal power is increased the gain reduces as the electrically pumped active material conduction band electrons are depleted by stimulated recombination with valence band holes. The normalized saturation output power $p_{o,sat}$ is defined as the normalized output power at which the amplifier gain is half the unsaturated gain. The saturation output power $P_{o,sat} = p_{o,sat}P_{sat}$. From **Fig. 5(a)** it can be seen that $p_{o,sat} \approx -3.8$ dB for 20 dB unsaturated gain and is almost independent of G_0.

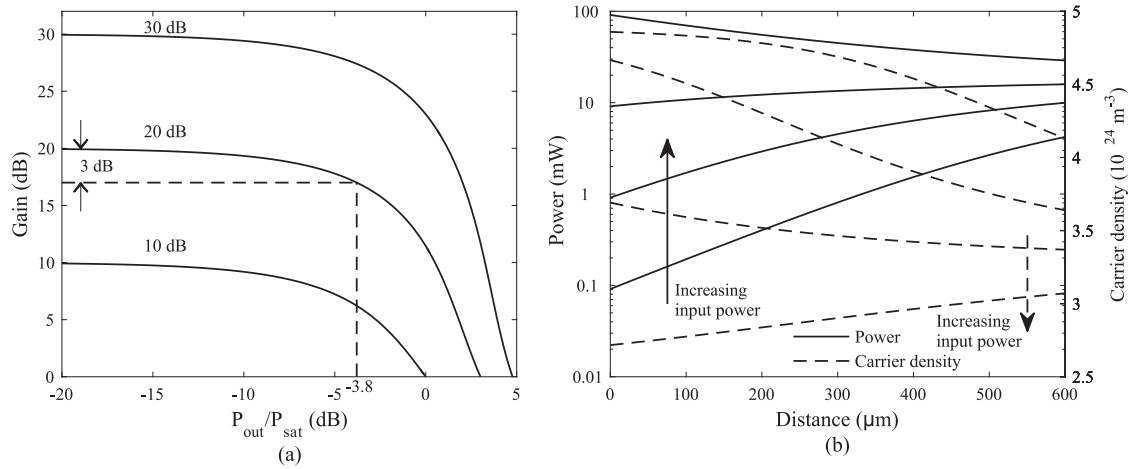


Fig. 5 (a) Gain vs. output normalized power. The parameter is the unsaturated gain. (b) Carrier density and signal distributions for normalized input powers of 0.01, 0.1, 1 and 10. The unsaturated gain is 20 dB.

Consider a 600 μm long square cross-section bulk SOA with w and d equal to 0.4 μm, $\Gamma = 0.45$, $\tau_c = 0.5$ ns, $a = 1 \times 10^{-20}$ m^2, $n_t = 2.5 \times 10^{24}$ m^{-3} and $\exp(\alpha L) = 10$. If the unsaturated gain is 20 dB at a current of 150 mA and $\eta = 1$, then $n_o = 4.9 \times 10^{24}$ m^{-3}, $g_0 = 1.1 \times 10^4$ m^{-1} and if the signal wavelength is 1550 nm, $P_{sat} = 9.6$ dBm and $P_{o,sat} = 5.8$ dBm. **Fig. 5(b)** shows the distributions of $P(z)$ and $n(z)$ for various values of $p(0)$. When $P_{in} \ll P_{sat}$, n is uniform throughout the amplifier. Models that include ASE show that when the amplifier is unsaturated n has a non-uniform symmetrical distribution with minima at the SOA ends and a maximum at the center. As the saturation level increases the n distribution becomes less symmetric; at high levels of saturation n approaches n_t. A further disadvantage of the model is that the spontaneous recombination term in Eq. (8) is taken to be a linear function of n. In practice spontaneous recombination is more accurately modelled as a polynomial function of carrier density so τ_c is actually n dependent. In power booster applications the most important parameter is $P_{o,sat}$. The model shows that $P_{o,sat}$ can be increased by reducing a, τ_c or Γ. One method for increasing $P_{o,sat}$ is to use a wide dilute waveguide, which has a low Γ; however disadvantages include increased device length and current. Compared to bulk materials QD materials have significantly higher values of differential gain, allowing even higher values of $P_{o,sat}$ to be achieved. Modeling of semiconductor nanostructure based SOAs is significantly more complex than for bulk and QW SOAs.

Time-Domain Model and Pattern Effects

SOAs are usually used for modulated signal amplification, so it is of interest to model time domain behavior. Eqs. (8) and (9) can be used to model SOA dynamics having timescales as short as 100 s of picoseconds (Connelly, 2002). First the carrier density is initialized to some suitable value. The carrier density $n(z,t+\Delta t)$ at the next time step is determined from the finite time difference solution of Eq. (8). Assuming that the propagation time of a lightwave through the SOA, which is typically of the order of $1-10$ ps, is much less than the time variation of the optical input, $P(z,t+\Delta t)$ can be determined from Eq. (9) as

$$P(z, t+\Delta t) = P_{in}(t+\Delta t)\exp\left(\int_0^z \Gamma g_m(z, t+\Delta t)dz - \alpha z\right) \tag{14}$$

The calculated $P(z,t+\Delta t)$ is then used in Eq. (8) to determine the carrier density at the next time step until the time span of interest is completed. **Fig. 6** shows the simulated normalized output power, spatially averaged carrier density and instantaneous gain for an amplified 5 Gb/s Non-Return-to-Zero (NRZ) data stream for an SOA having the same parameters as in the steady-state model above. When the input data is high (logical '1'), the carrier density is depleted and so the gain is reduced. If the gain response time (related to τ_c and the saturation level) is comparable to the inverse of the bit rate then signal distortion will arise within a bit period. If a long string of '1's occurs then the gain will saturate at a constant value. If a '1' is followed by a '0' the gain begins to recover; however the gain experienced by the next '1' may not have recovered to its unsaturated value. Hence the signal distortion is bit pattern dependent. Such pattern effects lead to intersymbol interference and can have a severe impact on system performance.

Because bulk and QD SOAs have relatively large τ_c PDPD will also be severe, which severely impacts on coherent system performance. Because PDPD is a nonlinear effect, compensation is very difficult to achieve using current digital signal processing

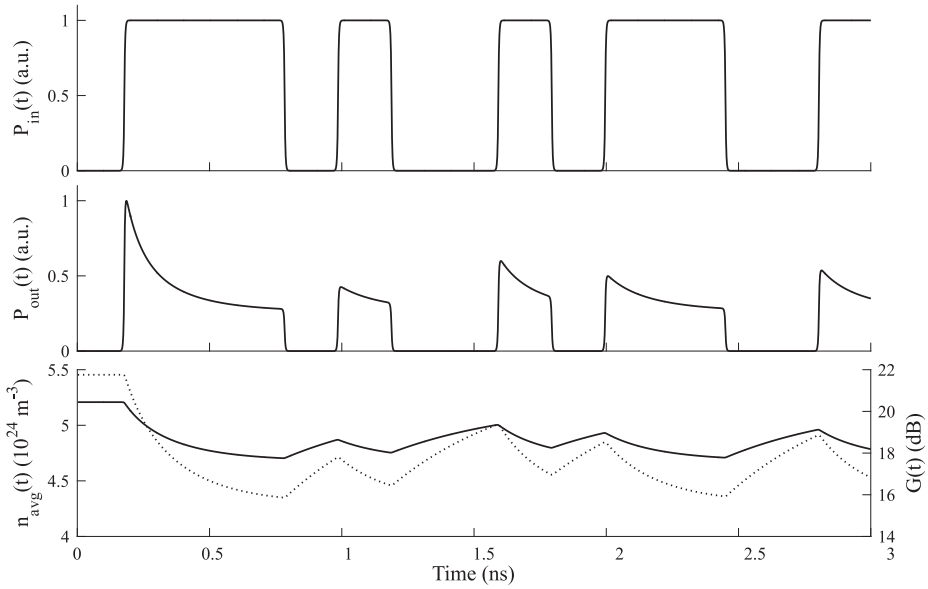


Fig. 6 Input bit stream and simulated SOA output power, spatially averaged carrier density and instantaneous gain for an input NRZ 5 Gb/s data sequence. The input data has high and low level powers of 0.2 mW and 0.2 μW respectively and a rise/fall time of 50 ps. The bias current, unsaturated gain and signal wavelength are 160 mA, 22 dB and 1550 nm respectively.

technology used in digital coherent receivers. QD SOAs have τ_c values significantly smaller than for bulk and QW SOAs so gain recovery is faster, which greatly reduces pattern effects. Pattern effect free amplification with QD SOAs has been demonstrated in Intensity Modulation (IM) systems at bit rates exceeding 160 Gb/s. PDPD is also much lower in QD SOAs, which allows phase transparent amplification of high-order modulation formats at extremely high data rates (Lange *et al.*, 2013).

Two time constants are necessary to describe QDash SOA dynamics: a slow component ($\sim$100 ps) and a fast component ($\sim$1–2 ps). The former is related to carrier transport between individual dashes; the latter to carrier dynamics within a single QDash. At high bit rates (>10 Gb/s), the carrier transport effect cannot follow signal intensity changes, so the gain response dynamics are similar to QD SOAs.

Pulse Amplification Model

Because SOAs have wide optical bandwidths and fast gain recovery lifetimes, they can be used to amplify optical pulses with Full-Width at Half Maximum (FWHM) pulsewidths in the picosecond range. By exploiting ultrafast SOA nonlinearities in materials such as QDs, it is possible to amplify femtosecond pulses (Sugawara *et al.*, 2004). We use a simple model (Agrawal and Olsson, 1989) applicable to pulsewidths much larger than the interband relaxation time (typically < 0.1 ps in bulk and QW SOAs). Intraband processes, which are not considered in this model, such as carrier heating, spectral hole burning, two-photon absorption and interband refractive index dynamics can be exploited to implement ultrafast optical signal processing functions. If it is assumed that $\alpha \ll g$, where $g = \Gamma g_m$, the output pulse instantaneous power $P_{out}(\tau)$ and phase $\phi_{out}(\tau)$ are given by

$$P_{out}(\tau) = G(\tau) P_{in}(\tau) \tag{15}$$

$$\phi_{out}(\tau) = \phi_{in}(\tau) - \frac{\alpha_H}{2} h(\tau) \tag{16}$$

where τ is the local time with respect to a time frame moving with the pulse, $P_{in}(\tau)$ and $\phi_{in}(\tau)$ the input pulse power and phase respectively. α_H is the LEF, which in general is carrier density and wavelength dependent. The instantaneous SOA gain $G(\tau) = \exp[h(\tau)]$. If the input FWHM pulsewidth $\tau_p \ll \tau_c$, then $h(\tau)$ is given by

$$h(\tau) = -\ln\left\{1 - \left(1 - \frac{1}{G_0}\right)\exp\left[-\frac{U_{in}(\tau)}{E_{sat}}\right]\right\} \tag{17}$$

where G_0 is the unsaturated gain, the saturation energy $E_{sat} = E_s W d/(\Gamma a)$ and

$$U_{in}(\tau) = \int_{-\infty}^{\tau} P_{in}(\tau')\, d\tau' \tag{18}$$

The temporal dependence of $G(\tau)$ induced by the amplified pulse power leads to dynamic changes in the pulse phase, i.e., SPM. The frequency chirp $\Delta\nu_{out} = (1/2\pi) d\phi_{out}/dt$ of the output pulse is given by

$$\Delta\nu_{out}(\tau) = \Delta\nu_{in}(\tau) + \frac{\alpha(G_0 - 1)}{4\pi\, G_0} \frac{P_{out}(\tau)}{E_{sat}} \exp\left[\frac{-U_{in}(\tau)}{E_{sat}}\right] \tag{19}$$

where $\Delta\nu_{in}$ is the input pulse chirp. The above closed form equations are applicable to pulsewidths of the order of picoseconds to 10s of picoseconds. To illustrate the effects of an SOA on an amplified pulse consider an unchirped Gaussian input pulse of zero phase and power $P_{in}(\tau) = E_{in}/(\tau_0\sqrt{\pi})\exp[-(\tau/\tau_0)^2]$, where E_{in} is the pulse energy. τ_o is related to τ_p by $\tau_p \approx 1.665\ \tau_0$. In this case $U_{in}(\tau) = (E_{in}/2)[1 + erf(\tau/\tau_0)]$, where *erf* is the error function. Simulated amplified pulse power, chirp and power spectrum are shown in Fig. 7. The amplified pulse is asymmetric because the leading edge of the pulse experiences a larger gain than the trailing edge. The amplified pulse is chirped, resulting in a broadening of the pulse spectrum. At high gains the amplified pulse spectrum exhibits a multipeak structure caused by SPM induced frequency chirp. The complex pulse structure and spectral broadening can lead to a significant increase in chromatic dispersion when the pulse is transmitted in an optical fiber and a consequent degradation in system performance. It is possible under appropriate operating conditions for an SOA to act as a pulse narrower or, by inducing an additive chirp whose sign is the opposite of the input pulse chirp, as a dispersion compensator.

Basic Applications

The three basic applications of SOAs are: booster amplifier, in-line amplifier and preamplifier. Systems using relatively inexpensive low-power lasers often need the signal power to be increased prior to transmission. Boosting laser power in an optical transmitter can also be used to overcome modulator losses, compensate for splitting and tap losses in optical distribution networks and to increase the power budget of optical links. The most critical requirement of a booster amplifier is a high $P_{o,sat}$. The *NF* and polarization sensitivity are of less importance but it is advantageous to keep them as low as possible. Post amplification optical filtering is also not critical because operating the amplifier in saturation greatly reduces the ASE.

Conventional optical transmission systems employ IM of a carrier lightwave. At the receiver end the signal is reconverted to the electrical domain by a photodetector, which is a phase and polarization insensitive process referred to as Direct Detection (DD). If an optical preamplifier is not used, the dominant receiver noise is thermal noise originating from the receiver load resistance and electronic amplifiers. If an optical preamplifier, followed by a narrowband optical filter, is placed in front of the photodetector the

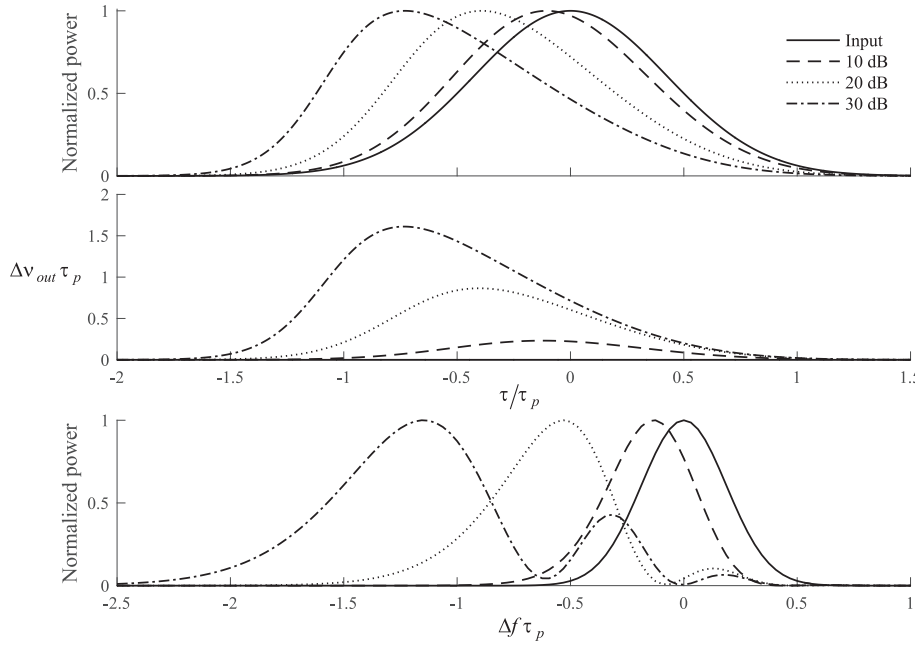


Fig. 7 SOA input and amplified unchirped Gaussian pulse normalized power, chirp and normalized power spectrum. $E_{in}/E_{sat}=0.1$. The parameter is the unsaturated gain and $\alpha_H=5$. Δf is the frequency deviation from the input lightwave center frequency.

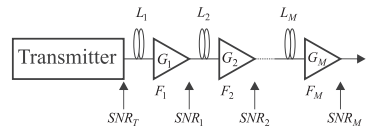


Fig. 8 Optical amplifier chain. The last amplifier in the chain, if immediately followed by a receiver, acts as an optical preamplifier.

dominant noise process is signal-spontaneous beat noise. The power required at the detector is larger than that required if no preamplifier were used; however the power required at the SOA input is lower so the overall receiver sensitivity is improved. It is important the sufficient gain is provided by the SOA such that signal-spontaneous noise dominates; low *NF* and polarization sensitivity are also important. In practical DD receivers typical sensitivities for a 10^{-9} Bit-Error-Rate (BER) are of the order of a few thousand photons per bit. The use of an SOA preamplifier can result in sensitivity improvements typically in the range of 10–20 dB.

In long-haul optical transmission systems, in-line SOAs can be used to compensate for link losses as shown in **Fig. 8**, where each amplifier compensates for the loss of the preceding fiber link. When several SOAs are cascaded the accumulated ASE will reduce the SNR. If the ASE is high the amplifiers will become saturated, which limits the signal amplification. To avoid excessive noise accumulation it may be necessary to follow each SOA by a narrowband optical filter; however this can limit the use of WDM to increase system capacity. The total noise factor F_{total} of a chain of M amplifiers, defined as the ratio of the transmitter SNR SNR_T, to the output SNR SNR_M of the last amplifier in the chain, is given by

$$F_{total} = \frac{F_1}{L_1} + \frac{F_2}{L_1 G_1 L_2} + \cdots \frac{F_M}{L_M \prod_{k=1}^{M-1} L_k G_k} \tag{20}$$

where L_k is the loss of the k-th fiber span and F_k and G_k are the noise factor and gain of the k-th in-line amplifier respectively. If the amplifiers are identical ($F_k=F, G_k=G$) and exactly compensate for their preceding fiber spans ($L_i=G^{-1}$) then

$$F_{total} = MGF \tag{21}$$

If SNR_T is known then the SNR at the receiver input can be determined using Eqs. (20) or (21). In addition to high gain, $P_{o,sat}$ and *NF* are important in-line amplifier parameters.

Signal distortion can occur even when an SOA is slightly saturated and can significantly increase in severity as the number of amplifiers in the chain increases. A gain-clamped SOA can be used to greatly reduce signal distortion and Cross-Gain Modulation (XGM) induced interchannel crosstalk that occurs when IM channels dynamically saturate the amplifier gain, and is of importance in WDM systems. SOA gain-clamping can be achieved by producing lasing action, at a wavelength remote from the signal wavelength, by introducing wavelength dependent feedback such as placing integrated distributed Bragg reflectors at opposite ends

of the SOA. In the lasing state the SOA carrier density is clamped at a fixed value. Changes in the input signal power lead to opposing changes in the lasing mode power, the effect of which is to keep the carrier density and resulting gain fixed. Gain-clamping can also be achieved by the use of an externally injected Continuous Wave (CW) lightwave: using this technique almost pattern free transmission of a 112 Gb/s four level pulse amplitude modulation data over a 40 km fiber link has been demonstrated (Chan and Way, 2015).

Large increases in system capacity and performance can be achieved by employing multi-level modulation formats such as PSK and QAM. The reception of multi-level optical signals requires coherent detection, whereby the signal is combined with a Local Oscillator (LO) laser and simultaneously detected; the resulting electrical signals are then processed to recover the received lightwave phase and amplitude. Coherent systems are now common in backbone networks. A major advantage of coherent detection is that the LO power can be increased to a level such that receiver thermal noise is negligible, in which case receiver sensitivity is determined by the SNR γ_s of the received symbols. In contrast to DD receivers, coherent receivers are linear so post-detection signal processing can be used to compensate for local and transmission link impairments. In the quantum noise limited regime, encountered in links not employing optical amplifiers, $\gamma_s = \eta n_S$ where η is the quantum efficiency of the receiver photodetectors and n_s is the average number of photons per received symbol. When in-line optical amplifiers are used to periodically compensate for the transmission fiber loss and where the last amplifier acts as a preamplifier, as shown in Fig. 8, the signal-spontaneous beat noise dominates in which case $\gamma_s = 2n_s/MF$. If NF is equal to the minimum possible value of 3 dB then $\gamma_s = n_s/M$.

Functional Applications

SOAs can be used to implement many optical signal processing functions especially important at ultra-fast data rates where real-time electronic processing is not possible. In this section the important SOA ultra-fast signal applications of wavelength conversion, optical logic and regeneration are described but also the low speed applications of optical remodulation and microwave phase shifting, all of which exploit various SOA nonlinearities.

SOA FWM Wavelength Converter

Wavelength conversion whereby data is transferred from one carrier wavelength to another is an important function required in WDM networks. Non-degenerate Four-Wave Mixing (FWM) in SOAs is one of the most promising wavelength conversion techniques as it is has the advantages of ultrafast operation and preserves the input data signal amplitude and phase modulation and as such can be used in IM and coherent communication systems. FWM is a coherent third-order nonlinear effect that can occur between two or more co-propagating co-polarized optical fields (Agrawal, 1988). If there are two fields, a high power pump at frequency v_o and a lower power probe at frequency $v_0 - \Delta v$, where Δv is the pump-probe detuning, a refractive index modulation at Δv will result that leads to the creation of a conjugate field at frequency $v_0 + \Delta v$. In the case of a CW pump and modulated signal the conjugate is identical to the signal, albeit with a different power level, as shown in Fig. 9 and can be selected using a bandpass filter at the SOA output. FWM in SOAs arises from different physical phenomena. At low $|\Delta v| (< 1/\tau_c)$ the refractive index modulation is caused by modulation of the carrier density and is only significant for $|\Delta v|$ less than a few 10s GHz. At higher frequencies, two other effects dominate: Spectral Hole Burning (SHB) and Carrier Heating (CH). SHB is caused by the pump creating a hole in the intraband carrier distribution, which modulates the occupation probability of carriers within a band leading to fast gain and refractive index modulation. CH is caused by stimulated emission and free carrier absorption. Stimulated emission subtracts carriers that are cooler than average while free carrier absorption moves carriers to higher energy levels in the band. The resulting increase in temperature decreases the gain. Both SHB and CH have characteristic times in the hundreds of femtosecond range, so FWM based wavelength converters can operate with $|\Delta v|$ in the 100s GHz range and at ultrahigh bit rates (100s Gb/s). FWM also allows the simultaneous conversion of multiple input wavelengths to another set of multiple output wavelengths which is useful in multicasting applications. The conversion efficiency η, defined as the ratio of the SOA output converted signal power to

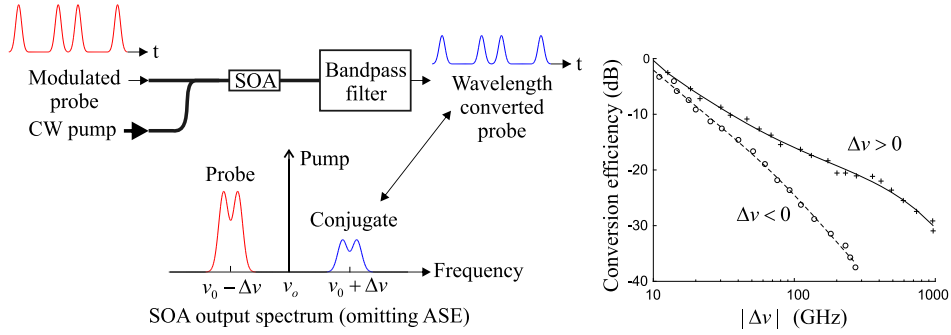


Fig. 9 SOA FWM wavelength converter and typical conversion efficiency vs. detuning frequency characteristic.

the input signal power, depends on a number of factors including the SOA gain, input pump and probe powers, Δv and the time constants associated with the physical processes described above. A typical experimental plot of η versus Δv is shown in Fig. 9.

Optical Logic

Optical logic gates have important applications such as addressing, header recognition, data encoding and encryption. Many different schemes for implementing different logic operations have been investigated, including several based on the exploitation of SOA nonlinearities such as XGM and FWM (Stubkjaer, 2000). Interferometric configurations containing SOAs are useful for implementing fast optical switching and logic functions. The most commonly used structure is the SOA Mach-Zehnder Interferometer (MZI), which for stable operation must be integrated. An SOA MZI XOR gate and associated truth table are shown in Fig. 10. The XOR gate is of special interest since it is the basic building block for a wide range of more complex logic functions. The data signals (A and B) input into the two SOAs modulates the SOA carrier densities leading to refractive index modulations and thereby XGM and Cross-Phase Modulation (XPM) of the co-propagating input CW light, having a different wavelength to the data signals. The input data powers and SOA operating current are chosen such that the CW light experiences the same phase shift in each SOA when A and B have the same power level (identical logic levels) leading to destructive interference (logic '0') at the interferometer output. Conversely when A and B have different logic levels the relative phase shift is π, which results in the CW light experiencing constructive interference (logic '1'). SOA MZIs often have tunable phase shifting elements placed in one or both of the interferometer branches to allow fine tuning and optimization of the switching process. Other logic operations can be implemented by choosing suitable input data power levels and SOA-MZI operating conditions. SOA interferometric logic gates have the advantage of having very steep transfer functions so the speed of operation is not limited by the slow carrier recovery times. Only low input signal power levels are needed to introduce a phase difference between the interferometer arms, so that efficient logic operation conversion is obtained almost independently of wavelength. Also because the carrier density modulation is small, the frequency chirp imposed on the output signal is not large. Bulk or QW SOA MZI logic gates can operate at speeds in the 10s Gb/s range. Much higher speeds (>250 Gb/s) can be achieved using QD SOAs. SOA-MZIs can also be used as wavelength converters but only for IM signals.

All-Optical Data Regeneration

Regeneration of a data signal is needed if the signal quality has been adversely impacted by noise or other impairments and involves one or more of three functions: amplification, reshaping (of the signal amplitude and/or phase) and retiming. Conventional SOA based regenerators for IM signals use bulk or QW SOA MZI switching structures. However, due to its faster gain dynamics a single QD SOA can be used to perform some regeneration functions such as noise compression at ultrafast bit rates (>40 Gb/s) as shown in Fig. 11 (Akiyama *et al.*, 2007). A noisy input NRZ data stream has an average binary '1' power level that induces gain saturation in the SOA. If the gain dynamics are much faster than the signal transitions the SOA effectively responds instantaneously to the input data thereby compressing the '1' level noise. This scheme does not reduce the '0' level noise; however it can be reduced by following the SOA with a saturable absorber.

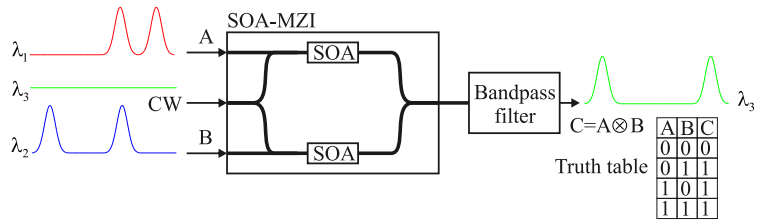


Fig. 10 SOA MZI XOR logic and truth table.

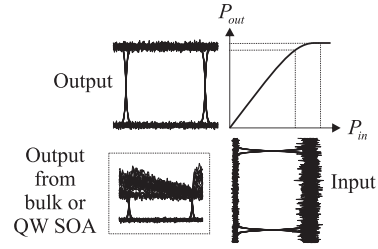


Fig. 11 Eye diagrams illustrating QD SOA noise compression at high bit rates. If a bulk or QW SOA is used, pattern effects will be present due to the much slower gain recovery time.

Regeneration of amplitude/phase modulated signals is significantly more challenging and requires the use of a nonlinear optical loop mirror or Phase Sensitive Amplifier (PSA), which precludes the direct use of phase insensitive SOAs. A PSA is capable of amplifying only one of the two quadrature components of a lightwave signal and attenuating the other. In theory a PSA can realize noise free amplification ($NF=0$); however most implementations are based on complex bulk crystal or fiber systems. Recently a monolithically integrated PSA based on a dual-pump degenerate FWM process that uses an SOA as the nonlinear element has been demonstrated (Li *et al.*, 2016), which shows great promise for use in coherent optical communications.

Reflective SOA based Passive Optical Network

There is an increasing effort to develop cost-effective Fiber-To-The-Home (FITH) networks. WDM Passive Optical Networks (PONs) are one of the most promising solutions for realizing next generation FTTH. For successful deployment of WDM-PONs the Optical Network Unit (ONU) located in the subscriber premises needs to be wavelength independent as well as being cost effective. One such WDM-POM which uses a carrier remodulation scheme based on the saturation properties of a Reflective SOA (RSOA) is shown in **Fig. 12** (Lee *et al.*, 2005). An RSOA is a TW SOA with a highly-reflective coating applied to one of the end facets. RSOAs have similar optical bandwidths to TW-SOAs but have a steeper gain vs. input optical power characteristic and are therefore easier to saturate. In the WDM-PON Central Office (CO), low extinction ratio (ER) NRZ downstream signals are generated by direct modulation of lasers and transmitted through an Arrayed Waveguide Grating (AWG) wavelength multiplexer, the fiber link and at the Remote Node (RN) separated by an AWG wavelength demultiplexer. At the ONU the downstream channel is split by a coupler; one output of which is sent to a receiver for conventional detection and processing, the other output is sent to an RSOA. The power level of the RSOA input is chosen such that the amplifier operates in saturation, thereby suppressing the downstream modulation component. The upstream channel (RSOA output) is generated by modulating the RSOA current with NRZ data. The downstream channel data rate is higher than that of the upstream. At the upstream receiver located in the CO, the unsuppressed component of the downstream signal, which would lead to a power penalty, is filtered out by an electrical lowpass filter. Error-free bidirectional transmission with 1.25 Gb/s downstream data rates was demonstrated for 20 km transmission distance.

SOA Microwave Phase Shifter

Microwave Photonics (MWP) is concerned with the generation, transport and processing of radio frequency, microwave and millimeter wave signals in the optical domain (Xue *et al.*, 2010). Wideband tunable microwave delays and phase shifters are a key element required in MWP, with applications in microwave filtering and optically fed phased array antennas, however their implementation in the microwave domain is very difficult. One of the most promising photonic based techniques uses Coherent Population Oscillations (CPOs) in SOAs, which allows tunability based on the SOA current or input optical power. CPOs are induced by injecting a sinusoidally modulated lightwave at frequency f_m as shown in **Fig. 13**. Therefore the input optical signal is a

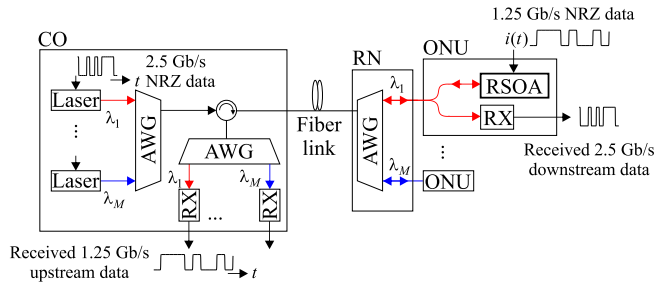


Fig. 12 WDM-PON topology. RX, receiver. Modified from Lee, W., Park, M.Y., Cho, S.H., *et al.*, 2005. Bidirectional WDM-PON based on gain-saturated reflective semiconductor optical amplifiers. IEEE Photonics Technology Letters 17, 2460–2462.

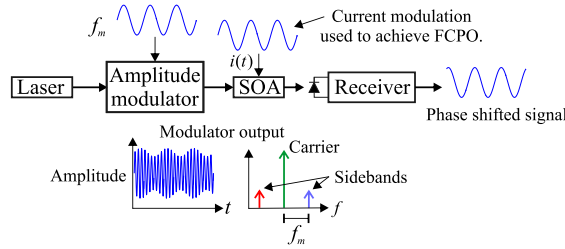


Fig. 13 SOA MWP phase shifter.

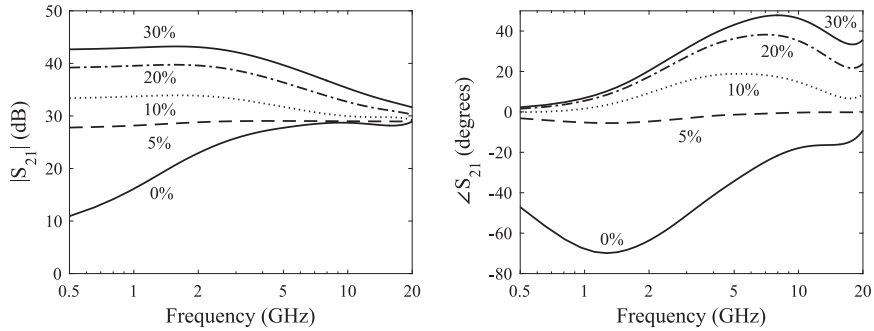


Fig. 14 Typical bulk SOA phase shifter transfer functions for different values of the current modulation index. The unsaturated gain at the lightwave wavelength of 1550 nm is 25 dB at the bias current of 175 mA. The average input optical power is −5 dBm, at which the SOA gain is 14 dB, with an optical modulation index of 10%.

double-sideband unsuppressed carrier amplitude modulated signal having a strong carrier and two relatively weak sidebands. This leads to the carrier density being modulated at the beat (modulation) frequency between the sidebands and carrier, which induces a frequency dependent change in the group index and therefore the group velocities of the sidebands. After propagation through the SOA the envelope of the optical signal will be delayed or advanced (slow or fast light). Detection of the output optical signal results in an output microwave signal that is phase shifted with respect to the modulator input. At modulation frequencies less than the inverse of the carrier lifetime, the beat-signal gain is low leading to a poor signal-to-noise ratio. The low frequency response can be enhanced by using Forced CPO (FCPO), which involves simultaneously modulating the input optical power and SOA current as shown in **Fig. 13**. The SOA phase shifter transfer function $S_{21}(f)$ is defined as the ratio of the SOA optical output to input detected beat signal currents. $|S_{21}(f)|$ and $\angle S_{21}(f)$ are the phase shifter signal gain and phase shift respectively; typical frequency plots of which for CPO (no current modulation) and FCPO operation are shown in **Fig. 14** for various values of the current modulation index. Adjustment of the current modulation index allows control of the level of advancement and to switch from fast to slow light. Further tunability can be obtained by adjusting the bias current and optical power.

Conclusions

SOAs have many applications in optical networks especially as optical signal processing devices. Although SOAs have been investigated for many years and are commercially available, usually as single devices fabricated from bulk or QW material, the development of new SOA structures, advanced materials such as QDs, PICs and exploitation of ultrafast non-linearities make SOA research and development an exciting field of study. The commercial deployment of ultrafast coherent optical communication systems requires new optical signal processing functions such as full regeneration of amplitude/phase modulated signals, and doubtless new SOA based technologies will be developed to meet this need.

See also: Erbium Doped Fiber Amplifiers for Lightwave Systems

References

Agrawal, A.P., 1988. Population pulsations and nondegenerate four-wave mixing in semiconductor lasers and amplifiers. Journal of the Optical Society of America B 5, 147–159.

Agrawal, G.P., Olsson, N.A., 1989. Self phase modulation and spectral broadening of optical pulses in semiconductor laser amplifiers. IEEE Journal of Quantum Electronics 25, 2297–2306.

Akiyama, T., Sugawara, M., Arakawa, Y., 2007. Quantum-dot semiconductor optical amplifiers. Proceedings of the IEEE 95, 1757–1766.

Chan, T.K., Way, W.I., 2015. 112 Gb/s PAM4 transmission over 40 km SSMF using 1.3 μm gain-clamped semiconductor optical amplifier. In: 2015 Optical Fiber Communications Conference and Exhibition (OFC), Los Angeles, CA, pp. 1–3.

Chuang, S.L., 2009. Physics of Optoelectronic Devices. New York, NY: Wiley.

Connelly, M.J., 2002. Semiconductor Optical Amplifiers. Boston, MA: Kluwer Academic.

Dutta, N.K., Wang, Q., 2013. Semiconductor Optical Amplifiers. Singapore: World scientific.

Lange, S., Contestabile, G., Yoshida, Y., Kitayama, K.I., 2013. Phase-transparent amplification of 16 QAM signals in a QD-SOA. IEEE Photonics Technology Letters 25, 2486–2489.

Lee, W., Park, M.Y., Cho, S.H., *et al.*, 2005. Bidirectional WDM-PON based on gain-saturated reflective semiconductor optical amplifiers. IEEE Photonics Technology Letters 17, 2460–2462.

Lelarge, F., Dagens, B., Renaudier, J., *et al.*, 2007. Recent advances on InAs/InP quantum dash based semiconductor lasers and optical amplifiers operating at 1.55 μm. IEEE Journal of Selected Topics in Quantum Electronics 13, 111–124.

Li, W., Lu, M., Mecozzi, A., *et al*., 2016. First monolithically integrated dual-pumped phase-sensitive amplifier chip based on a saturated semiconductor optical amplifier. IEEE Journal of Quantum Electronics 52, 1–12.
Olsson, N.A., 1989. Lightwave systems with optical amplifiers. Journal of Lightwave Technol 7, 1071–1082.
Piprek, J., 2017. Handbook of Optoelectronic Device Modeling and Simulation: Fundamentals, Materials, Nanostructures, LEDs, and Amplifiers. vol. 1. CRC Press.
Stubkjaer, K.E., 2000. Semiconductor optical amplifier-based all-optical gates for high-speed optical processing. IEEE Journal of Selected topics in Quantum Electronics 6, 1428–1435.
Sugawara, M., Ebe, H., Hatori, N., *et al*., 2004. Theory of optical signal amplification and processing by quantum-dot semiconductor optical amplifiers. Physical Review B 69, 235332.
Xue, W., Sales, S., Capmany, J., Mørk, J., 2010. Wideband 360° microwave photonic phase shifter based on slow light in semiconductor optical amplifiers. Optics Express 18, 6156–6163.

Wavelength Division Multiplexing

Klaus Grobe, ADVA Optical Networking SE, Martinsried, Germany

Introduction to WDM

Wavelength Division Multiplexing (WDM) is a multiplexing and transmission scheme in fiber-optical telecommunications where different wavelengths, emitted by several lasers, each carry dedicated information. These wavelengths are multiplexed by means of WDM filters. Likewise, they are separated, or demultiplexed, in the receive end by means of similar filters, or coherent detection using tunable local oscillators.

As a multiple-access mechanism, WDM allows separating different customers' traffic in the wavelength domain. This is called Wavelength Division Multiple Access (WDMA). WDM can be combined with any other multiplexing or multiple-access schemes, namely electrical Time Division Multiplexing (TDM, TDMA), Sub-Carrier Multiplexing (SCM, SCMA), and Orthogonal Frequency Division Multiplexing (OFDM, OFDMA, including the real-valued variant Discrete Multi-Tone).

WDM splits into high-performance Dense WDM (DWDM) with high channel count (>100) and channel spacing down to 12.5 GHz, and lower-cost Coarse WDM (CWDM) with up to 18 channels spaced at 20 nm. CWDM is used in access and backhaul, DWDM is used in all high-capacity and long-haul transport applications. Driven by improvements of components and modulation and equalization techniques, the total DWDM capacity has largely increased since the first experiments in the late 1980s, see **Fig. 1**.

DWDM systems have approached an area of slowed-down capacity improvement. Over the next years, DWDM on Standard Single-Mode Fibers (SSMF) will finally reach the Nonlinear Shannon Limit (Essiambre *et al.*, 2010). Further progress beyond this limit will possibly require new fiber types.

Overviews on WDM can be found, e.g., in Grobe and Eiselt (2014), Agrawal (1992).

Transmission Effects in Optical Fibers

WDM transmission depends on the fiber type that is used. SSMF show frequency dependence, time variance, and nonlinear behavior. The resulting transmission impairments are.

- Linear effects
  a. Attenuation
  b. Chromatic Dispersion (CD)
  c. Polarization-Mode Dispersion (PMD)
  d. Polarization-Dependent Loss (PDL)
- Nonlinear effects
  a. Kerr effects
  b. Scattering effects (nonlinear, stimulated)

Detailed discussions of these effects can be found in the literature, e.g., Agrawal (1992), (1995). For WDM long-haul transmission, all these effects and their interactions have to be considered.

Linear Effects

Attenuation

Attenuation in Silica single-mode fibers is mainly caused by intrinsic Silica-glass loss (Rayleigh scattering, Infrared absorption), extrinsic loss due to impurities and bending loss.

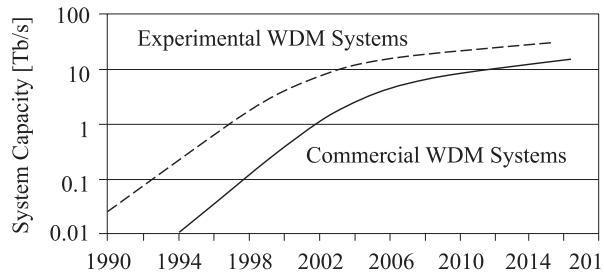


Fig. 1 Development of WDM Systems Transport Capacity over time.

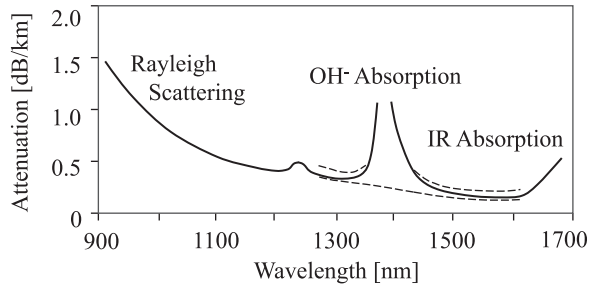


Fig. 2 Spectral loss in single-mode fibers.

The spectral loss caused by **intrinsic effects** and **impurity** is shown in Fig. 2. The attenuation peak around 1400 nm is caused by OH^- ions. In fibers vintages up into the 1980s, this peak was very pronounced, in new fibers it has almost been eliminated. The dashed lines in Fig. 2 indicate the loss tolerance as specified in ITU-T Recommendation G.695 (ITU-T, 2015) for SSMF.

Fibers exhibit loss caused by perturbations of the ideal waveguide geometry. Such perturbations can result from fiber bends, where bending radii with $R \gg \lambda$ (macro-bending) and $R \approx \lambda$ (micro-bending) with λ the optical wavelength have to be differentiated.

Macro-bending loss occurs when fiber-optic cables are bent, e.g., in patch frames etc. It increases exponentially with increasing wavelength and decreasing bending radii. For wavelengths <1300 nm and bending radii >20 mm, macro-bending loss can be neglected. WDM wavelengths around 1550 nm can already be affected significantly. The region >1600 nm is sensitive to macro bending.

Micro-bending loss is caused when the fiber is subject to radial pressure. This happens when a fiber (within a cable) is subject to mechanical pressure. This forces deformations of the fiber's core-cladding boundary which have similar dimension as the wavelengths. This causes destructive light interference and consequently loss.

Chromatic dispersion

Chromatic Dispersion (CD) causes signal distortions via the dependence of the velocity of propagation on the frequency of the respective spectral components. Since modulation always leads to spectral broadening of carrier waves, every information transmission in fibers is subject to CD. The resulting temporal spread of the signal leads to Inter-Symbol Interference (ISI).

In the absence of nonlinearity and birefringence, and with constant attenuation α, propagation of a light wave with envelope E is given by:

$$E(z,t) = E_0 \exp(-\alpha z) \exp(j\omega t - j\beta(\omega) z) \tag{1}$$

The phase constant $\beta(\omega)$ is usually developed into a Taylor series around a mean carrier frequency ω_0 (corresponding to, e.g., 1550 nm). From the third Taylor coefficient, β_2, the so-called Dispersion Parameter is derived (Agrawal, 1992, 1995):

$$D = -\omega^2 \beta_2 / 2\pi c_0 \tag{2}$$

D has the dimension (ps/(nm · km)). It describes the temporal spread (in ps) of signal pulses with a certain optical bandwidth (in nm) over a certain transmission distance (in km). The other coefficients are related to *phase velocity* ($v_p = \omega_0/\beta_0 = c_0/n$), group velocity ($v_G = 1/\beta_1$), and dispersion slope (β_3). More coefficients are typically not considered.

Fig. 3 shows D parameters for single-mode fibers according to ITU-T Recommendations G.652 (ITU-T, 2016d), G.653 (ITU-T, 2010b) and certain G.655 (ITU-T, 2009b) fiber brands (OFS, 2017; Corning, 2014; Ohsono *et al.*, 2003; Prysmian Group, 2010; Corning, 1998).

The D parameter can be used to define a transfer function which considers CD:

$$H_{CD}(z, j\omega) = \exp\left[j\left(\beta_0 + \beta_1(\omega - \omega_0) - \frac{1}{2} \frac{\lambda_0^2}{2\pi c_0} D(\omega - \omega_0)^2 \right) z \right] \tag{3}$$

The transfer function $H_{CD}(z, j\omega)$, complemented by a simple attenuation term, can be used for calculations or simulations of a linear fiber model (neglecting nonlinearity).

Polarization mode dispersion

Polarization Mode Dispersion (PMD) is caused by differences of the effective refractive index of the transmission fiber between (any) two orthogonal polarizations. This leads to birefringence and is caused by fiber geometry perturbation or lateral stress on the fiber. These cannot be perfectly avoided. While in a perfectly circular fiber no particular pair of orthogonal polarization modes is distinguished, birefringence leads to two particular modes developing, the *Principal States of Polarization* (PSP). They depend on wavelength. The PSPs are defined at the input to the fiber as those states of polarization, for which, when slightly varying the signal frequency, the output polarization remains constant. The time delay between both PSPs is called *Differential Group Delay* (DGD).

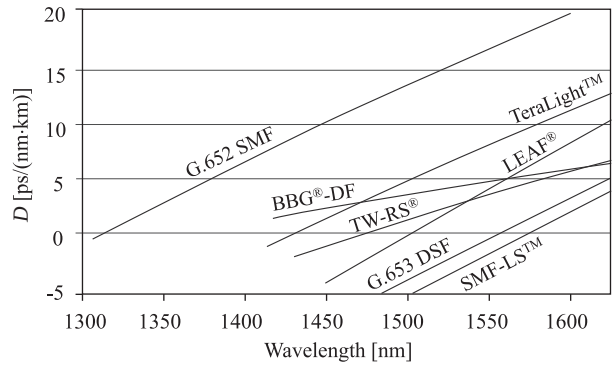


Fig. 3 D parameters of various single-mode fibers.

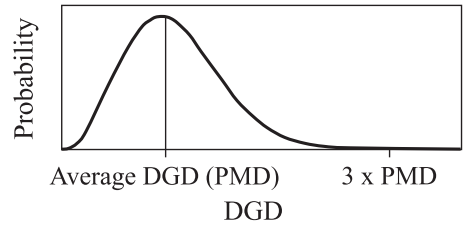


Fig. 4 Distribution function for DGD.

PMD results in pulse spread ΔT^{PMD}, depending on fiber length L and PMD Parameter D_{PMD}:

$$\Delta T^{\text{PMD}} = D_{\text{PMD}}\sqrt{L} \tag{4}$$

Long fibers can be modeled as cascades of k birefringent sections with random polarization orientations each. The resulting total PMD then yields:

$$\Delta T_{\text{tot}}^{\text{PMD}} = \sqrt{\sum_k \left(\Delta T_k^{\text{PMD}}\right)^2} \tag{5}$$

The PMD parameter D_{PMD} has the dimension (ps/$\sqrt{\text{km}}$). Typical values are listed in **Table 3**.

The PMD parameter is an average over time or wavelength. The actual pulse broadening is related to the instantaneous DGD and depends on the actual State of Polarization of the signal relative to the PSPs. The pulse broadening is therefore described by a stochastic process. The DGD has a Maxwellian distribution, as shown in **Fig. 4**.

The *average* DGD is referred to as PMD. As most transmission systems are impacted by the instantaneous DGD, a DGD tolerance is specified. Since DGD $> 3 \cdot \text{PMD}$ only occurs with a probability of $4 \cdot 10^{-5}$ ($\sim$21 min/year), the specified PMD tolerance is usually three times smaller than the DGD tolerance to guarantee low outage probability (Bülow, 1998).

Nonlinear Fiber Effects

Silica fibers exhibit weak nonlinearity. This effect becomes apparent at high power levels and long-distance transport without regeneration. The latter is achieved in all WDM long-haul transport, which is ultimately limited by this nonlinearity (Essiambre *et al.*, 2010).

Nonlinear fiber effects split into two classes, instantaneous Kerr effects and (stimulated, i.e., laser-like) scattering effects with specific gain spectra.

The **Kerr effects** can be classified into inter- (WDM-) channel and intra-channel effects, and into signal-signal or signal-noise interactions, see **Fig. 5** (Winzer and Essiambre, 2006).

The excitation of the individual Kerr effects, i.e., their relative relevance, depends on power level, fiber type (especially CD) and per-channel bit rate, as shown in **Fig. 6** (Winzer and Essiambre, 2006). For low CD, FWM (i.e., generation of new spectral components via parametric mixing) is dominant, and for very high Baud rates, the intra-channel effects dominate.

The propagation constant β is proportional to the effective refractive index n_{eff} seen by the electrical field, $\beta = 2\pi \cdot n_{\text{eff}}/\lambda$. The effective refractive index becomes dependent on the instantaneous power, causing nonlinearity:

$$n_{\text{eff}} = n_0(\lambda) + n_2 P/A_{\text{eff}} \tag{6}$$

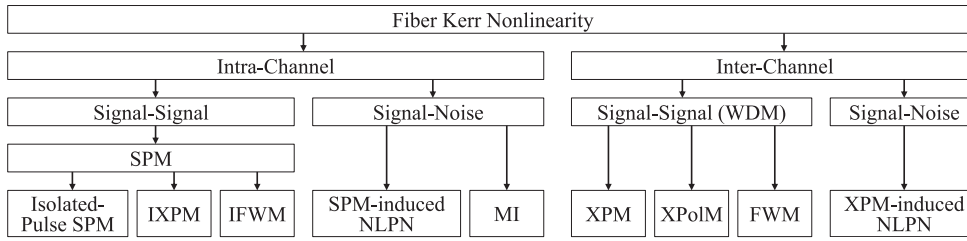


Fig. 5 Kerr-effect classification. MI, Modulation Instability; NLPN, Nonlinear Phase Noise; SPM, Self-Phase Modulation. XPolM: Cross-Polarization Modulation. (I)XPM: (Intra-channel) Cross-Phase Modulation. (I)FWM: (Intra-channel) Four-Wave Mixing.

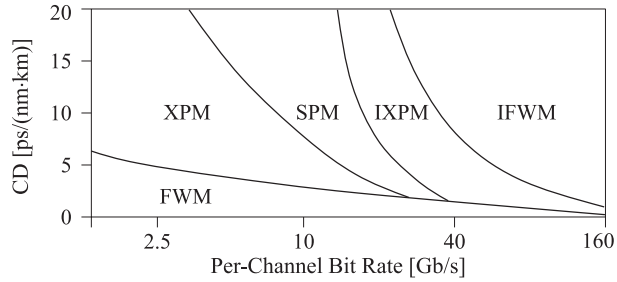


Fig. 6 Dominance of nonlinear effects.

Table 1 Values of nonlinearity for various types of transmission fiber

Fiber type	A_{eff} *typ.*	γ *typ. at 1550 nm*
SSMF G.652	80 μm^2	1.3/(W · km)
TrueWave® G.655 (OFS, 2017), DSF G.653	55 μm^2	1.8/(W · km)
LEAF® G.655 (Corning, 2014)	72 μm^2	1.4/(W · km)

Here, n_2 is the nonlinear fiber coefficient, A_{eff} is the effective area of the fiber, which is the average area occupied by the field in the fiber, and P is the instantaneous signal power. A nonlinearity coefficient γ is often used to describe the nonlinear properties of the fiber. It is given by $\gamma = 2\pi n_2/\lambda A_{\text{eff}}$. Values for different fiber types are listed in **Table 1**.

The Kerr effects can be calculated numerically using a nonlinear propagation equation, the so-called *Nonlinear Schrödinger Equation* (NLSE) (Agrawal, 1995):

$$\frac{\partial A}{\partial z} = -\frac{\alpha}{2}A - \frac{j}{2}\beta_2\frac{\partial^2 A}{\partial T^2} + j\gamma|A|^2 A \tag{7}$$

A is the complex envelope of the propagating field. The NLSE (7) does not consider scattering effects, and is limited in accuracy for ultra-broadband signals. A more general description has to be based on Wiener-Volterra series (Schetzen, 1980). It is possible to extend Eq. (7) by higher-order CD and a Raman shock term (Agrawal, 1995), resulting in the *Generalized Nonlinear Schrödinger Equation* (here, without attenuation term):

$$\frac{\partial A}{\partial z} = \sum_{i=1}^{3}\beta_i\frac{(-j)^{i+1}}{i!}\frac{\partial^i A}{\partial t^i} - \frac{j\beta_0 n_2}{n_0}\left(1 + \frac{j}{\omega_0}\frac{\partial}{\partial t}\right)\left(A\int_0^\infty R(\tau)|A(z,t-\tau)|^2 d\tau\right) \tag{8}$$

Eq. (7) can efficiently be solved numerically with the Split-Step Fourier Algorithm (Agrawal, 1995). The GNLSE (8) has to be solved with more CPU-demanding pseudo-spectral methods.

Raman Scattering is scattering of photons at glass molecules with certain nuclear and vibrational states. These states become dependent on the incoming intensity, causing nonlinearity. The molecular states cover a certain spectrum. Due to the amorphous structure of Silica glass, *harmonic broadening* occurs, and the scattered photons are down-shifted in frequency by continuous scattering spectra. During scattering, the system molecule-plus-photon is in a virtual, i.e., forbidden, energy state. Hence, scattering takes place quickly, within less than 1 ps (Agrawal, 1995).

For strong Stimulated Raman Scattering (SRS), amplification of signals of incident power $P_S(0)$ over length L is given by:

$$P_S(L) = P_S(0) \times \exp(G_R(\Delta f) I_0 L) \tag{9}$$

I_0 is the intensity of the optical pump, and $G_R(\Delta f)$ is the Raman gain as shown in **Fig. 7** (Agrawal, 1995; Chraplyvy, 1990).

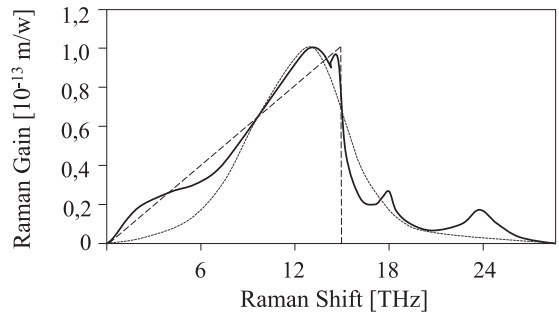


Fig. 7 Raman gain spectrum. Triangle (dashed) and single-Lorentzian (dotted) approximations were sometimes used for numerical approximations.

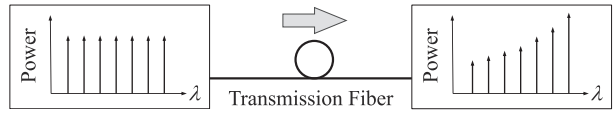


Fig. 8 Gain tilt in WDM channels due to SRS.

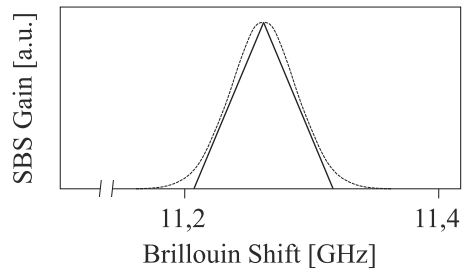


Fig. 9 Brillouin gain of a G.653 fiber with almost triangular shape according to Agrawal (1995), and approximation with a single Lorentzian (dotted).

In WDM systems, higher-frequency channels pump lower-frequency channels through SRS. This leads to power loss of higher-frequency channels and gain tilt, as shown in **Fig. 8**. In addition, SRS leads to crosstalk in the lower-frequency channels following the (intensity) bit patterns in the pump channels.

If the Raman-induced loss for the pump channels has to be <1 dB, the product of total bandwidth and total power of all channels has to be <50 GHz $\cdot$ W (Chraplyvy, 1990).

Brillouin Scattering is inelastic scattering caused by spatial refractive-index fluctuations which are generated by density fluctuations. The latter are caused by hypersonic sound waves or by electrical fields (electrostriction). Similar to SRS, the effect can be stimulated, leading to Stimulated Brillouin Scattering (SBS).

SBS has the lowest power threshold of all nonlinear fiber effects and can limit WDM systems in their maximum per-channel power. The maximum SBS gain is $G_B(\Delta f_B) \approx 4 \cdot 10^{-11}$ m/W and is achieved at a frequency shift of $\Delta f_B = \sim 11$ GHz. The SRS gain bandwidth for optical pumps around 1550 nm is limited to 20–100 MHz.

Similar to SRS, amplification of incident power $P_S(0)$ by SBS is given by

$$P_S(L) = P_S(0) \times exp(G_B(Df) I_0 L) \tag{10}$$

I_0 is the intensity of the optical pump, L is the fiber length, and $G_B(\Delta f)$ is the frequency-dependent SBS gain. An SBS gain example is shown in **Fig. 9**.

SRS causes *backscattering* which can be blocked by optical isolators. However, due to the low SBS threshold, per-channel launch power must be <10 dBm in order to avoid significant reflection and power loss (Chraplyvy, 1990). Due to the narrow-band gain characteristics, SBS is also reduced by widening the laser line widths of WDM channels. This can be achieved through additional slow amplitude modulation (so-called dither (Fishman and Nagel, 1993)).

Components and Sub-Systems

End-to-end WDM transmission consists of transmitters, transmission channel, and receivers. The WDM transmission channel always comprises SMFs. WDM transmitters and receivers consist of several functions, see **Fig. 10**.

Necessary functions in the transmit/receive end include channel coding/decoding (e.g., the application of Forward Error Correction), modulation/demodulation, and WDM and optional polarization multiplexing/demultiplexing. Optional functions in include amplification, encryption/decryption, and time-domain multiplexing/demultiplexing.

WDM Transmitters

Laser diodes

WDM systems use (tunable) Laser Diodes (LD) as optical transmitters. LD output spectra depend on the optical-cavity geometry and the gain-medium wavelength dependence. Semiconductor materials for wavelengths of 0.5 μm to 4 μm are shown in **Fig. 11** (Agrawal and Dutta, 1986).

LD direct current modulation causes intensity and parasitic frequency modulation. The latter is called (frequency) chirp. To avoid chirp, *external modulators* have to be used.

Tunable lasers are required for network-capacity optimization and remote reconfiguration, together with Reconfigurable Optical Add/Drop Multiplexers (ROADM). They also allow protection schemes based on wavelength switching. Tunable lasers further reduce the number of different transmitter variants, thus reducing device and inventory cost.

Mode number m and wavelength λ of a LD depend on the effective refractive index n and effective cavity length L, $m\lambda/2 = nL$. Consequently, LD tunability can be achieved by tuning any of these parameters

$$\frac{\Delta\lambda}{\lambda} = \frac{\Delta n}{n} + \frac{\Delta L}{L} - \frac{\Delta m}{m} \tag{11}$$

On the first two terms on the right-hand side, electronic, thermal or mechanical tuning can be applied (Coldren *et al.*, 2004). Electronic tuning refers to the application of an electric field which tunes the laser frequency. Several implementations exist, which are basically derivatives of the three-section Distributed-Bragg-Reflector (DBR) LD.

Thermal tuning can be done with certain Distributed-Feedback (DFB) LD. Mechanical tuning can be applied to Micro-Electromechanical Vertical Cavity Surface Emitting Lasers (MEM-VCSEL) or External Cavity Lasers. Electronic and mechanical implementations enable fast tuning speeds, thermal tuning is somewhat slower.

External modulators

External modulators are used for high-speed WDM transmission. They allow broad modulation bandwidth, (partly) avoid the chirp resulting from direct laser modulation, and can enable complex (I + Q) modulation.

Electro-Absorption Modulators (EAM) are similar to reverse-biased p-i-n diodes with a bulk active region or multiple quantum-wells (MQWs) as absorption layer. They can have low power consumption and low cost. However, they are mainly restricted to intensity modulation.

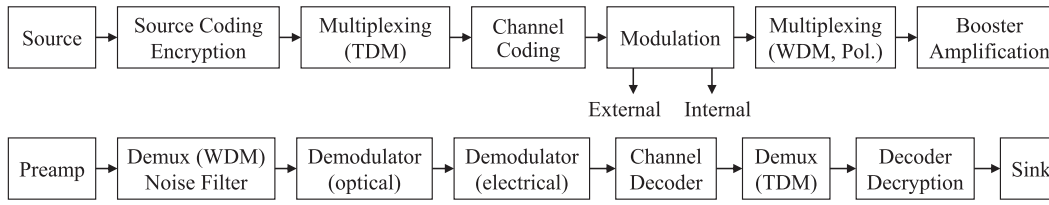


Fig. 10 Transmit-end (top) and receive-end (bottom) of WDM transmission system.

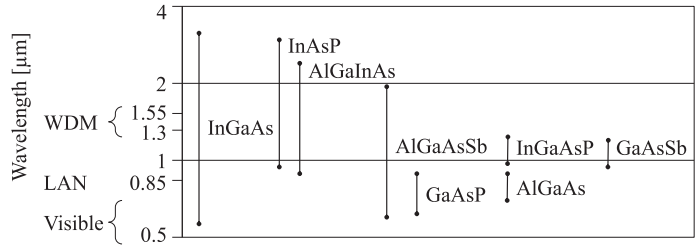


Fig. 11 Semiconductor materials for WDM laser diodes.

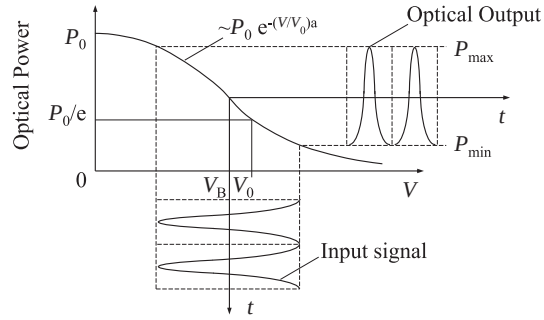


Fig. 12 Modulation of an EAM.

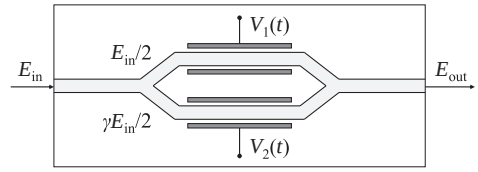


Fig. 13 MZM structure.

The output intensity of an EAM is a nonlinear function of the modulating voltage, see **Fig. 12**. For intensity modulation at high extinction ratio (ER) and low insertion loss, small bias voltage has to be used. For high bias voltage, ER decreases and insertion loss increases.

EAMs can be monolithically integrated with laser diodes, e.g., DFB lasers. This avoids adiabatic laser chirp. Since the same semiconductor system is used, transient chirp remains. EAMs are mainly used for 10-Gb/s intensity-modulated WDM transmission. Reach >100 km on SSMF without dispersion compensation can be achieved.

In **electro-optic modulators**, the distribution of electrons within the medium is distorted if an electrical field is applied so that the refractive index changes anisotropically. This leads to phase changes of traversing optical signals. With the change of refractive index Δn, the resulting phase change of a signal propagating through an electro-optic medium is given by $\Delta(\varphi) = \Delta\beta_0 L = k_0 \Delta n L$. A relevant material for electro-optic modulators is $LiNbO_3$ (Lithium Niobate).

Electro-optic modulators are used as phase modulators or as nearly chirp-free amplitude modulators. For amplitude modulation, a Mach-Zehnder Interferometer (MZI) structure according to **Fig. 13** is used. This structure is also called Mach-Zehnder Modulator (MZM).

The transfer function of the MZM is given as a function of the modulator voltages V_1 and V_2:

$$H(V_1, V_2) = E_0 \cos\left[\pi(V_1 - V_2)/2V_\pi\right] \cdot \exp\left[j\pi(V_1 + V_2)/2V_\pi\right] \tag{12}$$

The (nonlinear!) cosine term accounts for amplitude modulation, and the exponential term for phase modulation or chirp. MZM can achieve bandwidths of >100 GHz and high extinction ratio of 15–20 dB.

Single-Mode Fibers

Fibers have major impact on long-haul WDM transmission. Neglecting specialty and multi-mode fibers (the latter are relevant for LAN and inside data centers), relevant fibers are:

- Standard Single-Mode Fibers (ITU-T Recommendation G.652A-D (ITU-T, 2016d))
- Dispersion-Shifted single-mode Fibers (DSF, ITU-T G.653 (ITU-T, 2010b))
- Cut-off-Shifted single-mode Fibers (CSF, ITU-T G.654. A-C (ITU-T, 2016b))
- Non-Zero Dispersion-Shifted single-mode Fibers (NZ-DSF, ITU-T G.655A-E (ITU-T, 2009b))
- Ultra-broadband, non-zero dispersion-shifted single-mode fibers (ITU-T G.656 (ITU-T, 2010a))
- Low bending-loss single-mode fibers (ITU-T G.657A/B (ITU-T, 2016a))

SSMF are the oldest single-mode fibers. The first low-loss SSMF were built in 1978. Over time, several Recommendations (G.652A to G.652D) were produced which mainly improved water-absorption loss around 1400 nm and the PMD parameter D_{PMD}.

DSF combine lowest loss and lowest CD around 1550 nm. They were optimized of single-channel – *non-WDM* – transmission. Their close-to-zero CD is critical for WDM long-haul transmission since it allows phase matching between WDM channels which can lead to disastrous FWM. DSF have never been deployed on broad scale.

CSF have their cut-off wavelength shifted into the region of 1530 nm, and ultra-low loss. For wavelengths higher than cut-off, a fiber becomes single-mode. Cut-off shifting can be achieved by increasing the fiber core area, which also decreases nonlinearity. Fibers according to G.654B and G.654C have low PMD, and the G.654B type has CD *increased* as compared to SSMF. This makes G.654B CSF suitable for ultra-long-haul applications.

NZ-DSF are an improvement of DSF with regard to higher CD around 1550 nm. They aimed at a CD compromise between CD-limited reach and CD-supported nonlinearity suppression. Over time, CD was increased and PMD decreased. NZ-DSF also include large-area fibers.

G.656 fibers are the extension of G.655 fibers towards even higher CD. Recommendation G.656 further defines higher CD over increased bandwidth, low PMD and very low loss.

G.657 low bending-loss fibers were developed to decrease macro-bending loss resulting from cabling. G.657A fibers are splice-compatible with SSMF and have somewhat decreased macro bending loss. G.657B fibers are incompatible with SSMF and have even better bending loss.

Relevant dispersion parameters (D, dispersion slope S, zero-dispersion wavelength λ_0, PMD parameter D_{PMD}) are listed for relevant fibers in **Table 2**.

Some G.655 fibers have smaller effective core areas, compared to G.652 fibers, see **Table 1**. This results from refractive-index profile manipulations which often leads to multiple claddings, and also decreased core diameter, as shown in **Fig. 14**. Other G.655 fibers have increased effective areas (Large-Area (LA) or Large-Effective-Area fibers (LEAF) (Corning, 2014)).

The technically achievable maximum transport capacity of SMF is approaching (Essiambre *et al.*, 2010). *New fiber types* which possibly lead to WDM capacity increase are Few-Mode Fibers (FMF), Multi-Core Fibers (MCF), and Photonic Crystal Fibers (PCF), respectively.

In **Few-Mode Fibers**, several guided modes are modulated and demodulated independently. More guided modes are allowed by slightly increasing the core area, thus shifting cut-off such that the modes become allowed. FMFs are easily splicable, similar to SMF or CSF. All modes in an FMF can be amplified simultaneously by a single EDFA, given they are amplified equally. Due to increased core area, FMFs also have reduced nonlinear effects.

One challenge of FMF is the need for high-performance Multiple-Inputs Multiple Outputs (MIMO) receivers. These are required since mode de-/multiplexers, fiber distortion and other lumped devices will cause unavoidable mode coupling. For an FMF supporting the LP_{01} and the two degenerated LP_{11} modes, a 6×6 MIMO is required (each mode has two polarization modes). Further challenges relate to low-loss mode multiplexers and demultiplexers. The early devices used so far did not fully solve the low-loss requirement.

In **Multi-Core Fibers**, multiple single-mode cores are integrated in one fiber. Ideally, these cores are amplified by single fiber amplifiers (which can also be based on multiple cores), however challenges may occur with proper splicing.

Different MCF have been proposed and tested, originally based on three, seven, and 19 cores. Such MCF can have strong crosstalk between the cores. Then, MIMO receivers are required, similar to FMF. Crosstalk can also be suppressed by proper fiber design, e.g., by separating the cores with additional holey barriers. A MIMO may nonetheless be required if crosstalk is caused in fiber bends and lumped components.

Table 2 Dispersion characteristics of relevant single-mode fibers

	G.652A/C	*G.652B/D*	*G.654B*	*G.654C*	*G.655A*
D @ 1550 nm (ps/(nm · km))	16–21	16–21	22	20	0.1–6.0
λ_0 [nm] (Zero-Dispersion)	1311.5 ± 10	1311.5 ± 10	(~1300)	(~1300)	1480 ± 30
S @ 1550 nm (ps/(nm^2 · km))	≤0.092	≤0.092	0.07	0.07	≤0.05
D_{PMD} (ps/$\sqrt{\text{km}}$)	0.5	0.2	0.2	0.2	0.5
	G.655B	G.655C	G.655E	G.656	G.657A
D @ 1550 nm (ps/(nm · km))	1.0–6.0	1.0–6.0	5.5–10	2.0–14	16–21
λ_0 (nm) (Zero-Dispersion)	1480 ± 30	1480 ± 30	1440	1480 ± 30	1312 ± 12
S @ 1550 nm (ps/(nm^2 · km))	≤0.05	≤0.05	≤0.052	≤0.05	≤0.092
D_{PMD} (ps/$\sqrt{\text{km}}$)	0.5	0.2	0.2	0.2	0.2

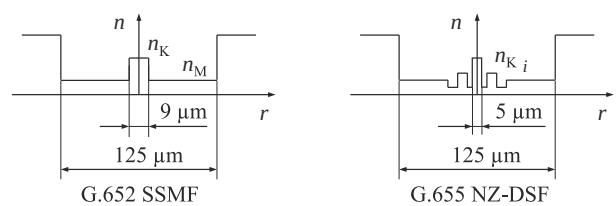


Fig. 14 Refractive index profile $n(r)$ for G.652 SSMF (left) and typical G.655 NZ-DSF (right).

In addition, new devices for separating the cores are required. Such devices can be based on extensions of fiber tapers. Further problems include the necessary splicing to amplifiers.

Photonic Crystal Fibers (PCF) have micro-structured cross sections, often comprising axial holes running along the fiber. They can either increase the fiber's confinement, or allow air guidance. This allows improving bending loss, PMD or nonlinear effects (which can be either decreased or increased, according to the needs). PCF split into two categories: index-guided and Photonic Bandgap- (PBG-) guided fibers (Russell, 2006). Index-guided PCF typically have a solid core and surrounding holes, the aim of PBG-guided fibers is hollow-core air guidance.

Solid-core, index-guided fibers with holey cladding have lower *effective* refractive index in the cladding and higher contrast than Silica fibers. They produce stronger confinement which allows stronger nonlinearities, birefringence, or bending sensitivity. The best attenuation reported so far is 0.28 dB/m at 1550 nm. Today, it is not clear if these fibers will play a more important role in WDM transmissions. One relevant application may be the crosstalk suppression provided by holey barriers in MCF.

PBG-guided fibers aim at air-guidance, allowing very low effective index (i.e., increased propagation velocity), and almost zero nonlinearity and PMD. Therefore, it is likely that they will play an increasingly important role in WDM transmissions. PBG-guided fibers split into Bragg fibers and hollow-core PCFs, see **Fig. 15**. Bragg fibers utilize the fact that 1D crystals can reflect light from all angles and polarizations. They form waveguides similar to hollow metal waveguides. They hence support the TE_{01} as the lowest-loss mode, which is immune to PMD and does not split into polarization modes.

Hollow-core PCFs consist of 2D-periodic photonic crystals and a hollow core. Here, the PBG is distorted, allowing guided fields to be confined into the hollow core region. This allows low nonlinearity, PMD and effective refractive index. The latter makes PCF suitable for "fast" fibers where the group velocity exceeds the one known from Silica SSMF. Best results reported so far are 1.2 dB/km at 1620 nm over limited fiber length (Roberts *et al.*, 2005).

Optical Amplifiers

In optical amplifiers, the WDM Optical Multiplex Section is amplified in the analog optical domain. Optical amplifiers are transparent with regard to payload protocols, Baud rates, and modulation schemes of the individual WDM channels. They do not provide equalization, pulse re-shaping or re-timing. Instead, they *add noise*. This noise is called Amplified Spontaneous Emission (ASE).

The optical amplifiers most relevant to WDM transmission are Erbium-Doped Fiber Amplifier (EDFA), Fiber Raman Amplifiers (FRA), and Semiconductor Optical Amplifiers (SOA). Parametric optical amplifiers (using the fiber's Kerr nonlinearity) so far have not proven relevance, and Brillouin fiber amplifiers cannot be used for WDM signals due to lack of bandwidth (see **Fig. 9**). WDM amplifier technologies are shown in **Fig. 16** (Hirano, 2002).

EDFAs are the most relevant amplifiers for WDM transmission (Desurvire, 1994). Originally developed for DWDM C-band transmission, similar fiber amplifiers exist for other wavelength bands as well. Examples include (co-doped) L-band EDFAs, and Thulium-doped and Praseodymium-doped amplifiers for S-band and O-band amplification, respectively.

Fig. 15 Air-guiding Photonic Bandgap (PBG) fibers.

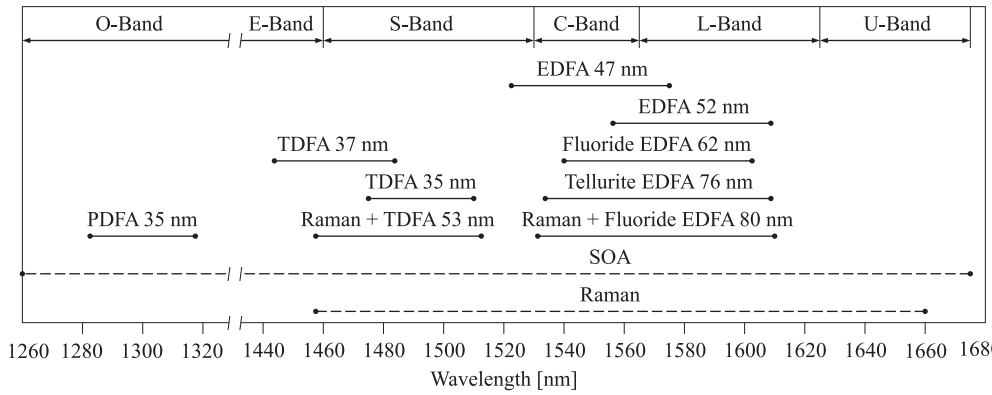


Fig. 16 WDM amplifier technologies.

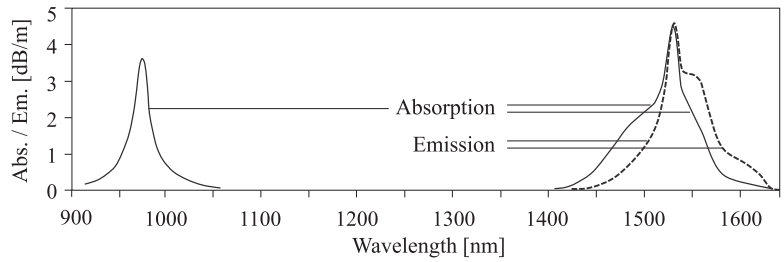


Fig. 17 Absorption and emission bands of an EDFA.

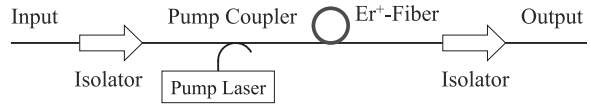


Fig. 18 Principle EDFA configuration with isolators.

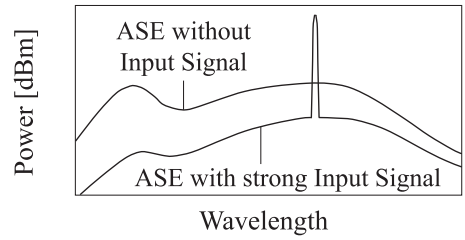


Fig. 19 EDFA ASE spectrum.

Erbium can produce gain at 1520–1610 nm, has absorption bands elsewhere which are accessible with low-cost pump lasers, and achieves high population inversion at reasonable pump-power levels. The absorption and emission bands of an EDFA are shown in Fig. 17. Two absorption bands are shown, one centered around 980 nm, the other overlapping with the emission band. Here, pumping is possible at 1480 nm, which is no relevant WDM payload wavelength.

Since an EDFA produces ASE, optical isolators are necessary to prevent back-propagation of noise. The resulting simplified block diagram of an EDFA is shown in Fig. 18.

Both, the ASE spectrum and the gain characteristics of an EDFA strongly depend on the (WDM) input signals, as seen from Fig. 19.

The gain dependence on input power necessitates power-level control specially for cascaded amplifiers. Power levels can be controlled at the transmit side and in in-line sites which perform equalization, e.g., with variable optical attenuators. This is often implemented in ROADMs.

The gain spectrum of **Fiber Raman Amplifiers** depends on the spectral location of the Raman pumps. Given the availability of cost-effective high-power pumps, Raman amplification is possible throughout the entire low-loss region of the optical fiber (it is difficult at best at strong water absorption around 1400 nm).

The SRS gain spectrum, relative to its pump wavelength, is shown in Fig. 7. With two pumps, gain bandwidths exceeding 90 nm can be achieved, as indicated in the top insert of Fig. 20. FRA use the transmission fiber itself as gain medium. Pumps can propagate co- and/or counter-directionally with the WDM signals to be amplified. Backward pumping is preferred since forward pumping requires higher pump-power levels. It is often used as pre-amplification for EDFAs. Both operation modes are shown in Fig. 20.

Distributed Raman amplification depends on the transmission-fiber type. On G.655 NZ-DSF, C-band amplification becomes less effective because the pumps fall into the zero-CD region and become subject to FWM.

Due to the ultra-short time constant of SRS, sophisticated transient and power-level control is required. Transient control can be based, e.g., on gain clamping.

Semiconductor Optical Amplifiers are similar in structure to index-guided Fabry-Pérot laser diodes (Agrawal, 1992). SOAs have broad gain bandwidth, and with different devices, the complete wavelength region of 1280–1650 nm can be covered. For CWDM (ranging from 1271 to 1611 nm), SOAs are the only cost-efficient amplification technology.

For multi-channel WDM operation, gain-control mechanisms are required. At system level, a saturating reservoir channel or small-signal operation (linear regime) can be implemented. At device level, laser gain clamping can be used. In a gain-clamped

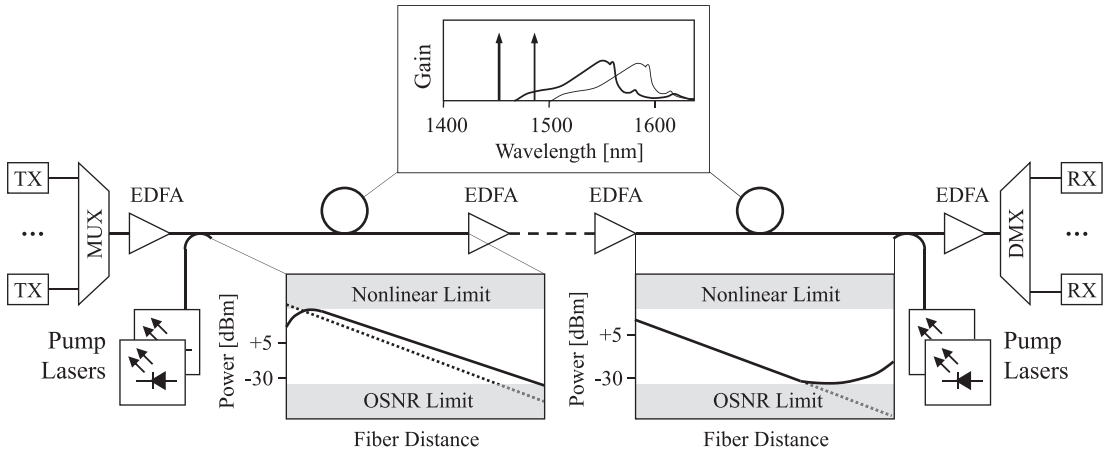


Fig. 20 Raman amplification with co- and counter-propagating pumps.

Table 3 Comparison of relevant SOA and fiber-amplifier parameters

	EDFA	*FRA*	*SOA*
Max. Gain (dB)	>40	>25	>30
Wavelength Region (nm)	1530–1610	1280–1650	1280–1650
3-dB Bandwidth (nm)	30–60	Pump-dependent	60
P_{Sat} (dBm)	22	0.75 × Pump	18
Polarization Dependence (dB)	0	0	<0.5
Noise Figure, typ. (dB)	4–6	3–5	6–8
Pump Power	25 dBm	>30 dBm	<400 mA
Time Constant	10 ms	fs–ps	0.2 ns
Switchable	No	No	Yes

SOA, the gain medium is shared between SOA and a laser. Hence, the device is operated above threshold and produces constant gain. Lasing occurs at wavelengths different from the amplification band.

Table 3 lists relevant SOA parameters in comparison with those of fiber amplifiers.

Chromatic-Dispersion Compensation

Compensation of CD became necessary in the 1990s with WDM per-channel bit rates increasing to 10 Gb/s. Due to its linear, deterministic nature, CD can be compensated quasi-statically in the optical domain. The most important techniques were based on (Agrawal, 1992, 1995):

- Dispersion Compensation Fibers (DCF)
- Dispersion-compensating gratings, e.g., Fiber Bragg Grating (FBG)

In addition, fiber nonlinearity can be used for (partial) CD compensation. This is achieved by mid-span spectral inversion or Self-Phase Modulation (Soliton effect).

CD design must provide correct amounts of net dispersion at each add/drop node in a network. Performance penalties caused by nonlinearity and CD for each *potential* connection must be as low as possible. Optimum CD design depends on per-channel bit rate and modulation, number of WDM channels and fiber type. For 10-Gb/s On-Off Keying (OOK), a dispersion map with Distributed Undercompensation (DUCS) in each span is preferred. In contrast, 40-Gb/s (NRZ-DPSK) signals prefer slight over-compensation in each span, and zero residual dispersion at the receiver. A DUCS dispersion map is shown in Fig. 21.

In addition to quasi-static compensation, various Tunable Optical Dispersion Compensators (TODC) have been developed, especially for the directly-detected 40-Gb/s WDM systems. TODCs became necessary to compensate the remaining receive-end residual CD. They may become relevant for direct-detect 400-Gb/s (based on Pulse-Amplitude Modulation with four levels (PAM-4)) again.

After Y2000, digital signal processing (DSP) became powerful enough to cope with real-time dispersion equalization in high-speed fiber-optic transmission. DSP is mostly applied in coherent receives, but pre-distortion at the transmitter is also possible. The most commonly used real-time equalizers are Feed-Forward Equalizers (FFE).

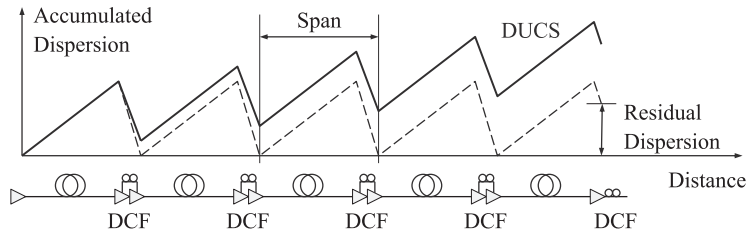


Fig. 21 Compensated multi-span long-haul link with DUCS.

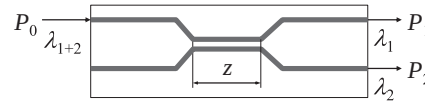


Fig. 22 Wavelength-selective fused fiber coupler.

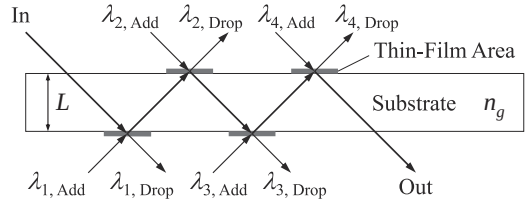


Fig. 23 Thin-film WDM filter.

Passive WDM Filters and Couplers

Passive static optical filters are used as de-/multiplex or add/drop filters and as noise-bandwidth limiters. They are now partially replaced by Re-configurable Optical Add/Drop Multiplexers (ROADMs).

Wavelength-selective directional couplers can be made from two fibers with different core diameters and refractive indices. These fibers are matched at only one wavelength each, the device can hence be used as *diplexer* (Senio and Jamro, 2009). The principle is shown in **Fig. 22**.

For the diplexer, the *isolation loss* is given as the power of the desired wavelength at a given output port divided by the power of the other wavelength at the same port: $\alpha_{ISO} = 10\ \log_{10}(P_{\lambda 1,port1}/P_{\lambda 2,port1})$. In multi-port filters, this is called *adjacent-channel isolation* and is given as ratio of desired wavelength at one port, divided by the power of the *two adjacent* wavelengths at the same port. The *total isolation* in a multi-port device is given by the power of the desired wavelength divided by the accumulated power of *all* other wavelength.

Dielectric **Thin-Films Filters** (TFF) are relevant for WDM de-/multiplexing and add/drop. TFF are based on Fabry-Pérot filters (FPF), the principle is shown in **Fig. 23**.

A FPF is essentially a resonator, bound on either end by a partial mirror. Light of frequencies other than resonant frequencies is mostly reflected internally. Light at resonant frequency enters the cavity and exits through the opposite mirror. In the TFF shown in **Fig. 23**, layers of dielectric thin films are used as mirrors. Thin-film reflection or transmission is wavelength-dependent. Hence, the add/drop locations depend on wavelength, and add/drop wavelengths are assigned individual fibers.

Relevant parameters of an FPF are the finesse and the Free Spectral Range (FSR). The finesse is a measure for the filter quality. Higher reflectivity of the thin-film mirrors leads to higher finesse. Resonant wavelengths are given by $\lambda = 2Ln_g/m$, where n_g is the refractive index of the substrate, and $m = 1, 2, 3\ldots$ The FSR is the distance between peaks in the filter response. It is given by $FSR = c/n_gL$. The finesse is given as the ratio of FSR and Full-Width at Half-Maximum (FWHM), $FSR/FWHM = \pi\sqrt{R}/(1-R)$. R is the reflectance of the mirrors.

Arrayed Waveguide Gratings (AWGs) can multiplex/demultiplex high WDM channel numbers (up to ~ 100) with relatively low loss (Dragone, 2005). **Fig. 24** shows an *N:M* AWG. It consists of two couplers (free-propagation regions, FPR) which are connected by a waveguide array with equal length difference ΔL between adjacent array waveguides.

Light coming from an input waveguide is diffracted and coupled into the arrayed waveguides by the first FPR. The optical path-length difference ΔL between adjacent array waveguides equals an integer multiple of the central wavelength λ_0 of the demultiplexer. As a consequence, the field distribution at the input aperture will be reproduced at the output aperture of the second FPR. Therefore, at the center wavelength, the light focuses in the center of the image plane if the input waveguide is centered in the input

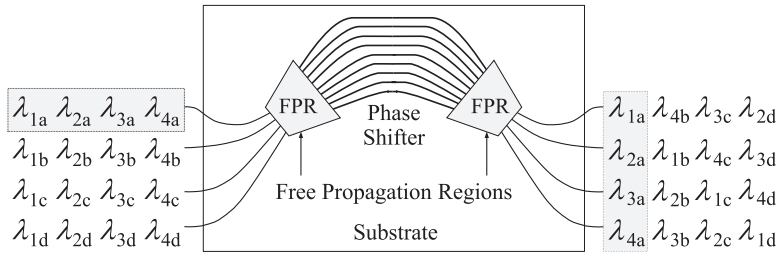


Fig. 24 *N:M* AWG as column-row converter.

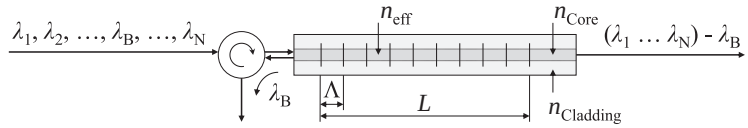


Fig. 25 Fiber Bragg Grating (FBG).

plane. If the input wavelength deviates from the central wavelength, phase changes occur in the array branches which increase linearly from the inner to outer array waveguides. This causes the wavefront to be tilted at the output aperture. Therefore, spatial separation of output wavelengths is given.

The FSR of an AWG is given by $FSR = c/n_2\Delta L$ (with n_2 the refractive index in the waveguides between the FPRs). AWGs can route one wavelength each in several wavelength bands, which are separated by the FSR. The respective devices are also called *cyclic AWGs*. Cyclicity can be used for capacity upgrades of WDM systems or in systems based on single-fiber working, utilizing, e.g., the C-band and the L-band. In ITU Recommendation G.698.3, cyclic AWGs with up to 48 ports (100 GHz in the C-band) are standardized.

As integrated waveguide devices, AWGs show strong temperature dependence. The temperature-induced wavelength shift can be expressed as (Grave de Peralta *et al.*, 2004):

$$\frac{\partial\lambda}{\partial T} = \frac{\lambda}{nL}\left(\frac{\partial(nL)}{\partial T}\right) = \frac{\lambda}{nL}\left(n\frac{\partial L}{\partial T} + L\frac{\partial n}{\partial T}\right) = \lambda\left(\frac{1}{L}\frac{\partial L}{\partial T} + \frac{1}{n}\frac{\partial n}{\partial T}\right) = \lambda\left(\alpha + \frac{1}{n}\frac{\partial n}{\partial T}\right) \tag{13}$$

For silica, $\partial n/\partial T$ is in the range of $7.5 \cdot 10^{-6}/^{\circ}$, and for silicon, $\alpha = 2.63$ ppm/°, respectively. A silica-on-silicon device has a wavelength drift (red shift) of $d\lambda/dT = 12$ 12 pm/°. A 50-GHz DWDM device (~400 pm channel spacing) is completely transparent every 400 pm, and opaque in between. The device becomes a wavelength stop if the temperature changes by ~17°. Hence, AWGs must be temperature-stabilized or athermalized. This can be achieved with mechanical and/or material compensation. Typical total insertion loss of athermalized 50-GHz AWGs is in the range of 6 dB, which is almost independent on the WDM port count.

Fiber Bragg Gratings (FBGs) are periodic perturbations of the refractive index in the propagating medium. An FBG in an add/drop configuration is shown in **Fig. 25**. Λ is the period of the grating (or perturbation). The Bragg phase-matching condition is satisfied for two waves with β_0 and β_1 if $|\beta_0 - \beta_1| = 2\pi/\Lambda$. If a wave with β_0 propagates through the grating, its energy is coupled onto a scattered wave traveling in the counter direction at the same wavelength if $|\beta_0 - (-\beta_0)| = 2\beta_0 = 2\pi/\Lambda$.

Fiber Bragg Gratings can be produced from photosensitive (Ge-doped) single-mode fibers. The grating is written into the fiber with UV lasers, where regions with higher UV intensity produce higher refractive index. The required refractive-index change is in the range of $\Delta n \approx 10^{-4}$.

Advantages of FBGs include low loss (down to 0.1 dB), ease of coupling to transmission fibers, polarization insensitivity, and high crosstalk suppression. The typical temperature coefficient is in the range of $\sim 1.2 \cdot 10^{-2}$ nm/°C (Lima *et al.*, 2005). Therefore, FBGs must be operated in temperature-controlled environment, or be temperature-compensated.

FBG with short period ($\Lambda \approx 0.5$ μm) or long period ($\Lambda \approx 100$–$1000 +$ μm) exist. *Short-period* FBGs are well-suited as filters or channalized CD compensation devices in WDM systems. *Long-period* FBGs can be used, e.g., for broadband EDFA gain equalization.

Interleavers are periodic filters used for combining or separating (de-interleaving) two WDM multiplex sections with the same channel spacing and a grid offset of half the spacing. For example, a 50-GHz DWDM signal can be composed of two 100-GHz signals. Due to the required periodicity, Mach-Zehnder Interferometers are suitable devices for interleavers. Similar to MZIs used as modulators (see **Fig. 13**), MZIs used as interleavers introduce a delay ΔL to one of the branches, as shown in **Fig. 26**.

The signal coming from Input 1 and going to Output 1 through the upper arm acts as reference. Then, the signals coming from Input 1 and going through the lower arm to Output 1 and Output 2 experience phase lags of $\pi/2 + \beta\Delta L + \pi/2 = \pi + \beta\Delta L$ and $\pi/2 + \beta\Delta L - \pi/2 = \beta\Delta L$, respectively. For $\beta\Delta L = k\pi$, $k = (2n+1)$, the signals in the upper output add in phase. At Output 2, the phase difference is $(2n+1)\ \pi$ which is out of phase, hence there is no signal at the output. Opposite behavior is achieved for

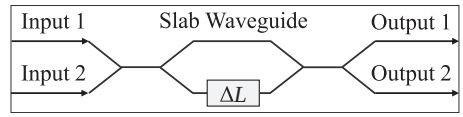


Fig. 26 MZI as (de-) interleaver.

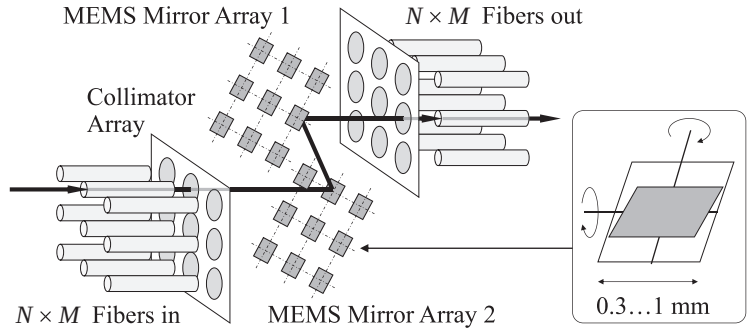


Fig. 27 3D-MEMS used for multi-dimensional fiber cross connect.

$\beta\Delta L = 2n\pi$. The MZI transfer function is given by:

$$\begin{pmatrix} T_{11}(f) \\ T_{12}(f) \end{pmatrix} = \begin{pmatrix} \sin^2(\beta\Delta L/2) \\ \cos^2(\beta\Delta L/2) \end{pmatrix} \tag{14}$$

For use as a de-interleaver, λ_1 and λ_2 (both at Input 1) are chosen to coincide with the minima and maxima of the transfer function. With $\beta = 2\pi n_{\text{eff}}/\lambda$, the delay ΔL and the corresponding phase lag $\beta\Delta L$ can be expressed as $\Delta L = m_i\lambda_i/2n_{eff}$, $\beta\Delta L = m_i\pi$. For $\lambda_1 = 2\Delta L n_{\text{eff}}/m_i$, m_i odd, and Output 1 has a signal whereas Output 2 has no signal. For $\lambda_2 = 2\Delta L n_{\text{eff}}/m_i$, m_i even, the output assignment is vice versa. Hence, the device acts as a de-interleaver. Operation as interleaver is similar.

Tunable WDM Filters

Tunable filters have been investigated since the early WDM days. Advantages of tunable filters include remote reconfigurability, and reduction of system items and spare parts. Tunability can be based on several technologies:

- TFF, thermally tuned.
- Liquid-Crystal filter, electrically tuned.
- MEM-tunable devices.
- FBG, temperature- or strain-tuned.
- Acousto-optic filter, tuned via electrostriction.
- Electro-optical filter, electrically tuned.
- AWG, thermally tuned.
- MZI, cascaded, with heaters.
- Ring-Resonator filters, thermally tuned.

So far, not all of these technologies have been commercialized.

Wavelength Switching Devices

Optical switching matrices in WDM systems are used for cross connects or Reconfigurable Optical Add/Drop Multiplexers (ROADMs). They must have low insertion loss, high isolation, and low back reflection. In addition, they need to be highly reliable, fast (for certain application), compact, low power-consuming, and scalable. Relevant techniques include Micro Electro-Mechanical Systems (MEMS) and Liquid Crystal Technologies (LCT).

MEMS refer to arrays of tiny tilting mirrors sculpted from semiconductor materials such as silicon. In 3D-MEMS arrays, the mirrors can be tilted in any direction. 3D-MEMS arrays can be used for cross connects on wavelength, wavelength-band or fiber level. A 3D-MEMS assembly is shown in Fig. 27.

As MEMS is based on mechanical movement, it offers the best isolation performance available. For the same reason MEMS cannot be used for drop-and-continue applications.

In **LCT**, liquid crystals alter their transmission characteristics when applying an electric field. LCT devices can throttle the amount of light that passes through them. They can hence be used as variable attenuators or power splitters with configurable splitting ratio. This enables drop-and-continue applications. LCT has low power consumption (as known from displays). It can be combined with electro-holographic technology for increasing switching speeds.

Liquid-Crystal-on-Silicon (LCoS) is the latest addition to LCT (Sakurai *et al.*, 2012). It allows WDM-channel switching based on many sub-elements. Variable numbers of these sub-elements can be bonded for varying channel bandwidth without attenuation dips, as shown in **Fig. 28**. LCoS can be used for *flexible-grid* ROADMs (with internal frequency granularity of 6.25 GHz).

ROADMs

Reconfigurable Optical Add/Drop Multiplexers are used in flexible WDM networks. Early technologies led to Degree-2 ROADMs for WDM rings or linear add/drop links. In meshed networks, Multi-Degree (MD-) ROADMs are required. Most devices are based on *Wavelength-Selective Switches* (WSS, based on MEMS or LCT/LCoS). This evolved from Degree-4 ROADMs (4 network ports plus local add/drop port) via Degree-8 to degrees exceeding 30.

The functionality of MD-ROADMs evolved from colored and directed add/drop via directionless and colorless ROADMs to contentionless devices, see **Fig. 29(A–D)**. Meanwhile, these variants can be implemented with *flexible-grid* technology.

In a **colored/directed ROADM**, each terminated network fiber (degree) requires dedicated add/drop filters. A client device hence must be connected to the correct add/drop filters and filter ports. This problem is solved in **directionless** ROADMs. An additional WSS (**Fig. 29(B)**) connects the add/drop ports of the line-terminating WSSs. An add/drop filter is connected to this additional WSS. Drop signals can now be selected from any of the terminated lines, and add signals be directed respectively. Add/drop filter ports are still colored, i.e., specific wavelengths terminate at specific ports. Due to the use of a single filter in the add/drop path, each wavelength can only be terminated once, leading to potential wavelength blocking.

In **colorless, directionless** ROADMs, a further WSS provides uncolored add/drop ports, see **Fig. 29(C)**. Different wavelengths can be terminated at any WSS port. This is one application area for WSS with (add/drop) port counts ≥ 20. Wavelength blocking can still occur. Colorless, directionless ROADMs allow *restoration* switching.

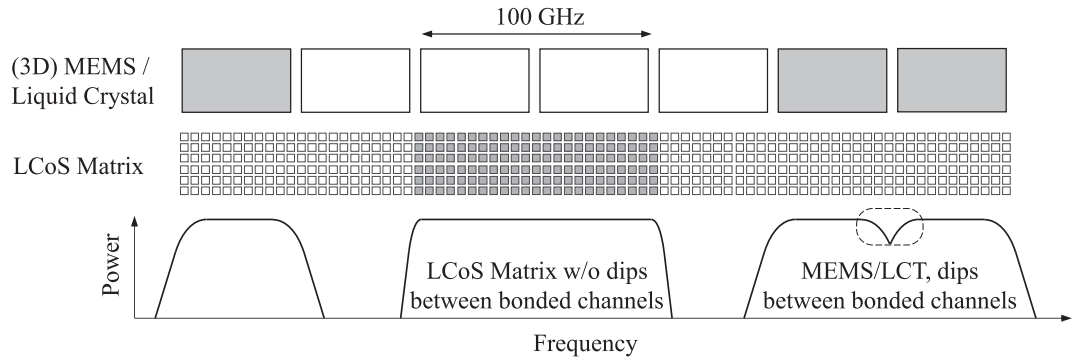


Fig. 28 LCoS flexible-grid matrix WSS.

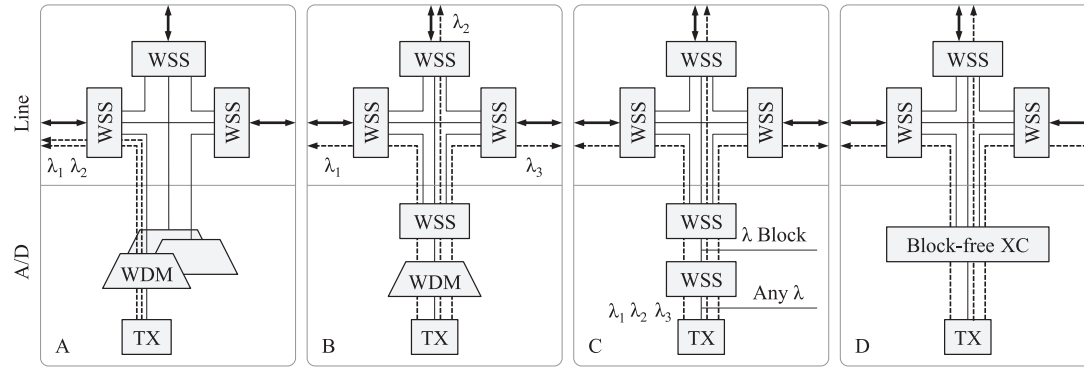


Fig. 29 Basic ROADM functionality. (A): colored per-degree add/drop. (B): directionless add/drop. (C): directionless and colorless add/drop. (D): directionless, colorless, and contentionless add/drop.

Wavelength-blocking is solved in **contentionless** ROADMs with a blocking-free cross connect (XC in Fig. 29(D)) in the add/drop. One implementation is shown in Fig. 30. It is based on WSSs in the outgoing (line) directions and power splitters for the incoming directions. The add/drop consists of WSS/power-splitter combinations, $N \times M$ switches, and a passive shuffle (i.e., patches) connecting all WSS/power-splitter combinations to all switches.

ROADMs, like other filters, exhibit bandwidth narrowing when being cascaded. This must be considered in transmission at high Baud rates. Depending on the filter and signal characteristics and number of cascaded filters, penalties on the receive-end OSNR must be applied. The effect can be counter-acted by proper spectral shaping, or with flexible-grid ROADMs.

Receivers

In WDM systems, two types of photo diodes are used as photo detectors. In a **PIN photo diode**, a p-n junction of suitable semiconductor material is used as high-speed photo detector. To improve responsivity, a lightly-doped intrinsic semiconductor is put between the p- and n-type semiconductors. To allow absorption in the intrinsic region, the p- and n-regions ideally are kept transparent.

The primary photocurrent resulting from absorption is given by $I_P = RP_0$. The responsivity R of the PIN diode results as $R = \eta q / h\nu$[A/W]. The external quantum efficiency η is the ratio of photo-generated electron-hole pairs and incident photons, $\eta = (I_p/q)/(P_0/h\nu)$. P_0 is the incident optical power, q is the electron charge, and $h(\nu)$ is the photon energy.

Fig. 31 shows the responsivity of InGaAs and Ge which are the relevant materials covering the relevant WDM wavelengths (Azadeh, 2009). For wavelengths $\lambda < 900$ nm, Si can be used.

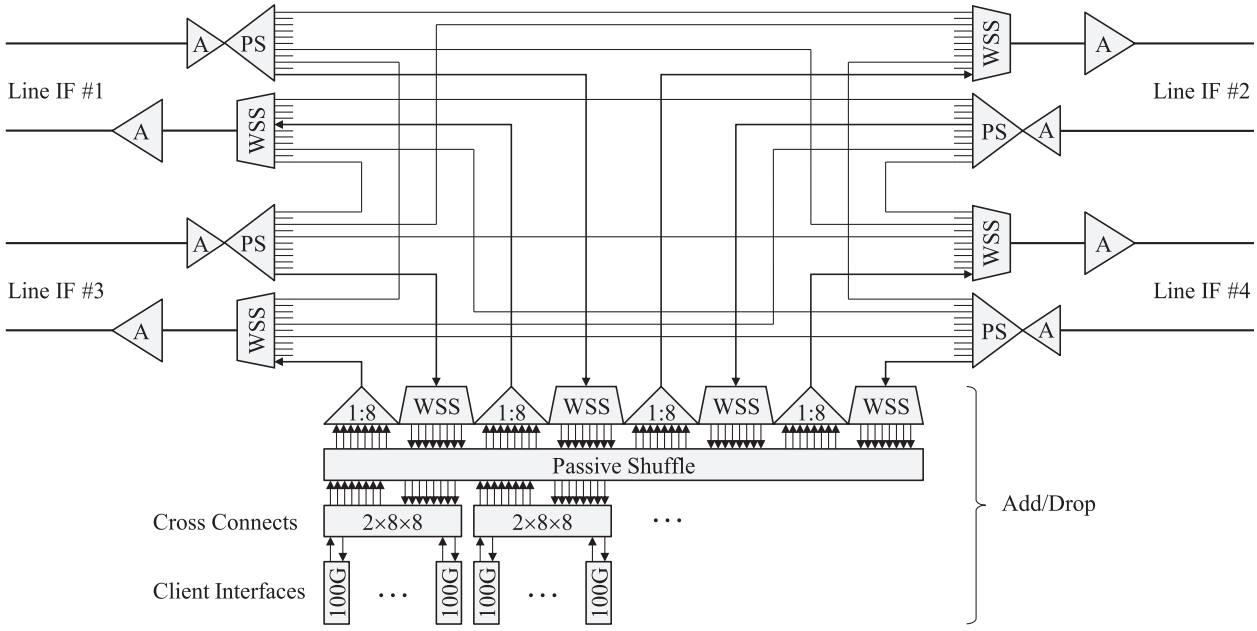


Fig. 30 Block diagram of Degree-N directionless, colorless, contentionless ROADM. PS: power splitter. A: amplifier (EDFA).

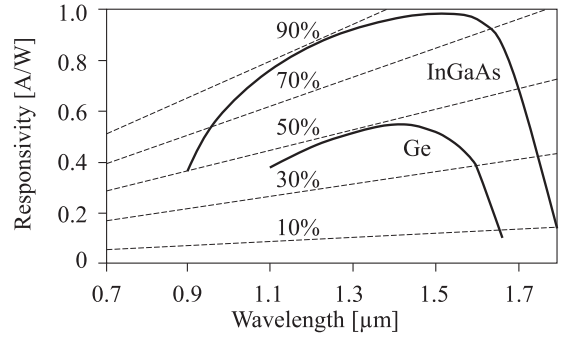


Fig. 31 Spectral responsivity of Ge and InGaas photo diodes.

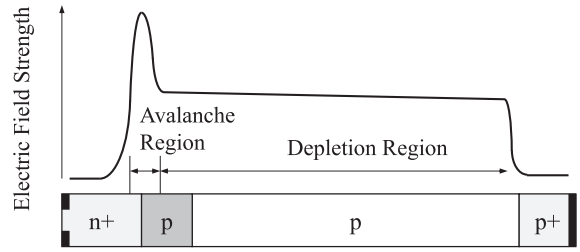


Fig. 32 Structure of an APD, and electrical field strength along its structure.

Table 4 Comparison of photo diodes

		Ge	*InGaAs*
Wavelength range, λ (nm)	PIN, APD	900–1650	1000–1700
Responsivity, R (A/W)	PIN	0.4–0.5	0.7–0.95
Quantum efficiency, η (%)	PIN	50–55	60–70
Avalanche gain, M	APD	50–200	10–40
k-Factor, k_A	APD	0.7–1.0	0.5–0.7
Bandwidth, B (GHz)	PIN	0.5–3	1–40
	APD	1.5 max.	1–5
Bias Voltage, V_B (V)	PIN	5–10	5–6
	APD	20–40	20–30

Reproduced from Azadeh, M. Fiber Optics Engineering. Dordrecht Heidelberg: Springer; 2009. ISBN 978-1-4419-0304-4.

Avalanche Photo Diodes (APDs) provide inherent gain. They are doped such that a region with very high field strength results when the diode is reverse-biased with high voltage. This region produces internal current gain due to impact ionization (avalanche effect). Higher reverse bias leads to higher gain, including stronger noise generation. The structure of an APD is shown in **Fig. 32**, together with the electrical field strength along the device (Azadeh, 2009).

APDs for WDM applications use the same semiconductor materials used for PIN diodes (see **Fig. 31**). InGaAs has less multiplication noise than Ge.

The responsivity of an APD can be expressed by the gain M, $R_{APD} = R_{PIN}M$ (M is set to 1 for PIN diodes). M is given by $M = I_M/I_P$, with I_M the average value of the total multiplied output current. Since the noise figure $F(M)$ increases with M, there is an optimum value of M that maximizes signal/noise.

A comparison of relevant photo-diode parameters is given in **Table 4** (Azadeh, 2009).

Non-Fiber-Related Effects

In Section Transmission Effects in Optical Fibers, effects related to the transmission fibers were discussed. Here, on overview is given on crosstalk (from demultiplexing) and noise (from amplification).

Linear Crosstalk

Linear crosstalk originates in WDM receivers from imperfect demultiplexing or faulty routing of certain channels. According to ITU Recommendation G.Sup39 (ITU-T, 2012b), two mechanisms exist:

- *Inter-channel crosstalk*. Ratio of total power in the disturbing channels to that in the wanted channel. Wanted and disturbing channels have *different* wavelengths.
- *Interferometric or intra-channel crosstalk*. Ratio of disturbing power *not* including ASE noise to the wanted power *within the same* wavelength.

Inter-channel crosstalk occurs when WDM channels are demultiplexed in a filter with non-perfect isolation. At the demultiplexer output port under consideration, the disturbing (neighbouring) channels are attenuated with respect to the wanted channel by the (adjacent) filter isolation$I$(dB). The worst case is given if the power of the channel under consideration is at its guranteed minimum, and all other channels are at their allowed maximum. This maximum difference is denoted d (dB) in **Fig. 33**. IL denotes the filter insertion loss.

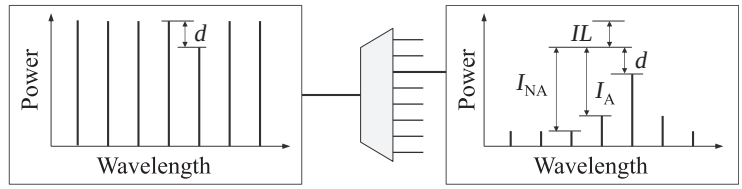


Fig. 33 WDM demultiplexer isolation.

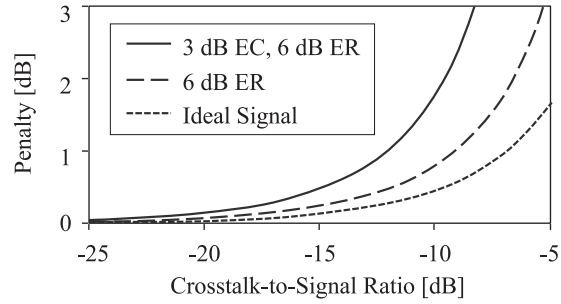


Fig. 34 Optical penalty vs. inter-channel crosstalk from single interferer.

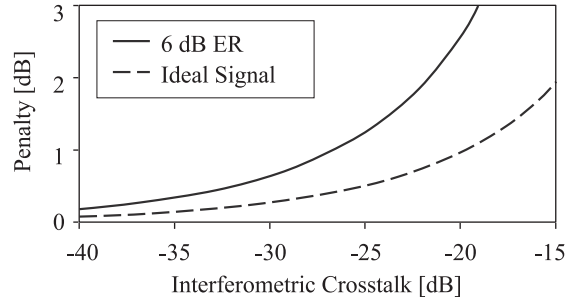


Fig. 35 Optical penalty vs. interferometric crosstalk (single interferer, optimum decision threshold).

In practical demultiplexers, the isolation I_A for the channels adjacent to the wanted channel is smaller than the isolation I_{NA} of the non-adjacent channels, $I_A < I_{NA}$.

The single-interferer inter-channel crosstalk penalty for various values of the eye closure (EC) and extinction ratio (ER) of the wanted signal is shown in Fig. 34. Penalties for real systems are somewhere below the highest curve. For multiple interferers, the penalty becomes smaller than the single-interferer ideal-signal penalty for small crosstalk, and higher for large crosstalk.

Interferometric crosstalk occurs when the disturbing channel and the wanted channel are at the same nominal wavelengths. Obviously, the disturbing signals cannot be suppressed with WDM filters. Interferometric crosstalk can occur:

- In a node where a wavelength is incompletely dropped before the new signal is added
- In a multiplexer where a transmitter emits at the wavelength of another channel. This can occur if channels are combined with power splitters rather than filters. Interference can then result from wrong laser tuning or insufficiently suppressed *laser side modes*.
- In a cross connect or MD-ROADM with poor isolation, where light from more than one network fiber (degree) can reach the respective receive port
- In any component, subsystem or network element with more than one light path that connects to the receiver. This effect is also called *Multi-Path Interference*.
- *Intentionally*. This is also referred to as crosstalk attack.

Interferometric crosstalk results from *field* beating. Therefore, lower crosstalk levels lead to the same penalties, compared to inter-channel crosstalk, see Fig. 35.

Interferometric multi-interferer crosstalk may have to be reduced by 2–4 dB as compared to single-interferer crosstalk for penalties in the range of 0.5–1 dB.

Noise in Optical Transmission Systems

Fiber and other components' loss leads to decreasing signal power levels. On long transmission links, this has to be compensated by optical amplifiers which amplify several WDM channels simultaneously. Like all amplifiers, optical amplifiers add noise (ASE). Together with the unavoidable noise sources in receivers, this decreases the Signal/Noise Ratio (SNR) which is a relevant measure for the achievable Bit-Error Rate (BER).

In order to derive the SNR which a photo diode generates, the primary noise sources must be considered. These are quantum (shot) noise and dark-current noise. Quantum noise arises from the statistical nature of photo-generated current. Dark-current noise results from the current that continues to flow through the bias circuit in the absence of light. It usually is not significant for PIN photo diodes, but can become significant in APDs. Further relevant noise sources include thermal noise from electronic receivers (Alexander, 1997) and ASE from optional optical pre-amplifiers. Consequently, PIN-based, APD-based and optically pre-amplified receivers must be considered separately.

Following, the achievable SNR and sensitivity of optical receivers (PIN, APD, optical pre-amplified) is given, followed by the noise figures of (chains of) optical amplifiers.

Achievable SNR (or, in the case of a pre-amplifier, noise figure) and sensitivities can be derived considering the relevant noise sources (Agrawal, 1992; Alexander, 1997). For weak and large PIN photo-diode input, thermal noise and shot noise dominate, respectively, leading to different SNR:

$$SNR_{\mathrm{PIN,therm}} = \frac{R_L {\sigma_P}^2}{4k_B TB},\ SNR_{\mathrm{PIN,Shot}} = \frac{{\sigma_P}^2}{2qI_P B} \tag{15}$$

R_L is the matched load resistance of the amplifier, $k_B = 1.38 \times 10^{-23}$ is Boltzmann's constant, T is the temperature, and B is the electronic (effective noise) receiver bandwidth. I_P is the photocurrent, q is the electron charge, and ${\sigma_P}^2$ is the input signal variance. These SNR values have to be decreased by the noise figure F_N of the electronic amplifier which follows the load resistor.

For APDs, SNR is dominated by shot noise, and avalanche gain M and excess noise factor $F(M)$ must be considered. The total shot noise (to replace $2qI_P B$ in Eq. (15)) is given by $\sigma_S^2 = 2q(I_P + I_D)BM^2F(M) + 2qI_L B$. I_D and I_L are the bulk dark current and surface current of the APD. The excess noise factor $F(M)$ is given by $F(M) = kM + (1-k)(2-1/M)$. The range for the factor k is $0 < k < 1$, see **Table 4**. For the avalanche gain M, also refer to **Table 4**.

Typical PIN receivers are dominated by thermal noise. They can be combined with *Optical Pre-Amplifiers* (OPA) in order to increase sensitivity. OPAs generate signal-ASE and ASE-ASE beat noise, respectively. For large OPA gain G, signal-ASE beat noise becomes the sole dominant noise. Then, the OPA noise figure F_N can be approximated as $F_N \approx 2n_{sp}$. Here, n_{sp} is the population-inversion factor. Large gain requires high population inversion, hence ideally $n_{sp} \approx 1$ and consequently, the optimum OPA noise figure is $F_{N,opt} \approx 2$. An ideal low-noise OPA has a noise figure of 3 dB and helps avoiding the thermal-noise limit of PIN diodes.

Sensitivity of an optical receiver is defined as the minimum average power which is required to achieve a BER of 10^{-12}. For On-Off Keying with direct detection, sensitivity of approximately -26 dBm, -36 dBm and -50 dBm results for receivers with PIN diodes, APDs, and OPAs, respectively. For the APD, $M = 10$ and $F(M) = 1.3$ have been assumed.

The achievable sensitivities heavily depend on the modulation and detection scheme. Sensitivity increases from direct detection via heterodyne detection to homodyne detection. In Tonguz and Wagner (1991), the equivalence between optically pre-amplified direct and asynchronous heterodyne detection has been shown. Quantum-limited sensitivity of 9 photons per bit is reached for coherent homodyne PSK reception.

Noise in cascaded optical amplifiers is most often derived in a semiclassical description using field-beating theory (Desurvire, 1994), considering signal-ASE beating.

A co- and counter-pumped dual-stage EDFA is shown in **Fig. 36**.

Here, *Friis' formula* for the total noise figure F_{tot} of cascaded amplifiers applies, $F_{tot} = F_1 + (F_2 - 1)/G_1$ (Friis, 1944). For sufficiently large gain G_1, the total noise figure is dominated by the first stage. The first EDFA stage should be pumped with 980 nm, which achieves better inversion as compared to 1480-nm pumping. This leads to noise figures of 980-nm-pumped EDFAs which are typically better by 1 dB, compared to 1480-nm pumping (Kaminow and Koch, 1997).

Cascaded amplifiers in long-haul links are shown in **Fig. 37**. Amplifiers have total noise figures F_{Ni} and gains G_i, and the intermediate spans have attenuations A_i (given by αL_i, with α the loss coefficient and L_i the respective span length).

The noise figure of a chain of amplifiers along the link is then given by:

$$F_{\mathrm{total}} \approx \frac{F_1}{A_1} + \frac{F_2}{A_1 G_1 A_2} + \ldots + \frac{F_n}{A_n G_1 A_1 \cdots G_{n-1} A_{n-1}} \tag{16}$$

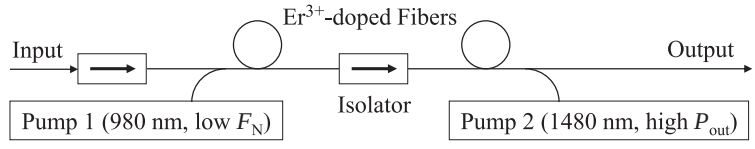


Fig. 36 Dual-stage EDFA.

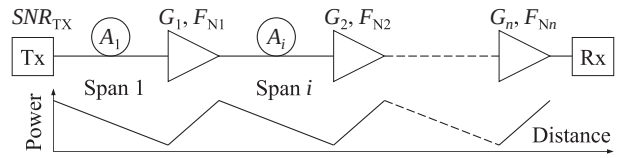


Fig. 37 Long-haul link with cascaded amplifiers.

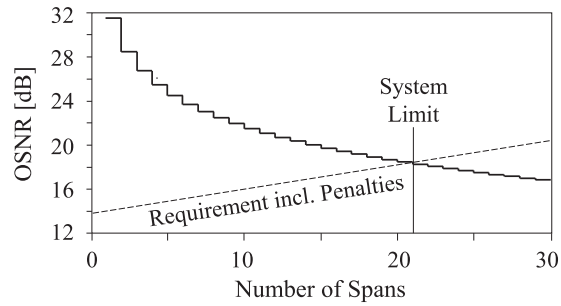


Fig. 38 Development of OSNR in long chains of (equidistantly spaced) amplifiers.

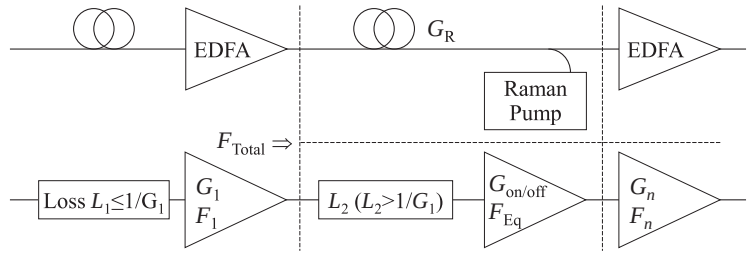


Fig. 39 Multi-span link with EDFAs and Raman pre-amplifiers, and equivalent model.

If $G_i \gg 1/A_i$, only the first amplifier dominates the receive-end noise behavior. If $G_i \leq 1/A_i$, then F_{total} is increased significantly by the respective spans (Loudon, 1985). If all amplifiers have equal gain G_i, and all links have equal loss A_i with $G_i = 1/A_i$, one gets $F_{total} = nF_N/A_i$. The resulting OSNR is often expressed in dB:

$$\mathrm{OSNR}_{0.1\mathrm{nm}} = P_{\mathrm{out,Ch}} - A - F_N - 10\log(n) + 58 \tag{17}$$

This definition relates to a reference bandwidth of $B_{o,ref} = 125$ GHz (or 0.1 nm). A is the mean span loss (now in dB), n is the number of spans (or amplifiers). The OSNR in long-haul systems, assuming $G = 1/A$, drops logarithmically with increasing span number, see **Fig. 38**. Here, noise figures of $F_N = 6.5$ have been assumed. The system limit is explained in Section Long-Haul WDM Systems.

In ultra-long-haul systems with low-noise design, several 100 amplifiers can be cascaded (Loudon, 1985). In regional systems with irregular span lengths (and spans with $A_i > 1/G_i$), the number of amplifiers is much smaller, often it is < 20.

The OSNR resulting from very long single spans or very long multi-span links can be improved with FRA. This is often done with backward-pumped Raman pre-amplifiers, acting as *distributed* pre-amplifiers to subsequent EDFAs, see **Fig. 39**. The EDFA input is prevented from getting a certain critical OSNR, as shown in **Fig. 20**. This results in improved noise figure of the combined amplifiers.

Often, distributed FRA are described via an *Equivalent* Noise Figure F_{Eq} and the Raman on/off gain $G_{on/off}$ of a *lumped* Raman amplifier (Bristiel *et al.*, 2006). These are given by $F_{Eq} = F_{N,FRA}\exp(-\alpha_S L)$ and $G_{on/off} = G_R(L)\exp(\alpha_S L) = \exp(g_R P_{P0} L_{eff})$. $F_{N,FRA}$ is the noise figure of the FRA, and $G_R(z)$ and g_R are the length-dependent Raman gain and the Raman-gain coefficient (see **Fig. 7**), respectively. P_{P0} is the pump power launched at $z = L$, and L_{eff} is the effective fiber length at the pump wavelength. α_S is the fiber loss coefficient at signal wavelengths. The on/off gain is explained in **Fig. 40** (also see **Fig. 20**).

Depending on fiber characteristics (loss, mode-field diameter, CD) and patch and splice losses between fiber segments, typical value of $F_{Eq} \approx -2.5$ dB and $G_{on/off} = 15$–20 dB result. The total noise figure F_{tot} of combined EDFA and Raman pre-amplification is given by:

$$F_{tot} = F_{Eq} + [(F_{EDFA}/\alpha_c) - 1]/G_{on/off} \tag{18}$$

Here, α_c is the (coupling) loss between the fiber output (at the end of L_2) and the EDFA.

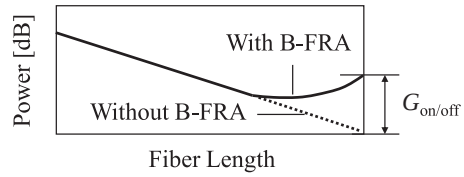


Fig. 40 On/off Raman gain. B-FRA: backward-pumped FRA.

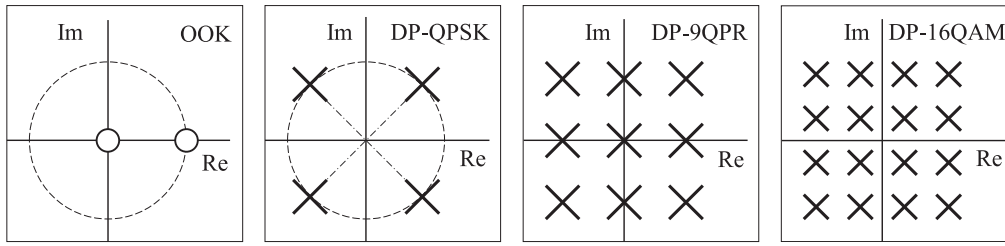


Fig. 41 Some relevant fiber-optic modulation schemes' constellation diagrams.

The receive-end OSNR in very long multi-span transmission can be improved by Raman-only amplification, using forward- and backward-pumped distributed FRAs. Noise figures for forward-pumped FRA can achieve $F_{N,FRA}=4–5$ dB (Premaratne, 2004).

Modulation of WDM Signals

Modulation Formats

Typically, WDM channels up to 100 Gb/s per channel are modulated individually, without considering other channels. Starting with 100 Gb/s, the respective capacity has also been split onto several *sub-carriers* which together form so-called WDM *super-channels*. Sub-carriers are modulated via a common entity.

Per WDM channel or sub-carrier, all parameters of the respective laser can be used for modulation. The simplest (and oldest) variant is *intensity modulation*. It can be detected non-coherently in direct-detection receivers, leading to Intensity Modulation with Direct Detection (IM/DD). For binary modulation, this is also called On-Off-Keying (OOK). Currently, a variant with four amplitude levels, Pulse-Amplitude Modulation-4 (PAM-4), is under consideration for shorter-reach 400-Gb/s WDM super-channel transport with direct detection.

If the instantaneous frequency of the laser is modulated, this is called *Frequency-Shift Keying*. It does not play an important role in WDM transmission.

If the carrier phase is chosen to carry the information, this is called *Phase-Shift Keying* (PSK). PSK has been used in commercial WDM system in the binary and quarternary, differentially modulated versions (DBPSK, DQPSK). DBPSK is still relevant for ultra-long-haul WDM. DQPSK is identical to differentially encoded 4QAM.

The combination of (multi-level) amplitude or intensity modulation and PSK leads to *Quadrature Amplitude Modulation* (QAM). Today, coherently detected QAM dominates ultra-high-speed WDM transmission.

In coherent detection, receivers must detect signals in orthogonal polarization planes, because the PSPs of the receive signal are generally not known. Then, two orthogonal polarizations can also be used for independent modulation and associated capacity increase. This is often called *Dual-Polarization* (DP) modulation.

Any modulation can use *full-response* or *partial-response* pulses. For Partial Response (PR), the pulse spreads over several symbols, causing intentional Intersymbol Interference (ISI). PR leads to smaller signal bandwidth and better CD tolerance. The added ISI can be eliminated with receive-end Maximum-Likelihood Sequence Estimation. The simplest form of PR in WDM transmission is Optical Duobinary (ODB), the PR variant of OOK. PR can be extended to complex symbol constellations, leading to *M*-ary Quadrature Partial Response (QPR).

Often, (de-) modulation is characterized graphically in constellation diagrams where the respective Inphase (I) and Quadrature (Q) components of each symbol state are shown. For DP modulation, this can be done for both polarizations. In **Fig. 41**, some examples are shown. Crosses (×) indicate DP modulation.

Independent from full vs. partial response, pulse or spectral shaping can be added. In some long-haul systems, this was done by Return-to-Zero (RZ) pulse shaping to increase tolerance against nonlinearity. Different RZ duty cycles were implemented, e.g., 33%, 50%, and 67% (called RZ33, RZ50 and RZ67, respectively). All RZ variants increase bandwidth over Non-Return-to-Zero (NRZ). Some resulting optical power spectra, normalized to the bit rate, are shown in **Fig. 42**.

Pulse (or spectral) shaping can also be done in order to decrease the bandwidth of (sub-) carriers. The intention here is to allow denser (sub-) carrier spacing, in order to increase spectral efficiency. This approach is sometimes referred to as Nyquist WDM (Bosco *et al.*, 2010). Spectral shaping for a super-channel with four sub-carriers is visualized in **Fig. 43**.

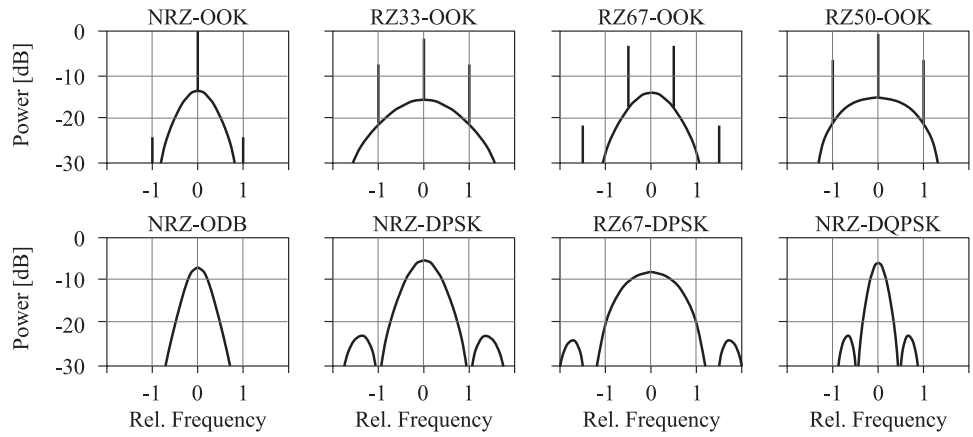


Fig. 42 Effect of modulation, pulse shaping and duty cycle on optical bandwidth.

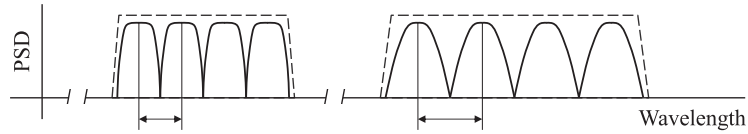


Fig. 43 Spectral shaping of sub-carriers to allow denser spacing.

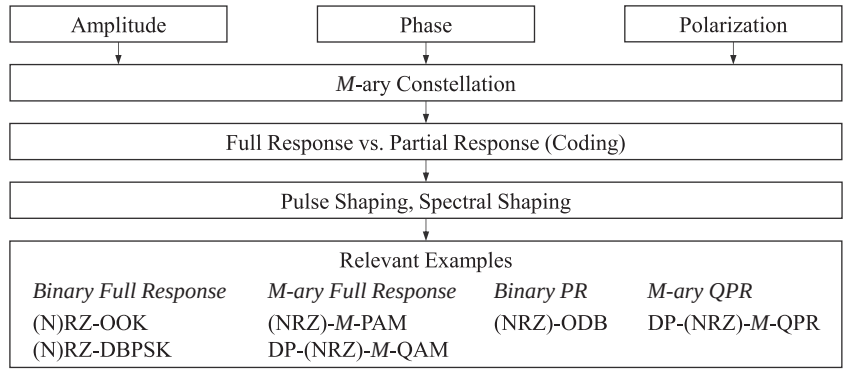


Fig. 44 Overview on fiber-optic modulation and pulse-shaping schemes.

The classification of modulation schemes suitable and relevant for WDM is given in Fig. 44, together with (commercially) relevant past and present examples.

Up to and including 10-Gb/s WDM transmission, NRZ-OOK (IM/DD) has been most important. For 40 Gb/s, RZ-OOK, ODB and DPSK have been used, and first DP-DQPSK systems been introduced. For 100 Gb/s per channel and beyond, long-haul systems are based on DP-QAM with intradyne detection. Shorter reach, e.g., for data-center interconnects, can also be based on ODB or PAM-4 with multiple sub-carriers.

Achievable Bit-Error Rates

Closed analytical calculation of the receive-end Bit Error Rate (BER) P_B of a WDM transmission channel including fiber nonlinearity, PMD, CD, crosstalk and noise is not possible. Hence, a simplified AWGN (Added White Gaussian Noise) system model must be used, e.g., Agrawal (1992), Linke and Gnauck (1988). All distorting effects must be considered via penalties on the receive-end OSNR.

For coherent detection, the analytical analysis leads to the Q-function which can be derived from the Complementary Error Function, erfc:

$$\operatorname{erfc}(x) = 2Q\left(\sqrt{2}x\right) = \frac{2}{\sqrt{\pi}\int_x^{\infty} e^{-y^2}\,dy} \tag{19}$$

Following, the AWGN-channel BER is derived for relevant (de-) modulation techniques.

Intensity modulation

Intensity modulation can be split into *binary modulation* (OOK), and multi-level modulation. The latter is called Pulse-Amplitude Modulation (PAM), with PAM-4 currently being the most relevant variant for WDM. OOK is mostly detected directly. Then, the BER results as:

$$P_B = \frac{1}{2}\exp\left(-\frac{E_b}{4N_0}\right) \tag{20}$$

E_b is the bit energy (in J or Ws), N_0 is the noise-power density (in W/Hz). The simplified block diagram of an NRZ-OOK system is shown Fig. 45. The transmitter consists of a continuous-wave Laser Diode (LD) and an external modulator. The receiver consists of pre-amplifier, Optical Band-Pass Filter (OBPF) and direct-detection receiver. The OBPF is necessary for WDM channel separation and as a noise-bandwidth limiter. The direct-detection receiver consists of a Sample-and-Hold (S + H) unit which approximates a matched filter, followed by a decision unit.

PAM-4 can basically use the same optical frontend as OOK. Added effort results from the 4-level signal generation (including serial-parallel (S/P) and parallel-serial (P/S) conversion), and the necessity to equalize the received signal, refer to Fig. 46.

PAM-4 is currently under investigation for WDM applications with ~100 km reach, e.g., for data-center interconnects. For 400-Gb/s transport, super-channels with 8×56 Gb/s can be used.

Phase-shift keying

PSK has first been considered for ultra-long-haul WDM transmission in the 1990s because of its superior OSNR. For homodyne detection, the BER of BPSK is given by:

$$P_B = Q\left(\sqrt{\frac{2E_b}{N_0}}\right) \tag{21}$$

This is the best BER achievable by the modulation formats discussed in Section Modulation Formats.

Early commercial BPSK systems were based on non-coherent detection with differential precoding and delay demodulation (DBPSK). Its BER is:

$$P_B = \frac{1}{2}\exp\left(-\frac{E_b}{2N_0}\right) \tag{22}$$

This BER is worse as compared to homodyne BPSK by ~3.5 dB.

The most relevant implementation of *multi-level PSK* is (Differential) QPSK. It has been used in early non-coherent systems, and today is the basis of the quasi-standardized intradyne 100 Gb/s transmission. The BERs for non-coherent and homodyne detection are 3 dB worse compared to BPSK, for the same symbol energies E_S.

Quadrature amplitude modulation

The term QAM is often applied to rectangular symbol constellations (see Fig. 41), however, other constellations based on multiple PSK stars or even non-symmetric constellations exist.

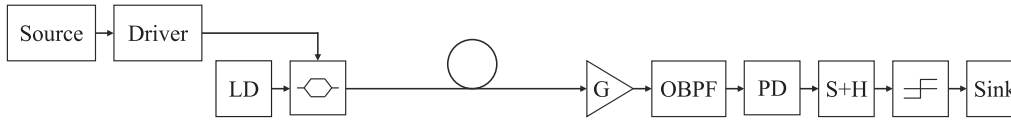


Fig. 45 NRZ-OOK with external modulation and Direct Detection (DD).

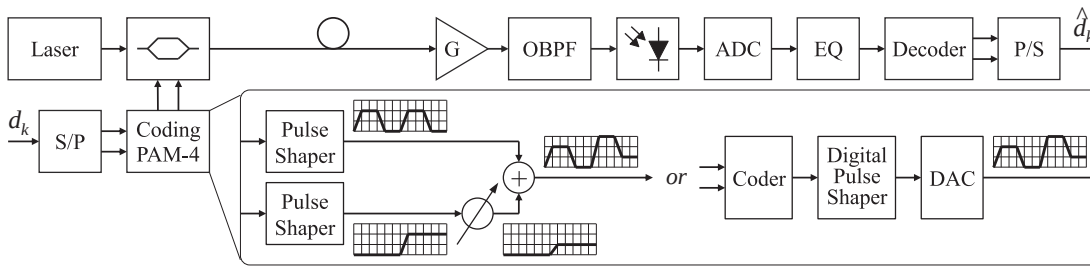


Fig. 46 PAM-4 system with two alternatives for 4-level signal generation.

For rectangular constellations and homodyne detection, the *symbol error rate* is given by:

$$P_S(M) = 4\left(1 - \frac{1}{\sqrt{M}}\right)Q\left(\sqrt{\frac{3\log_2 M}{M-1}\frac{E_b}{N_0}}\right) \tag{23}$$

The BER is given by $P_B \approx P_S/\log_2 M$ for $P_S \ll 1$.

Today, long-haul ultra-high-speed WDM relies on Dual-Polarization QAM with digital **intradyne detection**. For 100-Gb/s transport, this leads to DP-4QAM (DP-QPSK). For bit rates $\geq$200 Gb/s, DP-16QAM can be used, eventually with a multi-carrier approach. For tight sub-carrier spacing, this may be complemented by Partial-Response Signaling (QPR) or other spectral shaping. Typically, differential encoding is used to avoid error propagation.

Intradyne detection is based on tuning the local laser such that the resulting intermediate frequency is close to zero. Then, analog-to-digital conversion is performed, and exact phase tracking is done in the digital domain. This allows implementing digital dispersion-compensating filters which lead to almost zero penalty with regard to polarization and CD effects.

Coherent systems require polarization-diverse receivers. Therefore, they make use of polarization multiplexing. The resulting configuration for a Dual-Carrier system is shown in **Fig. 47**. For single carriers, a similar configuration was reported in Savory *et al.* (2007).

Each transmit laser is split into orthogonal polarizations by means of a polarization beam splitter (PBS). Both polarization signals are independently modulated and re-combined by a polarization beam combiner (PBC). At the receiver, the input signal is split into orthogonally polarized signals by another PBS. The local laser signal is also split and then combined with the input signal by means of two 90° hybrids. Each 90° hybrid has dual output ports for the respective inphase and quadrature components. The output signals are detected by four balanced receivers. These are followed by fast Analog-to-Digital Converters (ADC) which feed the digital signal processing (DSP). **Fig. 48** shows the DSP in more detail.

After detection and sampling at approximate Nyquist rate, the I and Q components have orthogonal but arbitrary polarization planes each. The four components are then processed in several filter stages. In a first stage, bulk CD compensation is performed. This stage can be complemented by Nonlinear Compensation (NLC). CD compensation is based on an n-tap Feed-Forward Equalizer (FFE). The upper bound for the tap number at a Baud rate of B GBd is given by $0.032 \cdot B^2$ per 1000 ps/nm of chromatic dispersion (Savory *et al.*, 2007). At 28 GBd (112-Gb/s DP-QPSK), this translates to 25 taps per 1000 ps/nm.

NLC primarily tackles intra-channel effects (SPM). Inter-channel effects (XPM, FWM) require equalizers which are currently too complex. Linear and nonlinear intra-channel compensation can be done iteratively, switching between time and frequency domain by means of FFT and IFFT, respectively, see **Fig. 49**. This approach is called Backward Propagation. The improvement in launch power (and therefore, OSNR) of NLC typically is limited to 1–1.5 dB.

In the next filter stage, Clock Data Recovery (CDR) and re-sampling to 2 samples/symbol are performed. CDR is based on a digital filter-and-square timing recovery. It also performs re-sampling to 2 samples/symbol.

PMD and residual-CD equalization, and polarization recovery (demultiplexing) is performed in a Multiple-Input Multiple-Output (MIMO) filter which follows re-timing. The filter tap weights are usually optimized using blind adaptation. This can be done with the Constant-Modulus Algorithm (CMA) (Johnson *et al.*, 1998). The CMA can equalize M-ary PSK signals with $M \geq 4$ only. Derivatives for BPSK, QAM and QPR exist.

The last filter stage performs Carrier Phase Estimation (CPE), i.e., the digital phase tracking. CPE for QPSK can be done with 4th-power Viterbi-and-Viterbi recovery. For BPSK, 2nd-power estimation can be used, whereas for (square) QAM, modifications are required. These can consist of considering symbols of equal magnitude separately.

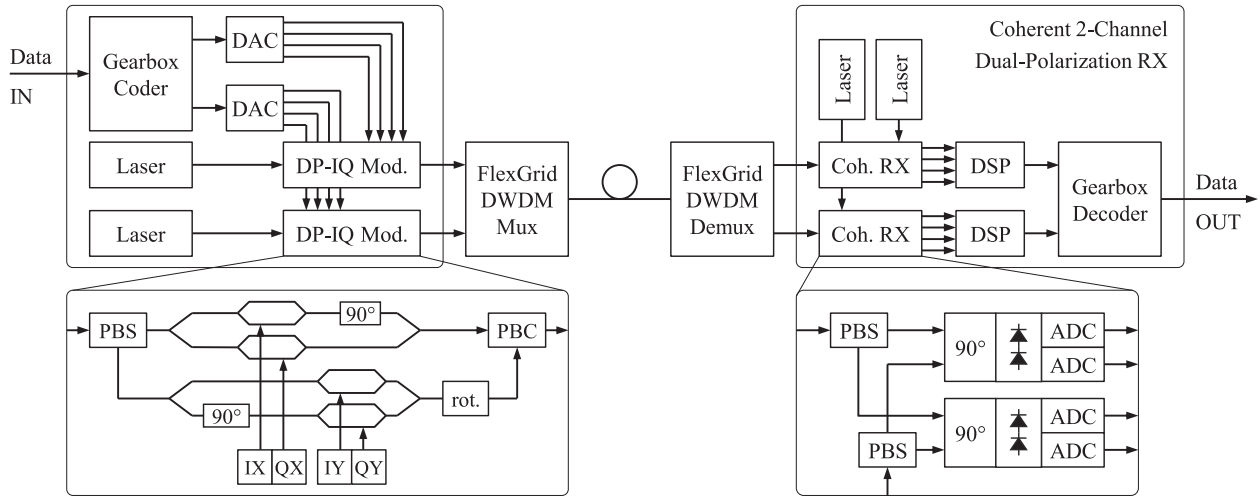


Fig. 47 Coherent intradyne Dual-Carrier DP-QAM (DP-QPSK) transmission.

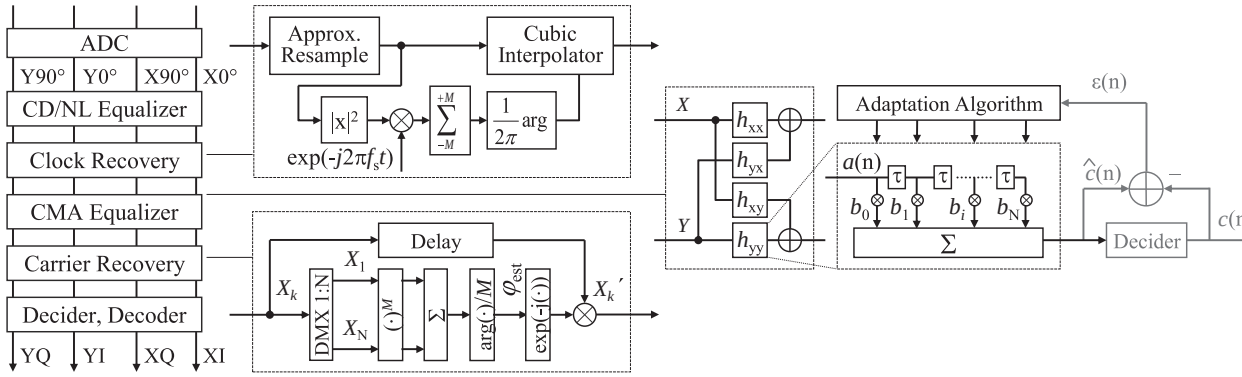


Fig. 48 Coherent intradyne receiver: digital realization.

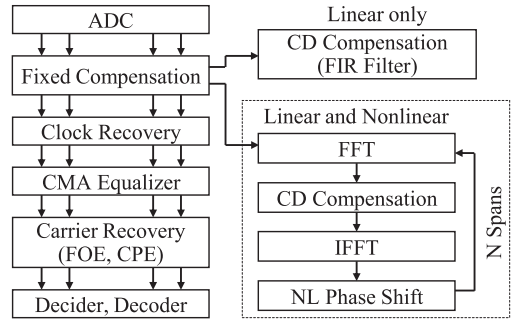


Fig. 49 Linear-only or linear plus nonlinear compensation (backward propagation).

Partial response signaling

Optical Duobinary, ODB, is the simplest form of *partial response* signaling in WDM transmission. It intentionally produces *known* ISI in order to achieve higher spectral efficiency. The ODB spectrum is compared to OOK and PSK in **Fig. 42**. PR signaling has been studied since the 1960s (Kobayashi and Tang, 1971).

ODB encoding can be achieved by a delay-and-add filter, followed by a Low-Pass Filter (LPF). The LPF bandwidth is 0.5-–0.75 · R_B (R_B the bit rate). This encoder can be replaced by a sole LPF with bandwidth in the range of ∼0.28 · R_B (so-called *filtered* Optical Duobinary). The ISI introduced at the transmitter can be unraveled at the receiver by differential decoding. Since this can produce error propagation, differential *precoding* at the transmitter is preferred, allowing symbol-by-symbol detection. The resulting transmitter is shown in **Fig. 50**.

With direct detection, the BER of ODBis in therange of the BER of OOK, depending on receiver implementation. ODB can also be detected coherently (Lyubomirsky, 2006). The BER is then given by:

$$P_B \approx \frac{3}{2} Q\left(\frac{\pi}{2}\sqrt{N_P}\right) \tag{24}$$

$N_P = E_b/hf_0$ is the average per-bit photon number. With optimized threshold, homodyne ODB achieves sensitivity (at BER$=10^{-9}$) of 15 photons/bit (compare to 9 photons/bit for BPSK).

ODB can be used for cost-effective medium-reach WDM transport of 100-Gb/s signals. Then, four DWDM sub-carriers carrying ∼28 Gb/s each are used, see **Fig. 51**.

Reach according to **Fig. 51** is OSNR-limited and in the range of ∼500 km. The sub-carrier spacing is 50 or 25 GHz. As compared to 50 GHz, 25 GHz leads to a reach penalty which is caused by higher sub-carrier crosstalk and filter effects. CD tolerance is in the range of ±300 ps/nm, mandating optical CD compensation (Section Chromatic-Dispersion Compensation) (ADVA, 2016).

Partial response coding can be applied to complex signals, leading to *Quadrature PR* (QPR). QPR is under frequent study for spectrally efficient WDM transmission (Lyubomirsky, 2010). Here, 9QPR has gained relevance so far. 9QPR has a symmetric 3 × 3 constellation, see **Fig. 41**.

The BER for coherent homodyne M-ary QPR ($M=9, 25, 49, \ldots$) is given by (Smith, 2003):

$$P_{B,M-\mathrm{QPR}} = \frac{4}{\log_2 L}\left(1-\frac{1}{L^2}\right) Q\left[\frac{\pi}{2}\sqrt{\frac{3E_b(\log_2 L)}{(L^2-1)N_0}}\right] \tag{25}$$

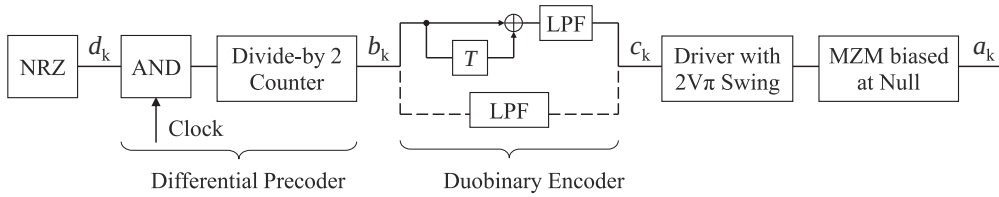


Fig. 50 Optical Duobinary coding.

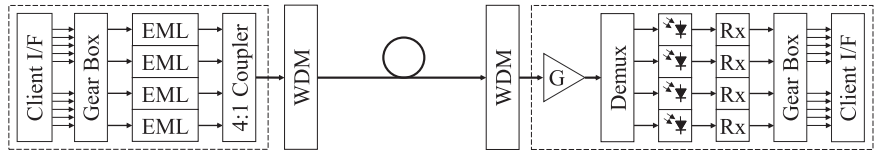


Fig. 51 High-speed 4-subcarrier Duobinary system with Externally Modulated Lasers (EML) and direct detection. Gear box is the synchronization circuit needed for 100-Gb/s client interfaces.

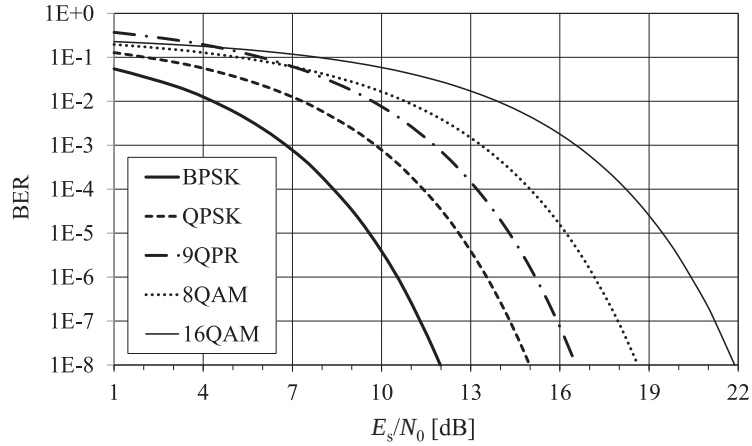


Fig. 52 BER of relevant modulation schemes (homodyne detection) for same symbol energy.

here, $L=\sqrt{M}$. 9QPR leads to ~2 dB penalty compared to QPSK, and a gain of ~2 dB over 16QAM, for coherent detection and same bit energy. This does not yet take an optimum receiver into consideration. For QPR, this is an MLSE which ideally can eliminate all penalty caused by PR coding. Simplified MLSEs have been shown for ODB and 9QPR. They are called Ambiguity Zone Detectors (AZD) (Kobayashi and Tang, 1971). AZD can improve the penalty for 9QPR against QPSK to ~1 dB at little implementation effort.

Comparison of modulation formats

In Fig. 52, bit-error rates are shown for the coherent modulation schemes that are most relevant for WDM transmission today. All schemes are normalized to the same symbol energy. The incoherent schemes DBPSK, OOK and ODB are not shown here. The resulting DBPSK BER curve is located very near to (~0.5 dB on the right-hand side of) the QPSK curve. OOK BER is 3 dB worse compared to DBPSK, and ODB BER is in the range of OOK BER.

These results imply ideal reception. Real systems (specially the receivers) lead to penalties over the numbers given in this chapter. These penalties are in the range of ≥3 dB, depending on implementation details.

System Realization

WDM in Metro and Access Networks

Large telecommunications networks are hierarchically organized. Residential and business access is based on wireless (2G, 3G, 4G, 5G upcoming) and wireline (fiber point-to-point, Passive Optical Networks (PON), Hybrid Fiber-Coaxial (HFC), copper twisted-pair) technologies. Access concentrators (base stations, PON Optical Line Terminations, Multi-Service Access Nodes, etc.) are

backhauled to aggregation sites of a first level. These sites are often referred to as Local eXchanges (LX) or Central Offices (CO). This backhaul is based on CWDM or DWDM. There is also a strong trend towards mobile *fronthaul* in LTE-Advanced (4G, 5G) networks. Here, several Baseband Units (BBU) are concentrated in BBU Hotels (BBUH) and connected to their respective antennas via digital high-speed links. Since these links must not be statistically multiplexed, and have tight latency and jitter requirements, they must run via point-to-point fiber or WDM channels.

Several LX can be connected, mostly via DWDM, to metropolitan-area core Points-of-Presence (PoPs). PoPs can accommodate aggregation switches of a secondary level, and they are connected via protected DWDM rings or mesh networks. Typically, the metro core network is connected to national and international backbones via two redundant PoPs (which accommodate core routers, Broadband Remote Access Servers (BRAS), etc.). The backbones use high-capacity, long-haul DWDM. An example network is shown in **Fig. 53**.

Real networks can deviate from **Fig. 53**. They may have less aggregation levels. They can be based on differing combinations of ring, mesh or point-to-point structures, and they can also use legacy technologies like SONET/SDH which are not shown here. However, most networks have in common that they use some sort of MPSL-over-WDM in the core, and some sort of Ethernet-over-WDM in the backhaul.

Metropolitan-area networks have been relevant applications for DWDM from the 1990s. Metro and regional networks are based on multi-span WDM systems. Many of these networks rely on *ring* architectures. One reason for (WDM) rings is that they require comparatively few fibers for redundant (i.e., protected) connections between a given number of sites. Rings are also relatively easy to manage (Maier *et al.*, 2002). A metro DWDM ring system is depicted in **Fig. 54**.

WDM nodes consist of add/drop filters or ROADMs, amplifiers, filters and termination for an Optical Supervisory Channel (OSC, which is part of the Data Communications Network), and optional CD compensation. The static filter structure shown in **Fig. 54** is increasingly being replaced by ROADMs. Main advantages of ROADMs include the capabilities of single-channel add/drop, which increases the number of logical connections without wavelength conversions (regenerations), and remote reconfiguration.

So far, most metro WDM systems make use of CD compensation. Main reason is the mixture of services, including transparent carriers' carrier services, which does not allow the exclusive use of coherent systems with in-built equalization. Due to distance restrictions, effects of accumulated nonlinearity are smaller than in long-haul.

Parameters for metro WDM link design are summarized in **Table 5** (Grobe and Eiselt, 2014).

A difficulty in metro WDM link design is the potential mixture of services and bit rates, and the related transceiver technologies and parameters. On certain links, it may be impossible to perfectly align receiver sensitivities, power levels or CD compensation requirements, which leads to additional penalties.

WDM in Data Center Interconnects

Relevant WDM applications are reach extensions for Storage Area Networks (SAN), and data center interconnects. Often, these are point-to-point applications which require very high WDM capacity over distances of <100 km.

Disk mirroring is primarily related to disaster recovery (DeCusatis, 2008). It provides redundancy for those data centers where failure/unavailability will cause significant costs. Disk mirroring can be performed *synchronously* or *asynchronously*. Synchronous disk mirroring means that the remote server can take on action immediately *without any interruption*. It has strict latency requirements which limit maximum distance between the data centers. Asynchronous mirroring allows some *limited* data loss or application interruption. It has relaxed latency requirements. In **Table 6**, typical latencies of WDM system components are listed.

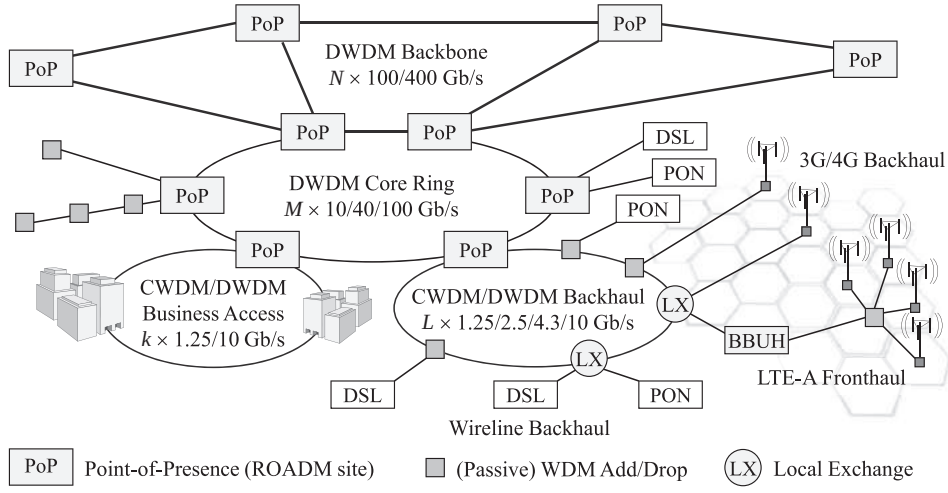


Fig. 53 Hierarchical fiber-optic telecommunications network. BBUH: Baseband Unit Hotel.

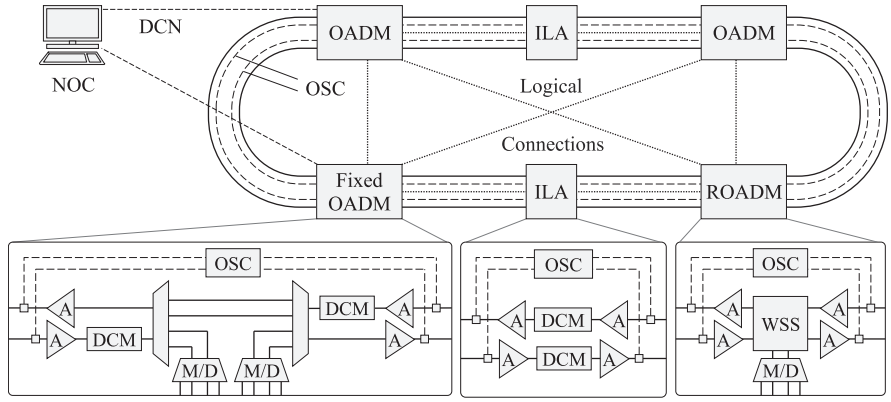


Fig. 54 Metro WDM ring. A: amplifier, DCM: Dispersion Compensating Module, DCN: Data Communications Network, ILA: In-Line Amplifier, M/D: Mux/Demux, NOC: Network Operations Center, OSC: Optical Supervisory Channel, WSS: Wavelength-Selective Switch.

Table 5 Parameters of metro and regional WDM link design

Parameter	*Value*
End-of-Life Penalty (Patches etc.) (dB)	2–3
Average Fiber Loss (dB/km)	0.25 (regio)–0.35 (metro)
Optical Path Penalty for CD, PMD, Nonlinearity (dB)	1 (2G5)–2.5 (10G, 40G)
Insertion Loss of WDM Filter plus OSC, east or west (dB)	7–10
Filtering Penalty for serial 40G Services for 50-GHz System	0.5 dB/Node
Typ. OSNR Requirement 2G5/10G/40G/coh. 100G (dB)	18 (no FEC) / 14/17/15 (FEC)
Typ. Sensitivity 2G5/10G/serial 40G/coh. 100G (dBm)	−28/−22/−20/−20
Typ. guaranteed Launch-Power Window (dB)	0–4
Typ. Amplifier Gain/Noise Figure (dB)	20–25/5.5–6.5

Reproduced from Grobe, K., Eiselt, M.H., Wavelength Division Multiplexing: A Practical Engineering Guide. Hoboken, NJ: John Wiley & Sons; 2014. ISBN 978-0-470-62302-2.

Table 6 Latencies of different transmission systems

Transmission system	*Latency/(group) delay*
Transparent WDM transponder w/o framing	$<$10 ns
TDM multiplexing with OTN framing	5–20 μs
FEC (10 Gb/s, 100 Gb/s)	5–10 μs
EDFA (single-stage)	$\sim$200 ns
100 km SSMF round-trip	$\sim$1 ms

Reproduced from Grobe, K., Eiselt, M.H., Wavelength Division Multiplexing: A Practical Engineering Guide. Hoboken, NJ: John Wiley & Sons; 2014. ISBN 978-0-470-62302-2.

In data centers, often Ethernet is used for the LAN, and Fiber Channel (FC) is used for the SAN. In addition, InfiniBand may be used for high-performance compute clusters (DeCusatis, 2008). In highly available data centers, key components are duplicated per site, and sites are connected for even higher availability, and to create virtual servers. Data-center interconnection then uses two separate WDM links, see **Fig. 55**.

Some applications require very high availability. In such mission-critical applications, all major components are duplicated, including the transmission paths. If necessary, more than two fiber connections (ducts) can be used. Options for redundant WDM transport are shown in **Fig. 56**.

The options shown here can be combined for highest path availability in the range of 99.9999%.

Long-Haul WDM Systems

In long-haul WDM transport, all system impairments (linear and nonlinear effects, noise) have to be considered. The influence of these effects varies with data rate, modulation format, fiber type, dispersion management, and the type of optical amplification.

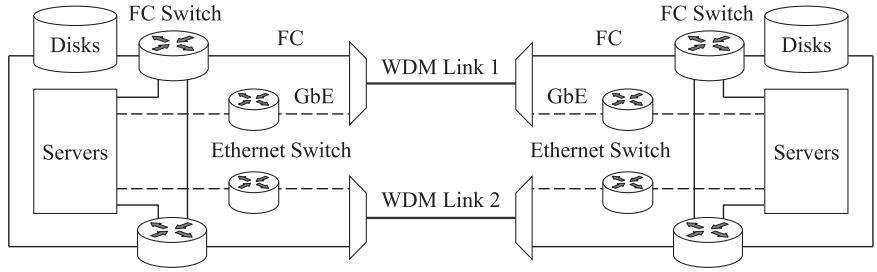


Fig. 55 Data-center interconnect via Ethernet (LAN) and FC (SAN) links.

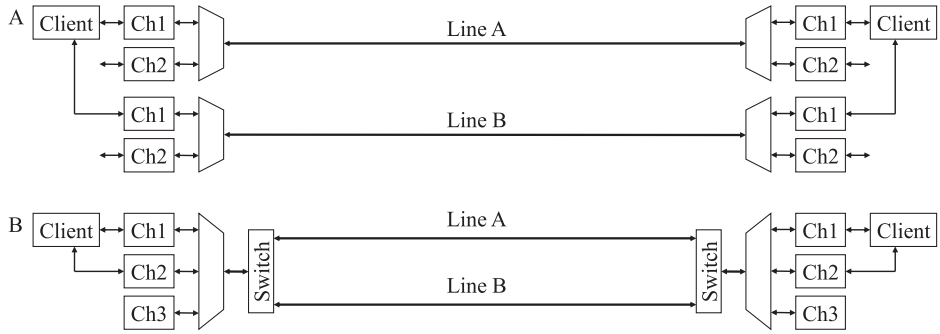


Fig. 56 Redundancy in data-center interconnects. A: Path protection, B: line protection.

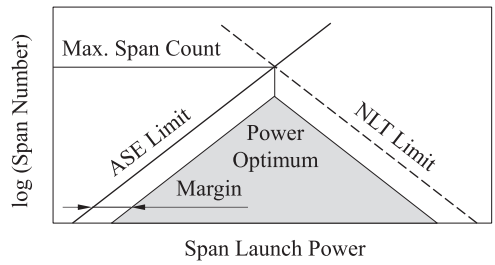


Fig. 57 The "working zone" of optical-layer design.

Since simulation of nonlinear fiber transmission causes immense calculation effort, there is the need for engineering rules enabling quick first-order performance estimates and identification of limiting effects in multi-path meshed networks. Each potential channel has to be in the power range above noise limitation and below severe nonlinearity accumulation. This limits the valid power-level range, see **Fig. 57**.

One relevant system limitation is given by the *Nonlinear Tolerance* (NLT). The NLT law $\log(N_{\text{Span}}) + P_{\text{Launch}}$ (dBm) = NLT (dBm) = const holds for different fiber types, data rates, modulation formats, and channel spacings (Mohs *et al.*, 2000). To achieve high transmission distances, the per-span launch power has to be kept low. This corresponds to shorter amplifier spans, leading to better noise figures at the cost of more amplifiers. Typical resulting reach is shown for (mixed) 10G/100G signals in **Fig. 58**.

For **Fig. 58**, fiber loss of 0.25 dB/km and a patch-panel and repair margin of 1 dB each were assumed for every span. Shorter spans improve total reach to a certain point only. Beyond this point, accumulated amplifier tilt and ripple lead to negative effects.

The loss of any components along a given WDM link is not constant. Spectral fluctuations induced by random loss non-uniformities cannot be fully compensated by optical amplification and accumulate as power ripples. These ripples arise from filter-loss and amplifier-gain non-uniformities, combined with transmitter launch-power variations. This leads to typical channel-power spread as shown in **Fig. 59**.

High-power channels are subject to stronger signal distortion by nonlinear effects, whereas the BER of the lowest-power channels is impacted by the low OSNR and/or low power at the receiver. *Power management* aims at controlling the WDM channel power levels at particular points in the networks to ensure optimum reach for all channels. It can be performed in

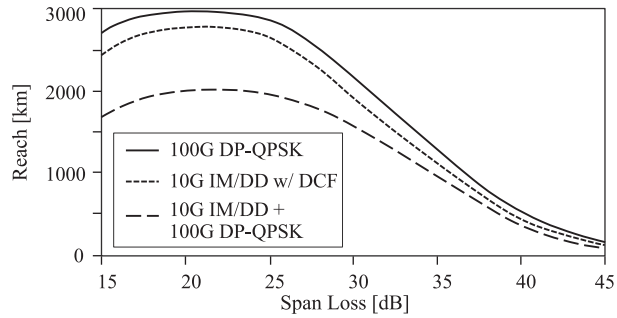


Fig. 58 Transparent reach for different span losses.

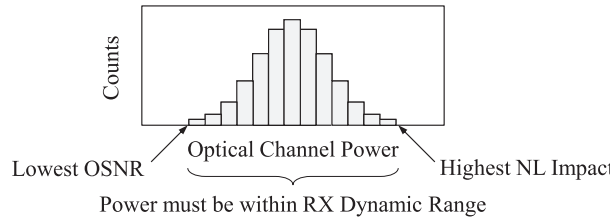


Fig. 59 Histogram of WDM channel power.

ROADMs which are located after every N spans, with $N = 5–10$ for terrestrial networks. This is typically supported by a distributed control plane.

The minimum OSNR calculated for each link including all OSNR penalties must be compared with the OSNR required by any of the receivers used. The system limit is reached if the OSNR of the worst link, including all penalties, is lower than the OSNR required at the receiver. After that, the optical signal must be electrically regenerated. This system limit was already shown in Fig. 38, where OSNR penalties (in dB) are assumed to grow linearly with the span count.

Mixed 10 Gb/s/100 Gb/s Design

In certain WDM networks, combined 10-Gb/s IM/DD and coherent 100 Gb/s transport is required. This causes potential problems:

- Coherent 100 Gb/s do not require optical CD compensation, whereas most 10-Gb/s IM/DD systems do
- Coherent 100 Gb/s systems can be subject to reach penalties when DCF are used. This results from increased XPM inside the DCF, and the fact that DCF compensate the channel walk-off (leading to stronger XPM in the transmission fibers).
- Phase-modulated signals suffer penalties caused by IM/DD signals. These are generated by XPM-induced Nonlinear Phase Noise (X-NLPN).

The DCF-related penalty can partly be reduced by carefully adjusting the DCF launch power. Further methods for reducing the DCF-induced XPM penalty are:

- Separate 10-Gb/s and 100 Gb/s transport on dedicated photonic layers (coherent overlay)
- Use CD compensators other than DCF, i.e., channelized FBGs

Reduction mechanisms for IM/DD-induced X-NLPN include:

- Reducing power of 10 Gb/s channels compared to 100 Gb/s channels.
- Using channelized FBGs instead of DCFs to allow walk-off between 10 Gb/s and 100 Gb/s channels.
- Increasing the spectral separation between 10 Gb/s and 100 Gb/s channels (guard band).

The last two methods decorrelate the 10 Gb/s from the 100 Gb/s signals along their paths.

Another method of partial nonlinearity suppression is based on using IM/DD with inbuilt pre-equalization. It avoids DCF and the related XPM, but does not avoid X-NLPN.

In addition, the Carrier Phase Estimation (CPE) in coherent receivers can be configured to achieve optimum results in the presence of nonlinearity.

For all methods, certain limitations or costs result. This is summarized in Table 7.

Beyond 100 Gb/s, further future functionality should be considered, e.g., super-channels for coherent 400-Gb/s and 1-Tb/s systems. For optimum performance, these systems require coherent-only, uncompensated photonic layers. This excludes both, coexistence scenarios based on optical CD compensation, and the general use of any IM (DD) technology.

Table 7 Cost of 10-Gb/s and 100-Gb/s coexistence

Method	*Generalized cost*	*Potential advantages*
Coherent overlay	Almost doubles cost of photonic layers, especially fibers.	Only safe solution for 400+ Gb/s super-channels
IM/DD with inbuilt EDC	10-Gb/s EDC does not reduce X-NLPN, and hence decreases 100-Gb/s reach.	Avoids compensated photonic layer and DCF-incuced penalties (XPM)
Use of FBGs instead of DCFs	Phase-ripple penalty to IM/DD channels (~1 dB penalty after 8–10 devices). *Channalized filters contradict FlexGrid*	Significantly reduces X-NLPN from IM/DD. Eliminates XPM
Guard bands between 10-Gb/s and 100-Gb/s channels	Decrease of system and network capacity, added network-planning complexity.	Can achieve almost optimum performance for IM/DD and QPSK.After gradual growth to all-100-Gb/s, the guard band can be eliminated.
Power-level reduction for IM/DD	Reduces IM/DD performance. Requires CD compensation for IM/DD (unless IM/DD uses inbuilt EDC).	Can lead to less or negligible IM/DD-induced penalty for QPSK.

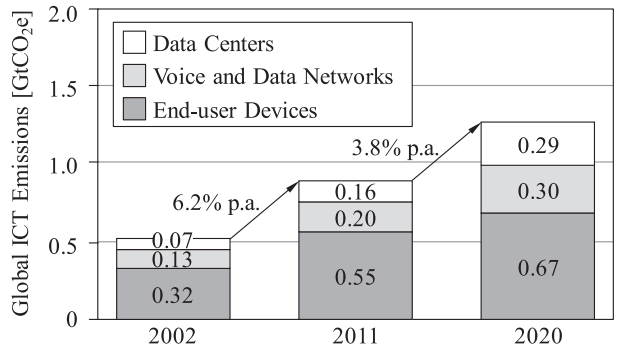


Fig. 60 Global ICT carbon footprint. Reproduced from Mohs, G., *et al.*, 2000. Maximum link length versus data rate for SPM limited systems. In: Proceedings European Conference on Optical Communication (ECOC) 2000, Munich.

WDM Sustainability

In the recent years, consideration of the environmental impact of the global Information and Communications Technologies (ICT) sector has become a major requirement by large network operators and legislation.

Global ICT produces a significant amount of the global CO_2 footprint. In 2007, the ICT portion of global CO_2 production was in the range of 2% (GeSI, 2020, 2012). This amount is projected to significantly increase over the next years, see **Fig. 60**.

From the global ICT CO_2 footprint, telecommunications infrastructure is in the range of ~25%. WDM transport is responsible for ~7% of network power consumption. Nonetheless, increased energy efficiency is relevant due to the exponentially increasing bandwidths.

The analysis of various environmental impact parameters of any systems is known as Life-Cycle Assessment (LCA). LCA according to ISO14040/14044 considers all relevant phases of systems or products, from extraction of raw materials via manufacture, distribution, the use phase to end of life (the latter preferably meaning recycling or partial reuse). Various environmental-impact parameters like Global Warming Potential (which is basically coupled to CO_2 footprint) can be derived. This requires detailed knowledge of the respective parameters of any components used, plus knowledge of the contributions from any logistics etc.

Fig. 61 shows an LCA of a WDM system in a simple point-to-point pre-amplified configuration. The respective system supports up to 80 channels and uses 10-Gb/s channel cards. Other WDM configurations (rings, 100-Gb/s cards,…) show similar results.

It can be seen that various environmental-impact parameters, in particular GWP and ODP, are dominated by the use phase. This is driven by energy consumption, assuming electricity mix of the year 2016. In this LCA, a use phase of 8 years has been assumed. The dominance of the use phase in LCA underpins the necessity of highest WDM energy efficiency.

Energy efficiency is ranked in so-called Telecommunications Energy-Efficiency Ratings (TEER). Definitions for WDM can be found in the ANSI Standard ATIS-0600015.02.2009 (ANSI/SCTE, 2009) and the Ecology Guidelines for the ICT Industry (ECO ICT 2017). InECO ICT (2017), the relevant metric (Figure of Merrit, FoM) is given as maximum throughput (in Gb/s) divided by average power consumption (in Watts). Average power consumption is defined as linear average of power consumption of a

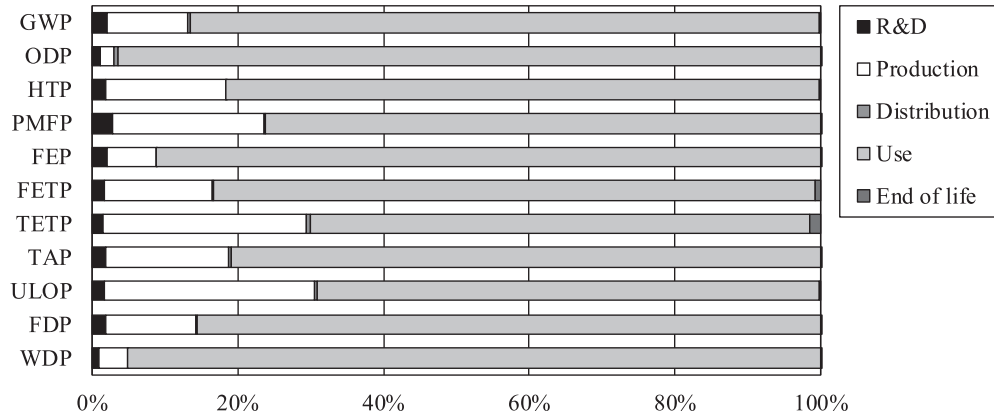


Fig. 61 LCA of a WDM system with 8 years of use phase. The diagram shows the potentials for Global Warming (GWP), Ozone Depletion (ODP), Human Toxicity (HTP), Particulate Matter Formation (PMFP), Freshwater Eutrophication (FEP), Freshwater EcoToxicity (FETP), Terrestrial EcoToxicity (TETP), Terrestrial Acidification (TAP), Urban Land Occupation (ULOP), Fossil Depletion (FDP) and Water Depletion (WDP), respectively.

Table 8 TEEER for WDM system FSP 3000 according to ECO ICT (2017).

ECO ICT Council TEER		*Ref. –Config. 2 DWDM w/ROADM, 80× 100G*					
		100G old (2013)		*100G new (2015)*		*400G (2016)*	
PC @ 1 Wavelength (W)		1296		725		597	
PC @ full Wavelengths (W)		13,863		8118		6719	
Max. Throughput (Gbps)		8000		8000		8000	
N.R. (Gbps/W)	**PC @ N.R. (W)**	**0.86**	**9302.32**	**0.86**	**9302.32**	**0.86**	**9302.32**
★★		0.96	8372.09	0.96	8372.09	0.96	8372.09
★★★		1.08	7441.86	1.08	7441.86	1.08	7441.86
★★★★		1.23	6511.62	1.23	6511.62	1.23	6511.62
★★★★★		1.43	5581.39	1.43	5581.39	1.43	5581.39
FoM (Gbps/W)	**Avg. PC (W)**	**1.06**	**7579**	**1.80**	**4435**	**2.19**	**3658**
Result		★★		★★★★★		★★★★★	

Source: Reproduced from ECO ICT, 2017. Ecology guideline for the ICT industry (Version 7.1), ICT Ecology Guideline Council. Available at: tca.or.jp/information/pdf/ecoguideline/guideline_eng_7_1.pdf.

system equipped with one wavelength and power consumption of a fully loaded system. Similar definitions are given in ANSI/SCTE (2009).

For relevant WDM configurations, Normative References (N.R.) have been defined which have to be matched by the respective WDM systems. Systems perform the better the clearer they exceed the N.R.

Table 8 lists the Figures of Merit (FoM) for a specific WDM transport system, the Fiber Service Platform 3000 (ADVA, 2016). The configuration analyzed here (Reference Configuration 2 of ECO ITC (2017)) is an 80-channel DWDM system with coherent 100-Gb/s OTN transponders and Degree-4 colorless, directionless ROADMs. Three generations of transport cards are analyzed, showing a strong increase of energy efficiency over time.

In order to further increase WDM energy efficiency, every layer of WDM transport systems and networks must be considered and optimized. This includes:

- Components (power supplies, ASICs, ADC/DAC,…).
- Thermal design (fans, also: Heating, Ventilation, Air Conditioning (HVAC)).
- Low-energy modes (partial deactivation, sleep modes,…).
- Networking aspects like grooming, flexibility, reconfigurability.

Even combination of these aspects may not be sufficient to keep WDM power consumption stable with exponentially increasing bit rates. **Fig. 62** displays power consumption of different WDM generations with channel bit rates of 2.5, 10, 40, 100 and 400 Gb/s, respectively. It can be seen that efficiency massively increased over time, now approaching 0.1 W/(Gb/s). However, absolute consumption also slightly increased.

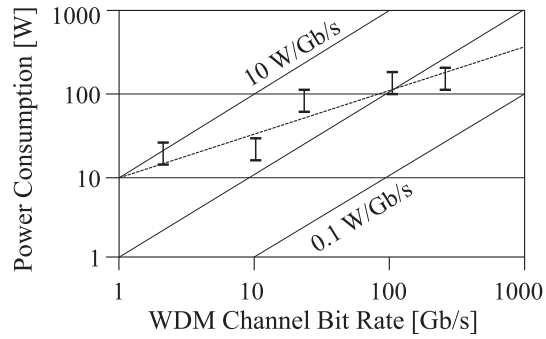


Fig. 62 Development of WDM power consumption over time, with increasing bit rates.

High savings in CO_2 footprint can be enabled by ICT, and especially telecommunications, outside the ICT sector. Examples include reduction of travel and truck rolls, dematerialization, control of energy grids, and many more. The telecommunications-related CO_2 abatement for 2020 is estimated to be five times higher as the ICT contribution to CO_2 production (GeSI, 2020, 2012).

Overview on the Optical Transport Network

OTN Layers

The Optical Transport Network (OTN) is specified in several ITU-T Recommendations (G.872 on architecture (ITU-T, 2017), G.709 on frames and formats (ITU-T, 2009a), G.798 on atomic functions and processes (ITU-T, 2012a)). It is sometimes also called Optical Transport Hierarchy (OTH). It combines TDM and WDM into a common transport system.

The TDM part is hierarchically structured, with Optical Channels (OCh) forming its basis. They are structured in hierarchical levels called Optical Channel Payload Unit (OPU), Optical Channel Data Unit (ODU) and Optical Channel Transport Unit (OTU).

The ODU signals are the end-to-end networking entities. Eight ODU signals have been defined in G.709. They address different bit rates and mapping schemes. The nominal ODU*k* rates are approximately 1.25 Gb/s (ODU0), 2.50 Gb/s (ODU1), 10.04 Gb/s (ODU2), 40.32 Gb/s (ODU3), and 104.79 Gb/s (ODU4). For 10 GbE LAN PHY services, an overclocked ODU2e with 10.40 Gb/s has been defined. In addition, two ODUflex signals for Constant Bit-Rate (CBR) clients and clients which are mapped via the Generic Framing Procedure (GFP, according to ITU-T Recommendation G.7041 (ITU-T, 2016c)) have been specified. Their nominal bit rates are $239/238 \times$ CBR client bit rate and $n \times 1.25$ Gb/s (with integer n), respectively. Jitter tolerances are ± 20 ppm for most ODUs, and ± 100 ppm for ODU2e and ODUflex for CBR services.

The layered structure of OTN is shown in **Fig. 63**.

The OTN point-to-point single-wavelength frame structure is called Optical channel Transport Unit (OTU). The nominal OTU*k* rates are approximately 2.67 Gb/s (OTU1), 10.71 Gb/s (OTU2), 43.02 Gb/s (OTU3), and 111.81 Gb/s (OTU4). The OTU*k* frame structure always contains 4×4080 bytes. Therefore, the frame duration is not constant across hierarchy levels.

Multiplexing of several OCh by means of WDM creates the Optical Multiplex Section (OMS) and the Optical Transport Section (OTS). The OMS and OTS can be accompanied by an Optical Supervisory Channel (OSC, i.e., a dedicated wavelength carrying supervision and management information only), as indicated in **Fig. 63**.

The OMS refers to sections between optical multiplexer and demultiplexer, and the OTS to sections between optical amplifiers, respectively (Gorshe, 2010). This is shown in **Fig. 64**. Here, two OTN interfaces, Inter-Division Interface (IrDI) and Intra-Division Interface (IaDI), are shown. These interfaces are defined in ITU-T Recommendation G.872. They allow interworking between (IrDI) and within (IaDI) network domains. The IrDI interfaces are defined with full regeneration at each end of the interface.

OTN Mapping and Multiplexing

All relevant client signals can be mapped efficiently into OTN. Via OTN multiplexing, they can also be (time-domain) multiplexed onto high-bit-rate wavelengths. The resulting mapping options are summarized in **Fig. 65**.

SONET/SDH client signals can directly be mapped into OTN OPUs (which perform the client adaptation). Although OTN does not require synchronization, it can support the synchronization requirements of SONET/SDH and of synchronous Ethernet.

For the different Layer-2 signals, different mapping mechanisms can be used. One option is GFP. It allows mapping of various Layer-2 signals into SONET/SDH or ODU frames. The client signals can be Protocol Data Unit-oriented (like IP/PPP or Ethernet MAC) or block-code-oriented CBR streams such as Fiber Channel (FC).

For 10 GbE, two physical implementations have been defined in IEEE 802.3ae (IEEE, 2002). The WAN PHY operates at 9.954 Gb/s and is bit-rate-compatible with SONET OC-192 and SDH STM-64. It can be mapped into ODU2. The LAN PHY, which

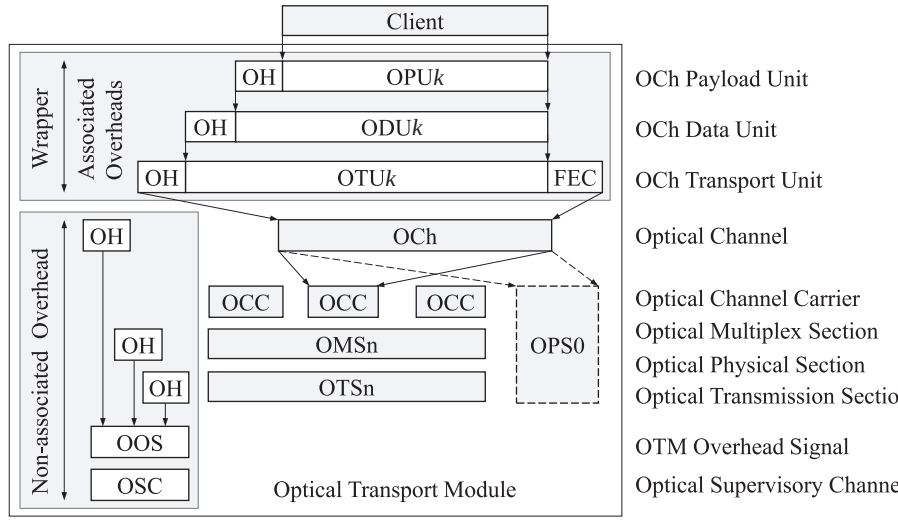


Fig. 63 OTN layering and monitoring. OH: Overhead.

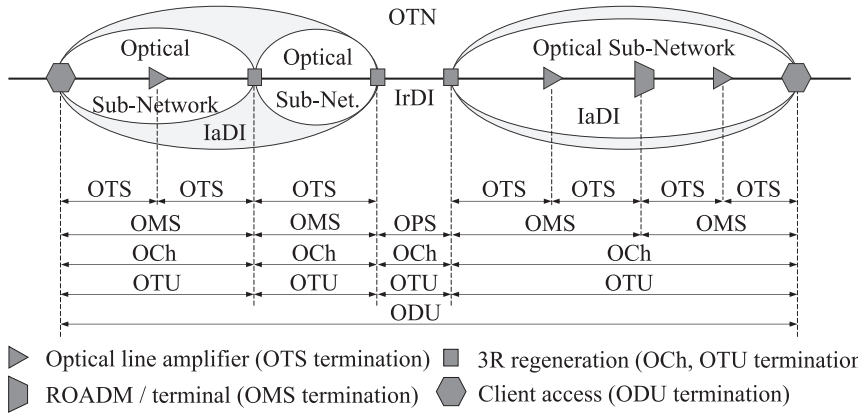


Fig. 64 OTN network layers and interfaces.

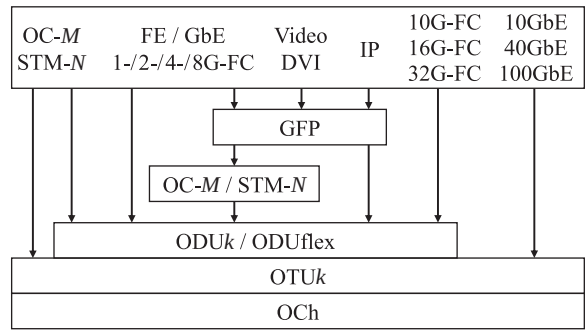


Fig. 65 OTN mapping options.

represents the majority of 10 GbE interfaces, operates at 10.3125 Gb/s. Direct mapping is possible into the overclocked ODU2e. ODU2e can be transported via ODU3 and ODU4, or directly on a wavelength with overclocked OTU2.

40 GbE and 100 GbE can directly be mapped into OTN. For Fiber Channel and other CBR services, the Generic Mapping Procedure (GMP) as defined in ITU-T G.709 can be used for mapping into ODUflex (CBR type), as an alternative to GFP.

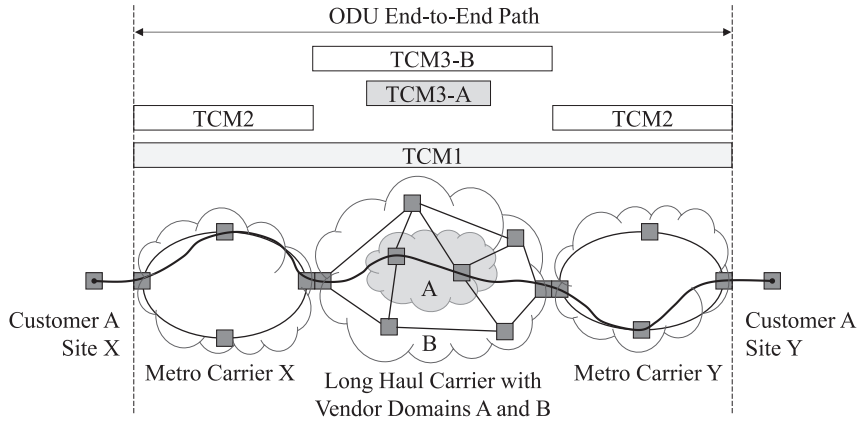


Fig. 66 OTN Tandem Connection Monitoring example.

OTN Operations and Monitoring Aspects

An important OTN feature is *Tandem Connection Monitoring* (TCM). TCM allows end-to-end monitoring across different administrative domains (i.e., network, operator or vendor domains). 6 independent TCM levels have been defined, allowing monitoring for nested and cascaded domains. **Fig. 66** shows a transmission example.

Two metro carriers and one long haul-carrier with two nested vendor domains provide the end-to-end connection. Every network domain and also the end-to-end connection (TCM1) can be monitored simultaneously.

Conclusion

Wavelength Division Multiplexing is a multiplexing and multiple-access technology, used in fiber-optic transmission in order to maximize transmitted bit rates. Its earliest beginnings, in the form of groundbreaking work on optical fibers and semiconductor lasers, are roughly 50 years old. Nonetheless, without WDM, the Internet as it is known today would not exist. Practically all long-haul transmission of data, video and voice is based on WDM. According to the Cisco Virtual Network Index (Cisco, 2017), global Internet traffic is approaching some 100 Tb/s around 2020. Today, latest commercial WDM systems have transport capacity, over up to 1000 km reach, of 20 Tb/s. Since this capacity exceeds the one of any other transmission channel by orders of magnitude, there are no alternatives to long-haul WDM.

In addition, backhaul of any end-customer access technology – including mobile – is increasingly based on WDM, and even in direct fiber-optic access, WDM has started to be used. With copper-based wireline access diminishing, the remaining Internet areas without WDM are dedicated-fiber, non-WDM ("gray") applications, the last mile in mobile networks and some microwave and satellite communications. Here, it is worth noting that even in free-space optical communications, WDM started to be used.

In the field of optics, WDM is strongly related to *modern optics* and *photonics*. Relevant areas are quantum optics and optoelectronics, but WDM also covers aspects of optomechanics and electro-optics. These areas also cover semiclassical approaches, including wave optics. The latter is obviously used for describing field propagation in fibers. Different semiclassical approaches are also used for describing optical-amplifier noise figures. As such, WDM is also one of the relevant applications – and success stories – of modern optics.

References

ADVA, 2016. ADVA optical networking SE: Fiber service platform 3000R7 module and system specification, Release 16.1, Issue A (5/3/2016).
ANSI/SCTE, 2009. Energy efficiency for telecommunication equipment: Methodology for measurement and reporting transport requirements, American National Standard for Telecommunications, ATIS-060015.02.2009.
Agrawal, G.P., 1992. Fiber-Optic Communication Systems. New York: John Wiley & Sons.
Agrawal, G.P., 1995. Nonlinear Fiber Optics, second ed. San Diego: Academic Press.
Agrawal, G.P., Dutta, N.K., 1986. Long-Wavelength Semiconductor Lasers. New York: Van Nostrand Reinhold.
Alexander, S.B., 1997. Optical Communication Receiver Design, vol. TT22. Bellingham: SPIE, (ISBN 0-8194-2023-9).
Azadeh, M., 2009. Fiber Optics Engineering. Dordrecht Heidelberg: Springer, (ISBN 978-1-4419-0304-4).
Bosco, G., Carena, A., Curri, V., *et al.*, 2010. Performance limits of nyquist-WDM and CO-OFDM in high-speed PM-QPSK systems. IEEE Photonics Technology Letters 22 (15), 1129–1131.
Bristiel, B., Jiang, S., Gallion, P., *et al.*, 2006. New model of noise figure and RIN transfer in fiber raman amplifiers. IEEE Photonics Technology Letters 18 (8), 980–982.
Bülow, H., 1998. System outage probability due to first- and second-order PMD. IEEE Photonics Technology Letters 10 (5),
Chraplyvy, A.R., 1990. Limitations on lightwave communications imposed by optical-fiber nonlinearities, invited. IEEE Journal of Lightwave Technology 8 (10), 1548–1557.
Cisco, 2017. The zettabyte era: Trends and analysis, Cisco White Paper. Available at: www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html.

Coldren, L.A., Fish, G.A., Akulova, Y., *et al.*, 2004. Tunable semiconductor lasers: A tutorial. IEEE Journal of Lightwave Technology 22 (1), 193–202.

Corning, 1998. Corning® SMF-LSTM CPC6 Single-mode non-zero dispersion-shifted optical fiber, corning incorporated, product information PI1050, Issued 9/1998.

Corning, 2014. LEAF® Optical Fiber, Data Sheet, Corning Incorporated. Available at: www.corning.com/media/worldwide/coc/documents/Fiber/LEAF%20optical%20fiber.pdf.

DeCusatis, C., 2008. Handbook of Fiber Optic Data Communication, third ed. Burlington, MA: Elsevier Academic Press, (ISBN 978-0-12-374216-2).

Desurvire, E., 1994. Erbium-Doped Fiber Amplifiers. New York: John Wiley & Sons.

Dragone, C., 2005. Low-loss wavelength routers for WDM optical networks and high-capacity IP routers. IEEE Journal of Lightwave Technology 23 (1), 66–79.

ECO ICT, 2017. Ecology guideline for the ICT industry (Version 7.1), ICT Ecology Guideline Council. Available at: tca.or.jp/information/pdf/ecoguideline/guideline_eng_7_1.pdf.

Essiambre, R.-J., Kramer, G., Winzer, P.J., *et al.*, 2010. Capacity limits of optical fiber networks. IEEE Journal of Lightwave Technology 28 (4), 662–701.

Fishman, D.A., Nagel, J.A., 1993. Degradations due to stimulated brillouin scattering in multigigabit intensity-modulated fiber-optic systems. IEEE Journal of Lightwave Technology 11 (11), 1721–1728.

Friis, H.T., 1944. Noise figures of radio receivers. Proceedings of the Institute of Radio Engineers 32 (12), 419–422.

GeSI, 2012. GeSI SMARTer 2020: The role of ICT in driving a sustainable future, GeSI and BCG. Available at: gesi.org/assets/js/lib/tinymce/jscripts/tiny_mce/plugins/ajaxfilemanager/uploaded/SMARTer%202020%20%20The%20Role%20of%20ICT%20in%20Driving%20a%20Sustainable%20Future%20-%20December%202012.pdf.

Gorshe, S., 2010. A tutorial on ITU-T G.709 optical transport networks (OTN), technology White Paper, PMC-Sierra, Inc., PMC-2081250, 1.

Grave de Peralta, L., Bernussi, A.A., Gorbounov, V., *et al.*, 2004. Temperature-insensitive reflective arrayed-waveguide grating multiplexers. IEEE Photonics Technology Letters 16 (3), 831–833.

Grobe, K., Eiselt, M.H., 2014. Wavelength Division Multiplexing: A Practical Engineering Guide. Hoboken, New Jersey: John Wiley & Sons, (ISBN 978-0-470-62302-2).

Hirano, A., 2002. Optical amplifiers and their standardization in ITU-T & IEC, ITU-T Workshop on IP/Optical. Available at: www.itu.int/itudoc/itu-t/workshop/optical/s8-p02r.html.

IEEE, 2002. IEEE Std 802.3ae™-2002 (Amendment to IEEE Std 802.3-2002), Media access control (MAC) parameters, physical layers, and management parameters for 10 Gb/s operation, August.

ITU-T, 2009a. Interfaces for the optical transport network, Recommendation ITU-T G.709.

ITU-T, 2009b. Characteristics of a non-zero dispersion-shifted single-mode optical fibre and cable, Recommendation ITU-T G.655, 11/2009.

ITU-T, 2010a. Characteristics of a fibre and cable with non-zero dispersion for wideband optical transport, Recommendation ITU-T G.656, 07/2010.

ITU-T, 2010b. Characteristics of a dispersion-shifted, single-mode optical fibre and cable, Recommendation ITU-T G.653, 07/2010.

ITU-T, 2012a. Characteristics of optical transport network hierarchy equipment functional blocks, Recommendation ITU-T G.798, 12/2012.

ITU-T, 2012b. Optical system design and engineering considerations, Recommendation ITU-T Series G, Supplement 39, 09/2012.

ITU-T, 2015. Optical interfaces for coarse wavelength division multiplexing applications, Recommendation ITU-T G.695.

ITU-T, 2016a. Characteristics of a bending-loss insensitive single-mode optical fibre and cable, Recommendation ITU-T G.657, 11/2016.

ITU-T,2016b. Characteristics of a cut-off shifted single-mode optical fibre and cable, Recommendation ITU-T G.654, 11/2016.

ITU-T, 2016c. Generic framing procedure, Recommendation ITU-T G.7041/Y.1303, 08/2016.

ITU-T, 2016d. Characteristics of a single-mode optical fibre and cable, Recommendation ITU-T G.652, 11/2016.

ITU-T, 2017. Architecture of optical transport networks, Recommendation ITU-T G.872, 01/2017.

Johnson, C.R., Schniter, P., Endres, T.J., *et al.*, 1998. Blind equalization using the constant modulus criterion: A review, invited paper. Proceeings of the IEEE 86 (10), 1927–1950.

Kaminow, I.P., Koch, T.L., 1997. Optical fiber telecommunications III. San Diego: Academic Press, (ISBN 0-12-395170-4).

Kobayashi, H., Tang, D.T., 1971. On decoding of correlative level coding systems with ambiguity zone detection. IEEE Transactions on Communication Technology COM-19 (4), 467–477.

Lima, M.J.N., Nogueira, R.N., Silva, J.C.C., *et al.*, 2005. Comparison of the temperature dependence of different types of bragg gratngs. Microwave and Optical Technology Letters 45 (4), 305–307.

Linke, R.A., Gnauck, A.H., 1988. High-capacity coherent lightwave systems. IEEE Journal of Lightwave Technology 6 (11), 1750–1769.

Loudon, R., 1985. Theory of noise accumulation in linear optical-amplifier chains. IEEE Journal of Quantum Electronics QE-21 (7), 766–773.

Lyubomirsky, I., 2006. Coherent detection for optical duobinary communication systems. Photonics Technology Letters 18 (7), 868–870.

Lyubomirsky, I., 2010. Quadrature duobinary for high-spectral efficiency 100G transmission. IEEE Journal of Lightwave Technology 28 (1), 91–96.

Maier, G., Pattavina, A., De Patre, S., *et al.*, 2002. Optical network survivability: Protection techniques in the WDM layer. Photonic Network Communications 4 (3/4), 251–269.

Mohs, G., *et al.*, 2000. Maximum link length versus data rate for SPM limited systems. In: Proceedings European Conference on Optical Communication (ECOC) 2000, Munich.

Ohsono, K., *et al.*, 2003. High performance optical fibers for next generation transmission systems. Hitachi Cable Review. 22),

OFS, 2017. TrueWave® RS Optical Fiber, Data Sheet, Doc ID: fiber-120, OFS Fitel, LLC, 06/2017. Available at: fiber-optic-catalog.ofsoptics.com/Asset/TrueWaveRSLWP-120-web.pdf.

Premaratne, M., 2004. Analytical characterization of optical power and noise figure of forward pumped Raman amplifiers. Optics Express 12 (18), 4235–4245.

Prysmian Group, 2010. TeraLightTM ultra optical fiber, draka communications, data sheet product type G.655.E, Issue Date 08/2010. Available at: www.prysmiangroup.com/sites/default/files/business_markets/markets/downloads/datasheets/SMF—TeraLight-Ultra-Optical-Fiber.pdf.

Roberts, P.J., Couny, F., Sabert, H., *et al.*, 2005. Ultimate low loss of hollow-core photonic crystal fibers. Optics Express 13 (1), 236–244.

Russell, P.S.J., 2006. Photonic crystal fibers. IEEE Journal of Lightwave Technology 24 (12), 4729–4749.

Sakurai, Y., Khan, S., Takamure, H., *et al.*, 2012. LCOS-based gridless wavelength blocker array for broadband signals at 100Gbps and beyond, OFC2012, L.A., paper OTh3D.

Savory, S.J., Mikhailov, V., Killey, R.I., *et al.*, 2007. Digital coherent receivers for uncompensated 42.8Gb/s transmission over high PMD fibre, ECOC2007, Berlin, September.

Schetzen, M., 1980. The Volterra and Wiener Theories of Nonlinear Systems. New York: John Wiley & Sons.

Senio, J.M., Jamro, M.Y., 2009. Optical Fiber Communications: principles and Practice, 3rd ed. Harlow: Pearson Education Limited.

Smith, D.R., 2003. Digital Transmission Systems, 3rd ed. Norwell, MA, USA: Kluwer Academic Publishers, (ISBN 1-4020-7587-1).

Tonguz, O.K., Wagner, R.E., 1991. Equivalence between preamplified direct detection and heterodyne receivers. IEEE Photonics Technology Letters 3 (9), 835–837.

Winzer, P.J., Essiambre, R.-J., 2006. Advanced modulation formats for high-capacity optical transport networks, invited paper. IEEE Journal of Lightwave Technology 24 (12), 4711–4728.

Coherent Lightwave Systems

Michael J Connelly, University of Limerick, Limerick, Ireland

Introduction

Most optical communication systems (typically operating in the 1300 nm or 1550 μm optical fiber communication bands) are based on intensity modulation (IM) of a carrier lightwave by an electrical data signal and Direct Detection (DD) of the received light. The simplest IM-DD schemes employ On/Off Keying (OOK) whereby turning on or off a carrier lightwave transmits a binary '0' or '1'. The lightwave is transmitted via an optical fiber and at the receiver detected by a photodetector. The resulting photocurrent is then processed to determine if a '0' or '1' was received. Photodetection is insensitive to the phase and polarization state of the received lightwave. In the ideal case, assuming a monochromatic carrier and noiseless receiver, the Bit Error Rate (BER) is related to the quantum noise of the received light; the number of received photons/bit required (sensitivity) to achieve a given BER is termed the quantum-limit; for example for a BER of 10^{-9} the quantum limit is 10 photons/bit. In practice the sensitivity of DD receivers is 10–30 dB less than the quantum limit, because of the relatively much higher receiver noise. The use of a high-gain optical preamplifier can greatly improve the sensitivity.

The capacity of IM-DD links can be greatly increased by the use of wavelength division multiplexing (WDM). Typical commercial WDM-IM-DD systems utilize an optical channel spacing of 50 GHz for a 40 Gb/s channel data rate. Erbium Doped Fiber Amplifiers (EDFAs) are used to compensate for transmission link losses (fiber attenuation and other losses such as splitting losses). The usable bandwidth of an EDFA is typically 30 nm, which is equivalent to 3.77 THz if centered at 1550 nm. The number of 50 GHz WDM channels than can fit into the EDFA bandwidth is 75 leading to a gross data rate of approximately 3 Tb/s. Higher gross data rates are achievable by using more complicated amplifier configurations to extend the usable bandwidth. The spectral efficiency of OOK modulation is 1 bit/s/Hz/channel.

Increased gross data rates can be achieved by employing spectrally efficient high-order modulation formats employing phase and amplitude modulation of the carrier lightwave (Nakazawa *et al.*, 2010; Seimetz, 2010; Winzer, 2012). A further advantage of high-order formats is that the symbol rate (R_s) corresponding to a specific data rate (R_b) is reduced thereby relaxing the bandwidth requirements of the transmitter and receiver components. System capacity can be further increased by Dual Polarization (DP) transmission, in which the two orthogonal polarizations of the carrier lightwave are separately modulated and recombined. Low-order phase modulated signals can be detected using differential detection receivers based on delay interferometers and DD. Such receivers have been employed in commercial systems but are not suitable for detection of more complex modulation formats. Detection of a phase/amplitude modulated lightwave necessitates the use of linear receivers that preserve the lightwave amplitude and phase. Linear receivers use coherent detection, whereby the received modulated lightwave is combined with a Local Oscillator (LO) laser and simultaneously detected; the resulting electrical signals are then processed to recover the received lightwave phase and amplitude. Because coherent receivers are linear, post-reception Digital Signal Processing (DSP) can be used to compensate for channel and receiver impairments and implement polarization demultiplexing, decoding and error correction. The combination of coherent optical reception and DSP is called a Digital Coherent Receiver (DCR) and forms the basis for all coherent optical receivers. This article gives an overview of optical modulation, high-order modulation formats, differential and coherent detection, DCRs and current developments in the field.

Optical Modulation

A monochromatic lightwave field can be expressed as

$$e_{CW}(t) = \sqrt{2P_s}\cos(\omega_s t) \tag{1}$$

where P_s is the power (averaged over an optical period) and ω_s is the angular optical frequency. An arbitrary modulation of the lightwave can be expressed as

$$e_s(t) = \sqrt{2P_s}a(t)\cos[\omega_s t + \theta_s(t)] \tag{2}$$

with time dependent amplitude modulation $a(t)$ and phase modulation $\theta_s(t)$. Dividing (2) by $\sqrt{2P_s}$ gives the normalized lightwave field

$$\frac{e_s(t)}{\sqrt{2P_s}} = a(t)\cos[\omega_s t + \theta_s(t)] \tag{3}$$

which can be expanded as

$$a(t)\cos[\omega_s t + \theta_s(t)] = a(t)\{\cos(\omega_s t)\cos[\theta_s(t)] - \sin(\omega_s t)\sin[\theta_s(t)]\} \tag{4}$$

Taking $\cos(\omega_s t)$ as the phase reference, the In-phase (I) and Quadrature (Q) components of (4) are $a(t)\cos[\theta_s(t)]$ and $a(t)\sin[\theta_s(t)]$ respectively. The phasor representation of (4) $A_s(t)=a(t)\exp[j\theta_s(t)]$, as shown in **Fig. 1**, is called the complex amplitude. Coherent lightwave communication systems always use the I and Q components of the carrier.

Modulators

Optical phase modulation can be implemented using an integrated Phase Modulator (PM) or Mach-Zehnder Modulator (MZM) as shown in **Fig. 2**. The PM comprises an optical waveguide embedded in an electro-optic material, usually Lithium Niobate ($LiNiO_3$). Phase modulation $\theta(t)$ of the input lightwave is achieved by applying a drive voltage $v(t)$ between the electrodes. Assuming a lossless device the relationship between the input $E_{in}(t)$ and output $E_{out}(t)$ complex fields is

$$E_{out}(t)=E_{in}(t)\exp[j\theta(t)]=E_{in}(t)\exp\left[j\frac{v(t)}{V_\pi}\right] \tag{5}$$

where V_π is the voltage required to obtain a π phase shift.

A MZM comprises two PMs embedded in a Mach-Zehnder Interferometer. In a dual-drive MZM, shown in **Fig. 2**, the PMs are independently driven. The input lightwave is split into two lightwaves of equal power. The phase shifted lightwaves in each arm of the interferometer are recombined resulting in interference, which depending on the drive voltages, can be varied between destructive and constructive. The modulator transfer function is given by

$$\frac{E_{out}(t)}{E_{in}(t)}=\frac{\exp[j\theta_1(t)]+\exp[j\theta_2(t)]}{2}=\exp\left\{j\frac{[\theta_1(t)+\theta_2(t)]}{2}\right\}\cos\left[\frac{\theta_1(t)-\theta_2(t)}{2}\right] \tag{6}$$

Each PM induces a phase shift $\theta_k(t)=v_k(t)\pi/V_{\pi,k}$, where $k=1,2$ denotes the upper or lower PM respectively. If the PMs are identical then $V_{\pi,k}=V_\pi$. If $v_1(t)=v_2(t)$ (push-push mode) then $\theta_1(t)=\theta_2(t)$ resulting in phase modulation $\theta(t)=\theta_1(t)$, without any accompanying amplitude modulation. If $v_2(t)=-v_1(t)$ (push-pull mode), the phase term in (6) is equal to unity so the modulator acts as an amplitude modulator. The power transfer function (magnitude squared of (6)), is

$$\frac{P_{out}(t)}{P_{in}(t)}=\frac{1}{2}\left\{1+\cos\left[\frac{v(t)}{V_\pi}\pi\right]\right\} \tag{7}$$

where $v(t)=2v_1(t)$ is the MZM drive voltage. Plots of (6) and (7) are shown in **Fig. 3** for push-pull mode operation. The MZM can be operated as an intensity modulator by biasing at the quadrature point $-V_\pi/2$ and driving with a V_π peak-to-peak modulation. By biasing at minimum transmission point $-V_\pi$ and driving with a $2V_\pi$ peak-to-peak modulation the MZM can be operated as a Binary Phase Shift Key (BPSK) modulator since there is a phase jump of π every time the minimum transmission point is crossed.

A major disadvantage of a single MZM is that independent modulation of the I and Q components is not possible, which complicates the drive voltage waveforms required to generate high-order modulation formats. An alternative modulator is based on two push-pull mode MZMs in parallel, one of which is in series with a $\pi/2$ phase shifter. Such an IQ Modulator (IQM), shown in **Fig. 2**, has a transfer function, normalized such that its maximum magnitude is equal to unity, given by

$$\frac{E_{out}(t)}{E_{in}(t)}=\frac{1}{\sqrt{2}}\cos\left[\frac{v_I(t)}{2V_\pi}\pi\right]+j\frac{1}{\sqrt{2}}\cos\left[\frac{v_Q(t)}{2V_\pi}\pi\right] \tag{8}$$

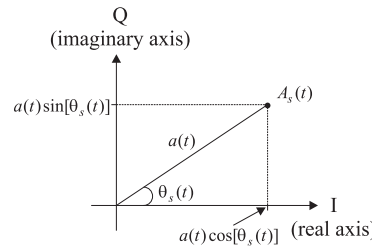


Fig. 1 I-Q representation of the complex amplitude of a modulated lightwave.

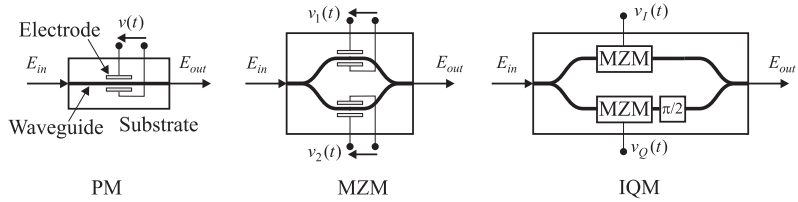


Fig. 2 Integrated PM, dual-drive MZM and IQM.

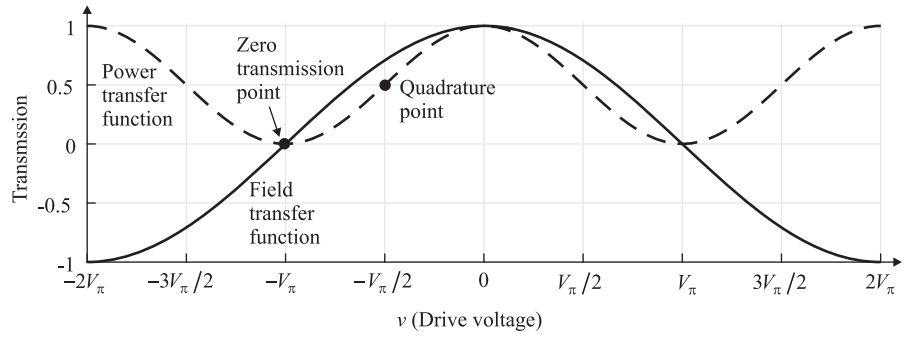


Fig. 3 Push-pull mode MZM transfer functions.

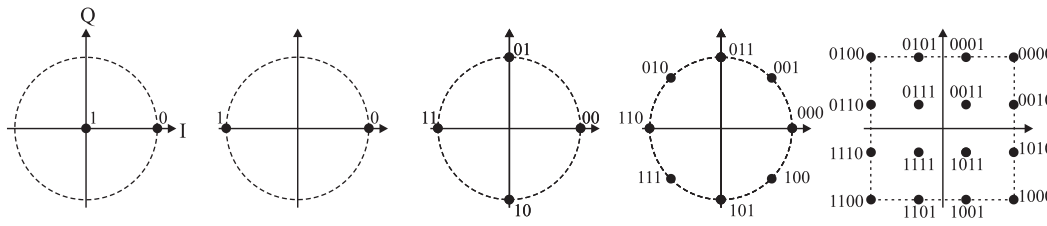


Fig. 4 Constellation diagrams with Gray coded symbols for common modulation formats assuming SP transmission. The symbol rates for SP and PDM transmission corresponding to 54 Gb/s data rate are also shown.

where $v_I(t)$ and $v_Q(t)$ are the I and Q branch MZM drive voltages respectively. The polar form of (8) is

$$\frac{E_{out}(t)}{E_{in}(t)} = a(t)\exp[j\theta(t)] \tag{9}$$

with amplitude modulation

$$a(t) = \frac{1}{\sqrt{2}}\left\{\cos^2\left[\frac{v_I(t)}{2V_\pi}\pi\right] + \cos^2\left[\frac{v_Q(t)}{2V_\pi}\pi\right]\right\}^{1/2} \tag{10}$$

and phase modulation

$$\theta(t) = \arg\left\{\cos\left[\frac{v_I(t)}{2V_\pi}\pi\right] + j\cos\left[\frac{v_Q(t)}{2V_\pi}\pi\right]\right\} \tag{11}$$

Any point within the normalized I-Q space (unit circle) can be reached by choosing suitable values of v_I and v_Q.

High-Order Modulation

In digital optical communications, the complex amplitude of the carrier lightwave can only take on discrete values, whereby b data bits are mapped onto M discrete symbol points located in the I-Q plane (constellation diagram). M is a power of 2 so $b = \log_2 M$. Bit-to-symbol mapping can be chosen arbitrarily; in most systems the Gray code is used, whereby only one bit per symbol differs from a neighboring symbol leading to optimum receiver BER performance. In M-ary PSK (m-PSK), symbols are mapped to M possible values of the carrier phase between 0 and 2π with a step size of $2\pi/M$; the phase values associated with BPSK are $[0, \pi]$; with 4-PSK (Quadrature PSK, QPSK), $[0, \pi/2, \pi, 3\pi/2]$ and with 8-PSK $[0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 7\pi/4, 9\pi/4]$ as shown in Fig. 4 (Gnauck and Winzer, 2005). In M-ary Quadrature Amplitude Modulation (m-QAM), symbols are mapped to one of M combinations of the carrier amplitude and phase. The constellation points can be arranged in a square (square QAM), as shown for 16-QAM in Fig. 4, or on multiple concentric circles (star QAM). 2-QAM and 4-QAM are the same as BPSK and QPSK respectively.

Demodulation of PSK signals using differential detection requires a phase reference, which is not available in DD based receivers. However, demodulation can be achieved if the data is pre-coded such that the information is carried in the phase changes between successive bits; such modulation is called Differential PSK (DPSK). The term DPSK is taken to mean differential binary DPSK. Differential Quadrature DPSK (DQPSK) is the most common differential PSK format used in commercial systems.

Single Polarization (SP) transmission is a 2D format since two degrees of freedom are exploited (I and Q). Two more degrees of freedom can be exploited by using Dual Polarization (DP) transmission, in which the two orthogonal polarizations of the carrier

lightwave are modulated by I-Q data (4D format). The two most common forms of DP transmission are Polarization Division Multiplexed (PDM) and Polarization Switched (PS) systems (Krongold *et al.*, 2012). In PDM, each polarization is independently modulated with half of the symbol bits, thereby halving the symbol rate compared with single-polarization transmission. For example in PDM-QPSK systems, each polarization is modulated with QPSK data so the number of possible symbols M = 16 (Roberts *et al.*, 2009). The principle of PS can be illustrated by considering the most power efficient format of all 4D modulation formats, PS-QPSK which has 8 possible symbols (3 bits/symbol) (Karlsson and Agrell, 2009). Two of the bits define a QPSK symbol, while the remaining bit, corresponding to whether the particular polarization component is turned on or off, indicates which polarization the QPSK symbol is transmitted on as shown in **Fig. 5**. The eight levels of the PS-QPSK format are not possible to Gray code, since each point in the signal constellation has 6 nearest neighbors.

High-order modulation of a carrier lightwave is generally implemented in a number of stages. First the data bits to be transmitted are encoded, typically using Forward Error Correction (FEC), and mapped onto the desired signal constellation (Djordjevic *et al.*, 2009). If required the symbols can be differentially encoded by a pre-coder. The resulting digitally generated I and Q signals are then converted to analog electrical signals by Digital-to-Analog converters (DACs), electrically filtered if necessary, and sent to the IQM modulator I and Q ports inputs via drive amplifiers. Transmitter DSP can also include functions such as Nyquist spectral shaping (to minimize the modulated lightwave bandwidth) and signal pre-distortion (to mitigate transmission impairments).

A specific example of high-order modulation is the Differential QPSK (DQPSK) transmitter, shown in **Fig. 6**. The input data stream is first encoded using FEC and then mapped to QPSK symbols, resulting in I and Q data bit streams u_k and v_k respectively, where the subscript denotes the k-th bit. u_k and v_k are differentially encoded, using a serial pre-coder. Serial pre-coder can be realized for speeds of up to 40 Gb/s; higher speeds require more complex parallel architectures. The pre-coder output bit streams I_k and Q_k are converted to analog Non-Return-to-Zero (NRZ) voltage signals by the DACs. These signals drive the I and Q ports of the IQM, such that its MZMs are operated as BPSK modulators, which results in QPSK modulation of the input lightwave. The constant amplitude output lightwave has a phase $\theta_k = \tan^{-1}[\cos(Q_k\pi)/\cos(I_k\pi)]$ relative to the input lightwave. The modulator output constellation diagram is the same as the QPSK constellation shown in **Fig. 4** but rotated 45° anticlockwise; however constant phase shifts are of no consequence in the demodulation process.

All modulators suffer from additive chirp (time dependent instantaneous frequency of the lightwave carrier frequency), especially at symbol transitions. When an MZM is operated in push-pull mode, the chirp is greatly reduced because the drive voltages are opposite in sign and so the induced chirp caused by each PM are opposite in sign and cancel. When the modulated lightwave is transmitted down an optical fiber, the chirp can lead to an increase in Chromatic Dispersion (CD) and a consequent degradation in system performance. In the MZM modulator, there are also undesirable power dips at the modulation signal transitions. Both of these effects can be greatly reduced by using Return-to-Zero (RZ) transmission. The required RZ pulses can be generated using a pulse carver, either prior to or after modulation. The most common RZ pulse carver uses a biased low-chirp MZM driven by a full or half symbol rate sinusoid. Commonly used RZ pulse streams have full-width at half maximum pulsewidths of 33%, 50% and 67% of the symbol period. The former two pulse streams have spectrums that contain a significant

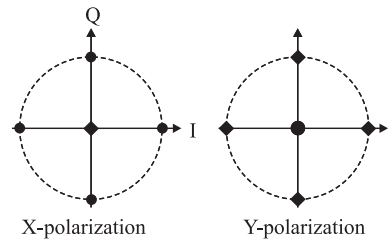


Fig. 5 Constellation diagrams for the two orthogonal polarizations (X and Y) of PS-QPSK. The circles indicate a QPSK symbol is transmitted on the X polarization, while a zero is transmitted on the Y polarization. The opposite is the case with the diamond constellation points. The symbol rate corresponding to 54 Gb/s data rate is 18 Gb/s.

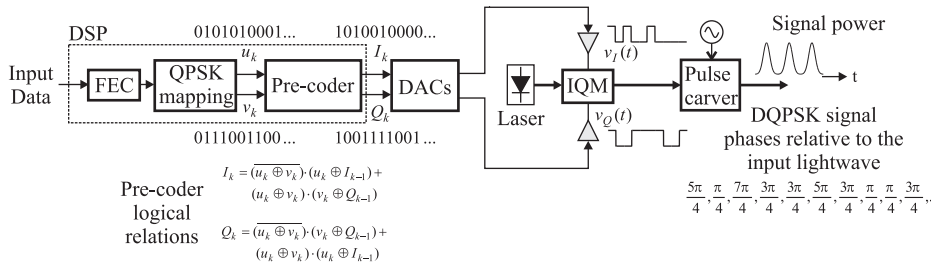


Fig. 6 DQPSK transmitter structure. The serial pre-coder logical relationships $\oplus$, $\cdot$ and $+$ are the XOR, AND and OR operators respectively. Lowpass Filters (LPFs) are often employed after the DACs.

carrier component, which is useful in retrieving the symbol clock at the receiver. The latter pulse stream has an alternating phase and is called Carrier Suppressed RZ (CSRZ), which compared to standard RZ-OOK is more tolerant to filtering and CD because of its narrower spectrum.

PDM-QPSK can be implemented by passing the orthogonal polarization components of the input lightwave through separate IQMs and recombining their outputs using a polarization beam combiner. PS-QPSK can be achieved in a similar way to PDM-QPSK except that the orthogonal outputs from each IQM are passed through on/off intensity modulators, to set the third bit of the symbol to be transmitted, before recombination.

Differential Detection

A differentially encoded PSK lightwave can be detected using differential detection and is especially applicable to DQPSK formats. The DQPSK demodulator shown in **Fig. 7** consists of two parallel Delay Interferometers (DIs). The incoming signal is split into two arms each with two sub-paths. In each arm the upper sub-path is delayed by the symbol period and the lower sub-path phase is shifted by $\pi/4$ or $-\pi/4$. The Balanced Detector (BD) output photocurrents corresponding to the k-th received symbol are given by

$$i_{r,k} = -\frac{RP_s}{2\sqrt{2}}[\cos(\Delta\phi_k) - \sin(\Delta\phi_k)] \tag{12}$$

$$i_{s,k} = -\frac{RP_s}{2\sqrt{2}}[\cos(\Delta\phi_k) + \sin(\Delta\phi_k)] \tag{13}$$

where R and P_s are the photodiode responsivity and the demodulator optical input power respectively and $\Delta\phi_k = \phi_{k-1}$ is the phase difference between the k-th and $(k-1)$-th symbols. The logical values corresponding to (12) and (13) are

$$r_k = \frac{-[\cos(\Delta\phi_k) - \sin(\Delta\phi_k)] + 1}{2} \tag{14}$$

$$s_k = \frac{-[\cos(\Delta\phi_k) + \sin(\Delta\phi_k)] + 1}{2} \tag{15}$$

Each of the four possible values of $\Delta\phi_k$, $[0,\pi/2,\pi,3\pi/2]$ results in the symbols $\{r_k,s_k\} = [\{0,0\},\{1,0\},\{1,1\},\{0,1\}]$. If received signal is distortion and noise free, r_k and s_k correspond to the transmitted data channels u_k and v_k (as in **Fig. 6**) respectively.

PDM-DQPSK transmission can be achieved by separately modulating the orthogonal polarizations of the carrier and recombining using a polarization beam combiner. When PDM signals are transmitted in an optical fiber they experience, Polarization Mode Dispersion (PMD) and Polarization Dependent Loss (PDL); which can be very severe in long-haul links and as such must be compensated for at the receiver. Demultiplexing of the two polarization streams is also necessary. In a DD based receiver polarization demultiplexing is accomplished optically using polarization tracking, which is also used for PMD and PDL compensation. Optically based polarization tracking configurations involve multiple optical and optoelectronic components (such as beam splitters and photodetectors) and electronic control and can be very challenging. The requirement for optically based tracking can be removed by using coherent detection and post-processing of the resulting electrical signals. The practical implementation of DD based receiver structures becomes impractical for higher order modulation formats. The alternative is to employ coherent detection, which allows polarization demultiplexing, demodulation and fiber channel compensation to be realized in the electrical domain thereby significantly reducing the receiver optical front-end complexity especially in terms of interferometric demodulation structures.

Coherent Detection

In the context of coherent lightwave systems the term coherent refers to techniques employing mixing between two optical waves on a photodetector (Ip *et al.*, 2008, Kikuchi, 2016). Optical receivers employing coherent detection are linear and thereby allow access, in the electrical domain, to all of the information – amplitude, frequency, phase and polarization – present in the detected

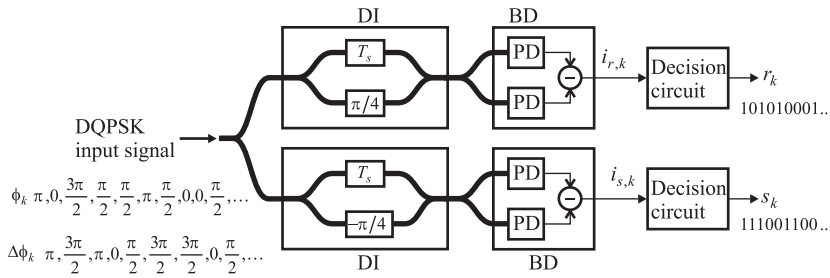


Fig. 7 DQPSK demodulator.

lightwave. The development of high speed Analog to Digital Converters (ADCs) and Application Specific Integrated Circuits (ASICs) has made it possible to carry out real-time DSP on multi-Gb/s data rate signals, which can be employed to equalize fiber impairments, carry out polarization demultiplexing and demodulate high-order modulation formats. DSP also allows the implementation of and flexible tunable WDM receivers. Photonic Integrated Circuit (PIC) technology has experienced rapid development to the point where most of the optical and electronic components required in coherent systems are commercially available or at an advanced state of research and development. This makes it possible to realize reliable and cost-effective solutions for meeting the demand for ultra-fast data rates.

Basic Principles

The fundamental concept in coherent detection is to multiply a modulated optical signal field $E_s(t)$ by a LO field $E_l(t)$ that acts as a phase reference, which enables full recovery of the amplitude and phase information of the signal. The complex optical fields of the signal and LO at the receiver input are

$$E_s(t) = \sqrt{2P_s}A_s(t)\exp\{j[\omega_s t + \theta_{sn}(t) + \theta_0]\} \tag{16}$$

$$E_l(t) = \sqrt{2P_l}\exp\{j[\omega_l t + \theta_l(t)]\} \tag{17}$$

where P_l is the LO power, ω_s and ω_l the received signal and LO optical angular frequencies respectively, $\theta_{sn}(t)$ and $\theta_l(t)$ the signal and LO phase noises respectively and θ_0 is the static phase shift between the signal and LO. The basic form of a balanced coherent receiver is shown in Fig. 8, in which the identical photodiodes have responsivity R. Assuming co-polarized signal and LO fields, the fields prior to detection by the upper and lower photodiodes as given by

$$E_1 = \frac{1}{\sqrt{2}}(E_s + E_l) \tag{18}$$

$$E_2 = \frac{1}{\sqrt{2}}(E_s - E_l) \tag{19}$$

with corresponding photocurrents

$$i_1(t) = \frac{R}{2}\left\{P_s a(t)^2 + P_l + 2\sqrt{P_s P_l}a(t)\cos[\Delta\omega\, t + \theta_s(t) + \theta_{sn}(t) - \theta_l(t) + \theta_0]\right\} \tag{20}$$

$$i_2(t) = \frac{R}{2}\left\{P_s(t)a(t)^2 + P_l - 2\sqrt{P_s P_l}a(t)\cos[\Delta\omega\, t + \theta_s(t) + \theta_{sn}(t) - \theta_l(t) + \theta_0]\right\} \tag{21}$$

Subtracting $i_2(t)$ from $i_1(t)$, gives the BD output

$$i(t) = 2R\sqrt{P_s P_l}a(t)\cos[\Delta\omega\, t + \theta_s(t) + \theta_n(t) + \theta_0] \tag{22}$$

having a total phase noise $\theta_n(t) = \theta_{sn}(t) - \theta_l(t)$. $\Delta\omega = \omega_s - \omega_l$ is the Intermediate Frequency (IF).

Heterodyne Detection

In heterodyne detection $\Delta\omega$ is non-zero. In order to avoid signal distortion caused by spectral folding, $|\Delta\omega|$ must be larger than the modulation bandwidth ($\approx 2\pi R_s$rad/s) of the optical carrier, i.e., $|\Delta\omega| > 2\pi R_b$, as shown in Fig. 9. It is possible to recover $a(t)$ using envelope detection, which involves electronically mixing $i(t)$ with itself and low-pass filtering to eliminate the component centered at $2\omega_{IF}$ and pass the baseband term proportional to $a^2(t)$. However, phase information cannot be recovered by this technique. Constant envelope formats such as M-PSK can be demodulated by using electrical differential detection which, in a manner similar to optical differential detection, determines the phase difference between consecutive symbols. However, pre-coding of the transmitted data is required.

Synchronous electrical detection can be employed to fully recover $A_s(t)$ such as in the receiver shown in Fig. 10. $i(t)$ is split into two and electrical LOs are used to translate the IF signals to baseband. LPFs pass the baseband signal and eliminate the terms centered at $2|\Delta\omega|$. An electrical Phase Locked Loop (PLL) provides carrier phase recovery by compensating for the IF signal phase noise. If the phase noise is totally eliminated then the output photocurrents $i_I(t)$ and $i_Q(t)$ are given by

$$i_I(t) = R\sqrt{P_s P_l}a(t)\cos[\theta_s(t)] \tag{23}$$

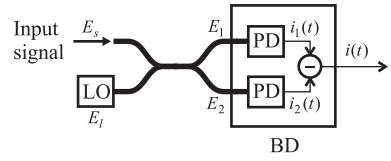


Fig. 8 Coherent receiver using balanced detection.

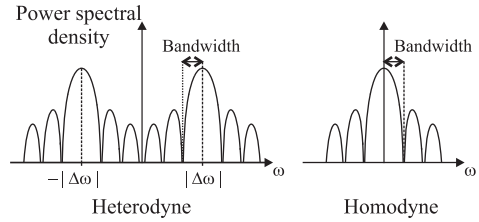


Fig. 9 Power spectrums of heterodyne and homodyne down-converted signals.

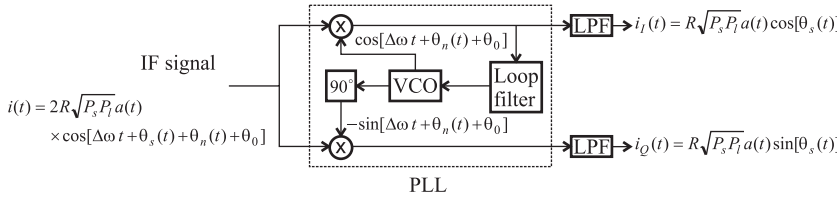


Fig. 10 Synchronous heterodyne receiver utilizing an electrical PLL. In the locked state the PLL tracks the IF signal phase noise to produce a Voltage Controlled Oscillator (VCO) output that acts as a phase reference for the I-channel.

$$i_Q(t) = R\sqrt{P_s P_l}\,a(t)\sin[\theta_s(t)] \tag{24}$$

which, when divided by $R\sqrt{P_s P_l}$, are equal to the I and Q-components of $A_s(t)$. A major disadvantage of the heterodyne receiver is the requirement that $|\Delta\omega| > 2\pi R_b$, so the minimum bandwidth requirement of the photodiodes is $2\Delta\omega$, which for high symbol rates can be very large. It is also difficult to design wideband low-distortion electrical mixers. An alternative to using an electrical PLL is to use an Optical PLL (OPLL) to match the LO phase to the input Signal Carrier (SC) phase; however the practical implementation of an OPLL is difficult mainly due to limitations in the achievable loop bandwidth. An additional difficulty is the relatively high phase noise of Distributed Feedback (DFB) semiconductor lasers, which are the preferred choice for the signal carrier and LO (an alternative is to use external cavity lasers, which however are of high cost compared to DFB lasers). Because of advances in high-speed DSP circuits, the practical implementation of wideband electrical PLLs is feasible.

Homodyne and Intradyne Detection

In homodyne detection $\Delta\omega = 0$, in which case (22) is given by

$$i(t) = 2R\sqrt{P_s P_l}\,a(t)\cos[\theta_s(t) + \theta_n(t) + \theta_0] \tag{25}$$

To achieve $\Delta\omega = 0$ requires synchronization of the LO frequency with the signal carrier frequency. In order to decode the transmitted symbols, the LO phase must track the SC phase noise. This can be achieved by an OPLL. When SC-LO phase synchronization is achieved, (25) gives the real part of $A_s(t)$; however the imaginary part of $A_s(t)$ cannot be detected. The latter can be detected by the use of a phase diversity receiver, discussed later in the context of intradyne reception.

In commercial coherent systems intradyne reception is employed, whereby $|\Delta\omega|$ is less than the symbol rate; but not necessarily equal to zero. This means that the SC and the LO do not need to be phase locked to each other thereby avoiding the need for an OPLL. In intradyne systems $\Delta\omega$ is referred to as the Frequency Offset (FO). The FO results in a steady angular rotation $\Delta\omega t$ experienced by the received signal constellation. With the use of commercial lasers, the magnitude of the FO may be up to several GHz. If the angular rotation over a symbol period is significant, FO estimation and compensation is necessary.

Phase Diversity Receiver

If the signal and LO are co-polarized, full recovery of $A_s(t)$ is possible by using phase-diversity detection as shown in **Fig. 11**. The receiver consists of a 90° Optical Hybrid (OH) and two BDs. The receiver has the major advantage of allowing carrier recovery to be carried out at baseband, which is now feasible with the availability of high-speed DSP circuits thereby negating the need for an OPLL.

Assuming non-zero $\Delta\omega$, the photocurrents from the BDs are given by

$$i_I(t) = R\sqrt{P_s P_l}\,a(t)\cos[\Delta\omega\, t + \theta_s(t) + \theta_n(t) + \theta_0] \tag{26}$$

$$i_Q(t) = R\sqrt{P_s P_l}\,a(t)\sin[\Delta\omega\, t + \theta_s(t) + \theta_n(t) + \theta_0] \tag{27}$$

Because $1/(2\pi|\Delta\omega|)$ is typically much smaller than the symbol period, (26) and (27) are effectively baseband signals, having bandwidths approximately equal to the modulation bandwidth. This means that the receiver bandwidth requirements are greatly relaxed compared to heterodyne detection, which negates the requirement for wideband electrical mixers, prior to any necessary

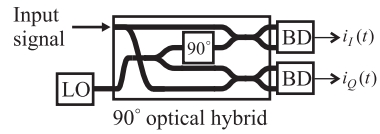


Fig. 11 Phase-diversity receiver.

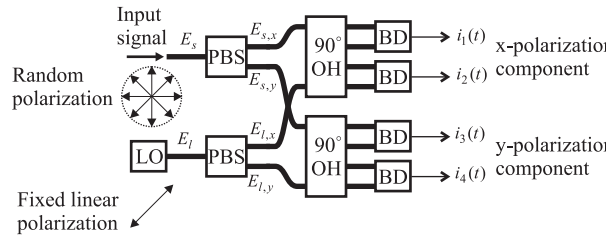


Fig. 12 Phase and polarization diversity receiver.

DSP. (26) and (27) can be combined to form a complex amplitude $A_c(t)$, given by

$$A_c(t) = \frac{1}{R\sqrt{P_l}}[i_I(t) + i_Q(t)] = \sqrt{P_s}a(t)\exp\{j[\Delta\omega\, t + \theta_s(t) + \theta_n(t) + \theta_0]\} \tag{28}$$

$A_s(t)$ can be recovered if the FO, phase noise and θ_0 are removed from (28) by applying suitable DSP.

Phase and Polarization Diversity Receiver

In practical systems the SC and LO polarizations are different. In the worst case scenario, when the SC and LO polarizations are orthogonal, the phase diversity receiver photocurrents will be zero. This problem can be overcome by the use of the phase and polarization diversity receiver shown in **Fig. 12**. The LO is polarized at 45° with respect to the Polarization Beam Splitter (PBS) polarization axes (x and y) to give equal LO powers at the PBS outputs. The split signal and LO components are then input to two 90° OHs. If it is assumed that the signal x- and y-polarization components are in-phase then the output photocurrents are given by

$$i_1(t) = R\sqrt{\frac{\alpha P_s P_l}{2}}\, a(t)\cos[\Delta\omega t + \theta_s(t) + \theta_n(t) + \theta_0] \tag{29}$$

$$i_2(t) = R\sqrt{\frac{\alpha P_s P_l}{2}}\, a(t)\sin[\Delta\omega t + \theta_s(t) + \theta_n(t) + \theta_0] \tag{30}$$

$$i_3(t) = R\sqrt{\frac{(1-\alpha) P_s P_l}{2}}\, a(t)\cos[\Delta\omega t + \theta_s(t) + \theta_n(t)] \tag{31}$$

$$i_4(t) = R\sqrt{\frac{(1-\alpha)\, P_s P_l}{2}}\, a(t)\sin[\Delta\omega t + \theta_s(t) + \theta_n(t)] \tag{32}$$

where α is the ratio of the x-polarization signal power to the total signal power. The complex amplitudes of the x- and y-polarizations corresponding to (29–32) can be written as

$$E_x(t) = \frac{1}{R}\sqrt{\frac{2}{P_l}}[i_1(t) + j\, i_2(t)] = \sqrt{\alpha\, P_s}a(t)\exp\{j[\Delta\omega\, t + \theta_s(t) + \theta_n(t) + \theta_0]\} \tag{33}$$

$$E_y(t) = \frac{1}{R}\sqrt{\frac{2}{P_l}}[i_3(t) + j\, i_4(t)] = \sqrt{(1-\alpha)\, P_s}a(t)\exp\{j[\Delta\omega\, t + \theta_s(t) + \theta_n(t) + \theta_0]\} \tag{34}$$

When the FO, phase noise and θ_0 are removed from (33,34), then $A_s(t)$ can be recovered in a polarization independent manner.

Signal-to-Noise Ratio

A major advantage of coherent detection is that the LO power can be increased to a level such that the receiver noise is negligible, in which case the receiver noise performance is determined by the Signal-to-Noise Ratio (SNR) of the received symbols. There is no difference in the performance of the phase diversity homodyne receiver and heterodyne receiver if in the latter an optical filter is used to reject optical frequency bands centered at $\omega_I \pm \Delta\omega$, which otherwise are down-converted to the same IF as the signal of interest and introduce WDM crosstalk. In a heterodyne system employing optical amplifiers, optical filtering avoids excess Amplified Spontaneous Emission (ASE) noise. The phase diversity receiver I and Q outputs can be combined to form a complex

valued signal, accompanied by additive complex valued white noise. The value of the SNR per symbol γ_s depends on the dominant noise source. In the quantum noise limited regime, encountered in links not employing optical amplifiers, γ_s is equal to the average number of photons per symbol n_s. In long-haul links, in-line optical amplifiers are used to periodically compensate for the transmission fiber loss as shown in Fig. 13, where the number of fiber spans N_a equals the number of (identical) polarization independent amplifiers. The detected noise (signal-spontaneous beat noise) is much larger than the quantum noise of the signal itself. Assuming that the noise figure of each amplifier is the minimum possible value of 3 dB, then γ_s is equal to n_s/N_a.

Spectral Efficiency and BER

The Spectral Efficiency (SE) (bit/s/Hz) of a communications link is the data rate that can be transmitted over a given bandwidth, excluding the use of error-detection codes. It is a measure of how efficiently the limited available spectrum is utilized. The maximum achievable SE (bit/s/Hz/polarization/channel) of a linear communication link in the presence of Additive White Gaussian Noise (AWGN) is given by the Shannon limit

$$SE_{\max} = \log_2(1 + SE_{\max}\gamma_b) \tag{35}$$

where γ_b is the ratio of the energy per bit to noise power spectral density, which is equal to the Signal-to-Noise Ratio (SNR) per bit. The symbol SNR $\gamma_s = b\gamma_b$. (35) can be rewritten as

$$\gamma_b = \frac{2^{SE_{\max}} - 1}{SE_{\max}} \tag{36}$$

which enables $SE_{\max}$ to be determined as a function of γ_b. The SE of a specific modulation format is given by

$$SE = \frac{\log_2(M)}{D/2} \tag{37}$$

where D is the dimension of the modulation format. In coherent systems, both the I and Q components of the carrier lightwave are used so $D=2$ for SP transmission and $D=4$ for DP transmission.

The principle metric for evaluating the performance of a communication link is the BER. In the absence of any transmission link impairments the BER only depends on γ_b. The symbol error rate and BER are approximately equal when Gray coding is used. Gray coding is not possible for PS formats; the best that can be done is to encode the levels so that constellation point pairs that are furthest away from each other have inverted binary code words, in which case the BER is approximately half the symbol error rate. Exact and approximate BER expressions for common modulation formats are given in Table 1, from which γ_b required to achieve a target BER can be calculated. A target BER of 10^{-3} is typical in systems employing FEC. Fig. 14 shows the SE as a function of γ_b for different modulation formats and also the Shannon limit, which shows that the use of high-order modulation formats increases SE at the expense of a larger SNR. It is also evident PS has the highest power efficiency among 4D formats.

Digital Coherent Receiver

In order to recover the transmitted symbols, the x- and y-polarization signals from the phase and polarization diversity optical receiver must be processed. The required processing can be carried out using a DCR, the main functional blocks of which are shown in Fig. 15. First the real and imaginary parts of the complex amplitudes of the x- and y-polarizations from the optical front-end are passed through Anti-Aliasing Filters (AAFs), digitized by a 4-channel ADC and combined to obtain the sampled complex amplitudes $E_x(n)$ and $E_y(n)$, where n is the sample number. The next stage is the removal, using a fixed equalizer, of Intersymbol Interference (ISI) caused by the fiber fixed Group Velocity Dispersion (GVD). Clock recovery, to determine the symbol rate and optimum sampling times, is usually performed after GVD compensation (Zhou, 2014). Linear polarization dependent fiber impairments are removed using adaptive equalization, and if necessary the DP signals (X and Y) can be demultiplexed. Next FO estimation and compensation are carried out followed by carrier phase detection and compensation after which the symbols are decoded and FEC applied.

Sampling

In order to minimize the DSP load, it is desirable to band limit the received signal to the minimum possible value that allows successful data recovery. The theoretical minimum bandwidth of the optical signal, termed the Nyquist bandwidth, is equal to the

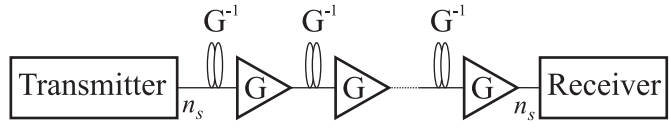


Fig. 13 Optical amplifier chain. The amplifier gain G compensates for the loss G^{-1} of the preceding fiber span.

Table 1 Exact and approximate BER expressions for various receivers assuming Gray coding, except in the case of PS-QPSK

Detection scheme	*BER*
Differential detection: DQPSK	$= Q(\alpha,\beta) - \frac{1}{2} I_0(\alpha\beta)\exp\left[-\frac{1}{2}(\alpha^2+\beta^2)\right]$, where $\alpha = \sqrt{2\gamma_b(1-\sqrt{1/2})}$ and $\beta = \sqrt{2\gamma_b(1+\sqrt{1/2})}$
Coherent detection: BPSK and QPSK	$= \frac{1}{2} erfc(\sqrt{\gamma_b})$
Coherent detection: M-PSK (M > 4)	$\approx \frac{1}{b} erfc\left[\sqrt{b\gamma_b}\sin\left(\frac{\pi}{M}\right)\right]$
Coherent detection: M-QAM	$\approx \frac{2}{b}\left(1-\frac{1}{\sqrt{M}}\right) erfc\left[\sqrt{\frac{3b\gamma_b}{2(M-1)}}\right]$
Coherent detection: PS-QPSK	$\approx \frac{1}{2}\left\{1-\frac{1}{\sqrt{\pi}}\int_0^\infty [1-erfc(x)]^3 \exp\left[-(x-\sqrt{3\gamma_b})^2\right] dx\right\}$

Q: Marcum Q-function; I_0: zero-order modified Bessel function of the first kind; *erfc*: complementary error function.

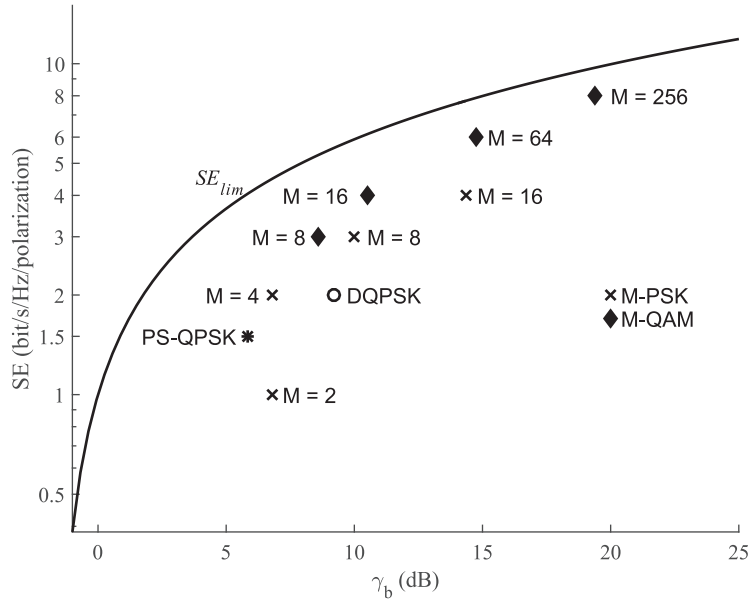


Fig. 14 Spectral efficiency vs. SNR per bit for different modulation formats for a BER = 10^{-3}. The Shannon limit is also shown. 4-QAM is equivalent to 4-PSK. PS based formats can only operate in DP mode.

symbol rate (per polarization), so the minimum bandwidth of the baseband I and Q signals is $R_s/2$, which means that the theoretical minimum required sampling rate is equal to R_s. It is not possible to construct an ideal AAF that band limits the polarization diversity receiver outputs to a bandwidth equal to $R_s/2$; practical AAFs have an amplitude response that extends beyond $R_s/2$. Symbol-rate sampling is also susceptible to sampling time errors. This limitation can be overcome by using a sampling rate pR_s, where $1<p<2$ is a rational oversampling ratio. This also avoids aliasing up to a signal bandwidth of $pR_s/2$. Clock recovery and adaptive equalization require two-times oversampling, so if fractional sampling is used, the sampled signals must be resampled at a rate of $2R_s$, which increases the processing complexity.

Group Velocity Dispersion Equalization

GVD is caused by CD, which is a combination of waveguide and material dispersion. The fiber refractive index $n(\omega)$ is frequency dependent, which leads to a frequency dependent group velocity $v_g(\omega) = c/n(\omega)$. Each optical pulse contains a continuum of components having frequencies within the pulse bandwidth, which travel at different group velocities. Thereby each of these components have different arrival times at the receiver, resulting in pulse broadening and ISI, the severity of which depends on the

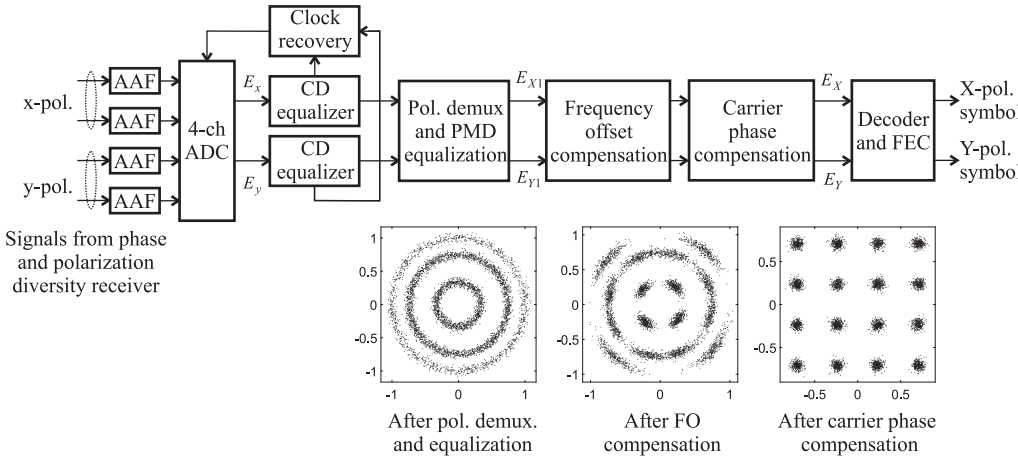


Fig. 15 DCR configuration showing a square 16-QAM signal constellation at various stages in the signal recovery process.

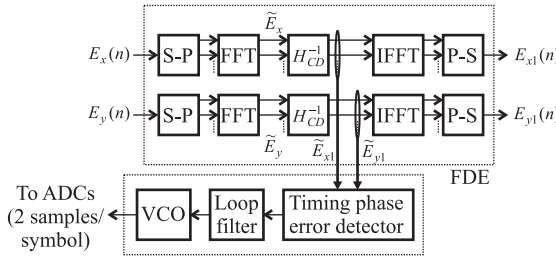


Fig. 16 General configuration of a FDE and frequency domain based clock extraction.

pulse bandwidth, transmission length and the fiber dispersion at the operating wavelength. In the frequency domain the fiber CD has the scalar transfer function

$$H_{CD}(\omega) = \exp\left[-jL\left(\frac{\beta_2\omega^2}{2} + \frac{\beta_3\omega^3}{6}\right)\right] \tag{38}$$

where ω is the angular frequency deviation from ω_s and L, β_2 and β_3 are the fiber length, dispersion parameter and dispersion slope respectively at ω_s. Because the GVD is time invariant it can be removed from each polarization tributary by the use of a digital filter, operating at two samples per symbol, having a transfer function H_{CD}^{-1} (equal to the complex conjugate H_{CD}^{*}). There are two ways of implementing the filter; a Finite Impulse Response (FIR) filter or a Frequency Domain Equalizer (FDE). FIR filters operate on serial data but are difficult to implement if the number of required taps is large. FDE equalization greatly reduces the computational load by using block-by-block processing; with the additional advantage that it is especially suited for use with some common clock recovery schemes. The general structure of a FDE is shown in Fig. 16. The digitized x- and y-polarization complex amplitudes are divided into blocks (of suitable size N) using a Serial-to-Parallel (s-P) converter and transformed to the frequency domain by calculating their Fourier transforms $\tilde{E}_x$ and $\tilde{E}_y$, using the Fast Fourier Transform (FFT) algorithm. If two-times oversampling is used, the discrete frequency k components of the FFTs range from $-1/T_s$ to $1/T_s$, with a resolution equal to $2/NT_s$. The spectral components are then multiplied by H_{CD}^{-1}, reconverted to the time domain by applying an Inverse FFT (IFFT), and serialized by a Parallel-to-Serial (p-S) converter. In practical systems, the GVD value is monitored by a pilot tone, added at the transmitter, and the required filter tap coefficients determined from the monitored GVD value. A major advantage of using electrical GVD compensation is that it eliminates the need for dispersion-compensation fiber.

Clock Recovery

After GVD compensation, clock recovery is carried out to determine the symbol clock frequency and optimum sampling times. Clock recovery can be carried out using well known techniques such as the Gardner method, which is particularly useful if the received signal contains some energy at the clock frequency. Two samples per symbol and knowledge of the previous symbol timing are required in order to estimate the timing error for the current symbol. If the x-polarization signal is considered, the timing error τ_e is calculated, after CD equalization, as

$$\tau_e = \sum_{n=0}^{N/2-1} \left[E_{x1}(2n-1) - E_{x1}(2n+1)\right] E_{x1}^{*}(2n) \tag{39}$$

which is computed in the frequency domain as

$$\tau_e = \sum_{k=0}^{N/2-1} \mathrm{Im}\left[\tilde{E}_{x1}(k)\tilde{E}_{x1}^{*}(k+N/2)\right] \tag{40}$$

where E_{x1} and $\tilde{E}_{x1}$ are the CD compensated x-polarization signal and its FFT respectively. The estimated timing error is then passed through a loop-filter, whose output controls the frequency and phase of the clock VCO, in such a way as to optimize the sampling times. This method works well when the received optical signals do not have significant PMD. In order to tolerate large PMD, both x- and y-polarizations need to be considered. A modified Gardner method that uses both polarizations, calculates the frequency domain timing phase error as

$$\tau_e = \sum_{k=0}^{N/2-1} \mathrm{Im}\left\{\left[\tilde{E}_{x1}(k) + \tilde{E}_{y1}(k)\, e^{j(\varphi)_U}\right]\left[\tilde{E}_{x1}^{*}(k+N/2) + \tilde{E}_{y1}(k+N/2)\, e^{j(\varphi)_L}\right]\right\} \tag{41}$$

which includes rotational phase angles ϕ_U and ϕ_L in the positive and negative frequency components respectively. Appropriate values of ϕ_U and ϕ_L are chosen to mitigate the effects of first-order PMD by introducing a relative phase between the x- and y-polarizations.

PMD and PDL Equalization and Polarization Demultiplexing

PMD is caused by the differential arrival time of the different polarization components of an input light pulse at the fiber output (Galtarossa and Menyuk, 2006). In single-mode fiber PMD is due to the fiber birefringence (polarization dependent effective refractive index) caused by noncircularity of the core and inhomogeneity of the fiber as well as external stress-induced material birefringence, such as bends and twists. The birefringence changes randomly along the fiber length and is wavelength and time dependent. Modeling the propagation of a pulse through a long length of fiber is complicated because of random polarization mode coupling and pulse splitting at every change in the local birefringence.

PMD is commonly described using the principal states model. For a given length of fiber and in the absence of PDL there exist two orthogonal polarization modes (Principal States of Polarization, PSPs) at the input for which, to a first-order approximation, the output polarization is independent of the frequency content of the lightwave. The PSPs propagate at different speeds according to a slow and fast axis induced by the birefringence of the fiber. For each pair of input PSPs there is a corresponding pair of orthogonal PSPs at the fiber output. When a light pulse is launched into any polarization state other than a PSP, the two polarization components slowly separate so the resulting composite pulse is broadened and distorted as shown in Fig. 17. The time difference between the fast and slow polarization states after a given propagation distance is called the instantaneous Differential Group Delay (DGD) $\Delta\tau$. Together, the frequency independent DGD and the PSPs are the fundamental manifestations of first-order PMD. The instantaneous value of $\Delta\tau$ (at a particular wavelength) at the fiber output is a random variable. From experiment and theory it has been shown that $\Delta\tau$ has a Maxwellian probability distribution entirely specified by a single parameter, the mean DGD $\langle\Delta\tau\rangle$, called the PMD of the fiber. For fiber lengths usually encountered in optical transmission systems, $\langle\Delta\tau\rangle$ is proportional to the square root of the fiber length. The constant of proportionality is called the PMD coefficient PMD_{coeff}, which for new types of fiber is typically of the order of 0.1 $ps/\sqrt{km}$. Older installed fiber has PMD_{coeff} values which are usually at least a magnitude higher. The average value of the frequency interval over which the PSPs are frequency independent is called the PSP bandwidth $\Delta\nu_{PSP}$, which is inversely proportional to $\langle\Delta\tau\rangle$. A reasonably accurate value of $\Delta\nu_{PSP}$ in the 1550 nm region is 125/

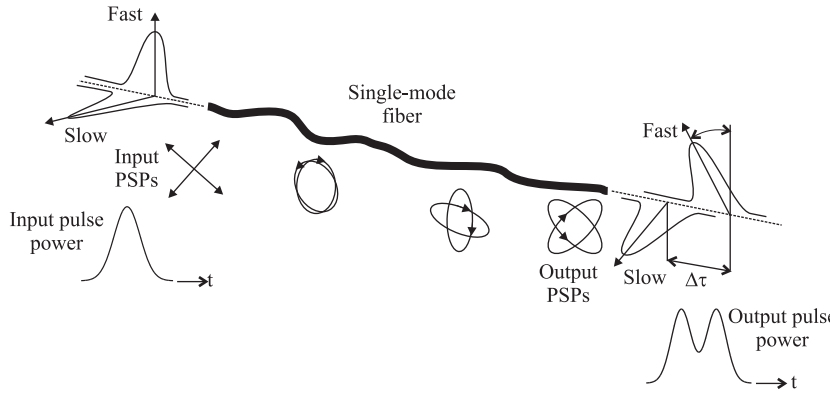


Fig. 17 First-order PMD and pulse distortion in an optical fiber. It is assumed that the input pulse power is equally distributed into the fiber input slow and fast PSPs. In the absence of PDL, the output PSPs are orthogonal and rotated with respect to the input PSPs. Slowly varying birefringence along the fiber leads to rotation of the PSPs. The differential group delay between the PSPs leads to pulse distortion.

$\langle\Delta\tau\rangle$ GHz, when $\langle\Delta\tau\rangle$ is expressed in ps. Considering a fiber with $PMD_{coeff}=0.1ps/km$ and a length of 1000 km, $\langle\Delta\tau\rangle=1ps$ and $\Delta\nu_{PSP}=125$ GHz, so in this case it can be assumed that the first order PMD approximation is valid for data having symbol rates of less than 125 Gbaud.

When the spectral width of the signal exceeds the PSP bandwidth, both the DGD and the PSPs vary across the signal spectrum, which leads to higher order PMD effects. Second order PMD has two components. The first component is polarization dependent CD, which adds or subtracts from the fiber CD depending on the input polarization state to the fiber. The second and dominant component is rotation of the PSPs across the pulse spectrum, which results in pulse distortion such as overshoots and generation of satellite pulses.

For a given modulation format and data rate, the impact of first-order PMD depends on the DGD and relative intensities of the light in the PSPs. The impairment is most severe when equal amounts of the transmitted light pulse are coupled into the input PSPs, and the impairment is negligible when all of the light is coupled to a single PSP. The transmitted pulses experience distortion and broadening, with the possibility of ISI, the consequence of which is a limitation in the acceptable transmission distance. While some simple practical rules for the maximum tolerable mean DGD in IM-DD schemes are available; in general evaluation of the effects of PMD on signal transmission requires complex simulations. The effects of PMD can be reduced by the use of low PMD_{coeff} fiber. The use of high-order modulation formats, for a given data rate, leads to a reduction in the optical signal bandwidth, thereby reducing the impact of PMD. Optical or electrical compensation offers the potential of reducing PMD impairment to acceptable levels, but because PMD can fluctuate on time scales of the order of a millisecond the compensators need to be rapidly adjustable. The availability of fast DSP circuits has made electrical compensation the preferred option. In the DCR, GVD compensation is followed by PMD equalization and polarization demultiplexing.

The complex amplitude of the optical signal at the input of the transmission system can written as a vector $[E_{x,in}(t),E_{y,in}(t)]^T$, where T denotes the transpose and $E_{x,in}(t)$ and $E_{y,in}(t)$ are the x- and y-polarization complex amplitudes respectively. The Fourier transform of the received complex amplitude is given by

$$\begin{bmatrix}\tilde{E}_x(\omega)\\ \tilde{E}_y(\omega)\end{bmatrix}=\mathbf{H}_o(\omega)\begin{bmatrix}\tilde{E}_{x,in}(\omega)\\ \tilde{E}_{x,in}(\omega)\end{bmatrix} \tag{42}$$

where $[\tilde{E}_{x,in}(\omega),\tilde{E}_{x,in}(\omega)]^T$ is the Fourier transform of $[E_{x,in}(t),E_{y,in}(t)]^T$, ω is the angular frequency deviation from ω_s and the 2×2 matrix $\mathbf{H}_o(\omega)$ is the link transfer function. In the linear region $\mathbf{H}_o(\omega)$ can be modeled as $\mathbf{H}_o(\omega)=H_{CD}(\omega)\mathbf{H}_p(\omega)$, where the link polarization properties are taken into account by the matrix $\mathbf{H}_p(\omega)=\mathbf{U}(\omega)\mathbf{TK}$. The frequency dependent PMD matrix $\mathbf{U}(\omega)$ is given by

$$\mathbf{U}(\omega)=\mathbf{R}_1^{-1}\begin{bmatrix}\exp(j\omega\Delta\tau/2) & 0\\ 0 & \exp(-j\omega\Delta\tau/2)\end{bmatrix}\mathbf{R}_1 \tag{43}$$

where $\mathbf{R}_1$ is a unitary matrix converting each PSP into the x- or y-polarization component. The frequency independent PDL matrix $\mathbf{T}$ is given by

$$\mathbf{T}=\mathbf{R}_2^{-1}\begin{bmatrix}\sqrt{\alpha_{\max}} & 0\\ 0 & \sqrt{\alpha_{\min}}\end{bmatrix}\mathbf{R}_2 \tag{44}$$

where $\mathbf{R}_2$ is a unitary matrix converting each PDL eigenmode (polarization states for which the PDL is either a maximum or minimum) into an x- or y-polarization component. $\alpha_{\max}$ and $\alpha_{\min}$ are the power transmission coefficients of the PDL eigenmodes. $\mathbf{K}$ is a frequency independent 2×2 unitary matrix which mixes the two polarizations at the receiving end of the fiber.

Because all of the above impairments are linear, they can be removed by passing $[\tilde{E}_x(\omega),\tilde{E}_y(\omega)]^T$ through an equalizer having a transfer function equal to $\mathbf{H}_o^{-1}(\omega)$. Because the fiber CD is fixed or varies very slowly it is first removed by using FDE equalization as described above. Polarization demultiplexing and polarization impairment equalization can be implemented by filtering the sampled CD compensated signal $[E_{x1}(n),E_{y1}(n)]^T$ with an equalizer having a transfer function $\mathbf{H}_{eq}(\omega)=\mathbf{H}_p^{-1}(\omega)$. Because $\mathbf{H}_p(\omega)$ is time dependent, adaptive equalization is required. $\mathbf{H}_{eq}(\omega)$ can be written as

$$\mathbf{H}_{eq}(\omega)=\begin{bmatrix}h_{xx}(\omega) & h_{xy}(\omega)\\ h_{yx}(\omega) & h_{yy}(\omega)\end{bmatrix} \tag{45}$$

$\mathbf{H}_{eq}(\omega)$ can be implemented as a 2×2 butterfly structured filter, each element of which is realized by FIR filters with tap coefficient vectors $\overrightarrow{h}_{xx}$, $\overrightarrow{h}_{xy}$, $\overrightarrow{h}_{yx}$ and $\overrightarrow{h}_{yy}$, of equal lengths k, as shown in Fig. 18. The most common method for adaptively finding the optimum tap coefficients is the Constant Modulus Algorithm (CMA). The input signal is normalized such that $|E_{x1}(n)|=|E_{y1}(n)|=1$, as is the case for NRZ M-PSK type signals. The tap coefficient vectors are updated on a symbol by symbol basis as follows:

$$\overrightarrow{h}_{xx}(n+1)=\overrightarrow{h}_{xx}(n)+\mu\, e_x(n)\overrightarrow{E}^*_{x1}(n) \tag{46}$$

$$\overrightarrow{h}_{xy}(n+1)=\overrightarrow{h}_{xy}(n)+\mu\, e_x(n)\overrightarrow{E}^*_{y1}(n) \tag{47}$$

$$\overrightarrow{h}_{yx}(n+1)=\overrightarrow{h}_{yx}(n)+\mu\, e_y(n)\overrightarrow{E}^*_{x1}(n) \tag{48}$$

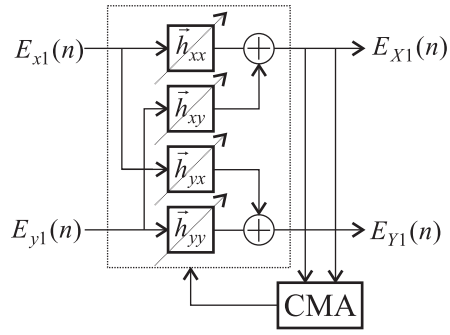


Fig. 18 Configuration of a 2×2 butterfly structured adaptive filter.

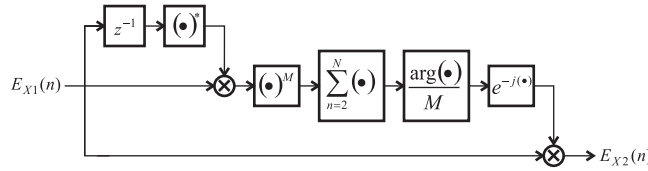


Fig. 19 DSP implementation of the time-domain differential phase M-th power FO estimation and compensation.

$$\vec{h}_{\gamma\gamma}(n+1) = \vec{h}_{\gamma\gamma}(n) + \mu\, e_\gamma(n)\, \vec{E}^*_{\gamma 1}(n) \tag{49}$$

where $e_x(n) = [1 - |E_{X1}(n)|^2]E_{X1}(n)$ and $e_y(n) = [1 - |E_{Y1}(n)|^2]E_{Y1}(n)$ are error signals, μ a step-size parameter and $\vec{E}_{x1}(n)$ and $\vec{E}_{y1}(n)$ are input vectors comprising the current input sample and the previous k samples. When the algorithm converges (such that $e_x(n) \to 0$ and $e_y(n) \to 0$) the polarization dependent impairments are compensated and polarization demultiplexed X- and Y-polarization components $E_{X1}(n)$ and $E_{Y1}(n)$ respectively of the fiber input PDM signals will be present at the equalizer output ports. If SP transmission is used, the signal of interest will be present at only one of the ports. CMA based equalization works well for M-PSK signals but is difficult to apply to M-QAM signals, whose amplitude has multiple levels.

FO Compensation

The adaptive equalizer does not remove the FO or the phase noise from the received signal. A non-zero FO results in the signal experiencing a phase rotation $\Delta\omega T_s$ over a symbol period. After adaptive equalization, the demultiplexed X- and Y-polarization signals are downsampled to one sample per symbol. If we consider the X-polarization component, the downsampled complex amplitude at the equalizer output has the form

$$E_{X1}(n) = a(n)\exp\{j[\Delta\omega\, t_n + \theta_s(n) + \theta_n(n) + \theta_0]\} \tag{50}$$

where t_n is the sampling time and $a(n)$, $\theta_s(n)$ and $\theta_n(n)$ are the sampled modulation signal amplitude and phase and phase noise. In order to remove the FO, it must first be estimated and then (50) reverse rotated by the phase offset $\Delta\omega T_s$ accumulated over a symbol period. For M-PSK modulated signals a common technique for estimating the FO is the time-domain differential phase M-th power algorithm. Considering an M-PSK signal constant $a(n)$, the phase modulation is removed by raising (50) to the M-th power to give

$$E^M_{X1}(n) = a^M \exp\{jM[\Delta\omega\, t_n + \theta_n(n) + \theta_0]\} \tag{51}$$

The complex conjugate of the previous sample raised to the power of M is given by

$$[E^*_{X1}(n-1)]^M = a^M \exp\{-jM[\Delta\omega\,(t_n - T_s) + \theta_n(n-1) + \theta_0]\} \tag{52}$$

Multiplying (51) by (52) gives

$$E^M_{X1}(n)[E^*_{X1}(n-1)]^M = a^{2M}\exp\{jM[\Delta\omega\, T_s + \Delta\theta_n(n)]\} \tag{53}$$

where $\Delta\theta_n(n) = \theta_n(n) - \theta_n(n-1)$ is the phase noise induced random phase within a symbol period. The phase noise can be removed by summing (53) over a suitably long number $N-1$ of single estimates and the FO estimate $\Delta\omega_e$ calculated as

$$\Delta\omega_e = \frac{1}{M\, T_s} \arg\left[\sum_{n=2}^{N} E^M_{X1}(n)[E^*_{X1}(n-1)]^M\right] \tag{54}$$

The implementation of this algorithm and subsequent removal of the FO is shown in **Fig. 19**. Because the arg operator has a range of $[-\pi, \pi]$, the maximum resolvable FO is $\pm 1/2MT_s$, which for a 40 Gbaud 8-PSK signal is ± 2.5 GHz. Therefore coarse tuning of the LO may be necessary to ensure that the FO is within the resolvable range. The performance of the M-power method

for high-order QAM is poor because only a small amount of the constellation points having equal phase separation is useable for FO estimation.

The FO can also be estimated by computing the FFT of the data erased signal, which has a peak at $M|\Delta\omega|$. The accuracy of the estimated FO can be greater than that for the time-domain method; however the computational complexity is much greater. It is not possible to determine the sign of the FO; to do so, necessitates further signal processing. Other FO estimation methods include blind frequency search (no training sequence is required), where the estimated FO is scanned over a particular range and the optimal FO determined using a minimum phase or Euclidian distance method. Another method uses a starting training sequence to obtain an initial FO estimate and then uses the recovered phase angle from the following carrier phase recovery unit to track the FO change. Both of these techniques work for arbitrary modulation formats.

Carrier Phase Compensation and Decoding

After the FO has been compensated, the phase difference between the received signal and LO must be removed before the final decoding stage. The X-polarization complex amplitude after FO compensation is given by

$$E_{X2}(n) = a(n)\exp\{j[\theta_s(n) + \theta_n(n) + \theta_0]\} \tag{55}$$

The linewidth of DFB lasers used in the transmitter and receiver is usually in the range of 100 kHz to 10 MHz, so the phase noise varies much more slowly than the phase modulation. In practice (55) is accompanied by additive noise, originating from the optical signal quantum noise, ASE noise if optical amplifiers are present in the transmission link and receiver electronic noise. It is possible to obtain an accurate estimate of the phase noise and offset by averaging the phase of (55) over a large number N of symbol periods. For M-PSK signals the feedforward M-th power method can be used; as shown in **Fig. 20**. First, each symbol in the l-th block $E_{X2,1}$ of N symbols is raised to the M-th power to erase the phase modulation. The phase estimate over the l-th block is then calculated as

$$\theta_l = \frac{1}{M}\arg\left\{\sum_{n=1}^{N} E_{X2,l}[n + (l-1)N]\right\} \tag{56}$$

Averaging reduces the effect of additive noise but the longer the averaging period the worse the phase estimate; hence the optimal block length is a trade-off between the additive noise and the phase noise; narrower linewidths requiring longer lengths. The symbols are then corrected by reverse rotating the symbols in the block by θ_l. Because the arg operator has a range of $[-\pi,\pi]$, the phase estimate is limited to values between $\pm\pi/M$ and so the resulting symbols will have a phase ambiguity of $2\pi/M$. This can be overcome by the use of differential modulation; however this results in a doubling of the error rate. After phase estimation, $E_{X2,1}$ is reverse rotated by an angle $-\theta_l$ to obtain the phase corrected block $E_{X,1}$, with corresponding complex amplitudes

$$E_X(n) = a(n)\exp[j\theta_s(n)] \tag{57}$$

which represents the correctly aligned X-polarization symbols. In practice (57) will also have an additive noise component due to the intensity noise of the optical signals and receiver noise. The complex amplitudes $E_X(n)$ are then decoded and FEC applied to recover the X-polarization symbols with an acceptable BER.

The M-th power method cannot be directly applied to high-order QAM formats because of the multiplicity of symbol amplitude levels. In the case of square 16-QAM the symbols can be divided into two classes: Class I and Class II as shown in **Fig. 21**. Class I symbols can be regarded as QPSK (M = 4) signals having two amplitude levels; the phase can be estimated by using only these symbols in the 4-th power technique and the estimated phase then used to correct the entire QAM constellation. Class II symbols are not used for the phase estimation.

Nonlinear Impairments

The dominant nonlinear impairment in optical fiber is due to the Kerr effect, which causes a refractive index change proportional to the signal intensity, resulting in nonlinear waveform distortion. This can severely limit the maximum transmission distance of high-order QAM signals. FIR based adaptive filters cannot be used to compensate for such nonlinear effects. The propagation of a DP signal though an optical fiber, in the presence of fiber attenuation, CD and Kerr nonlinearity is described by the nonlinear

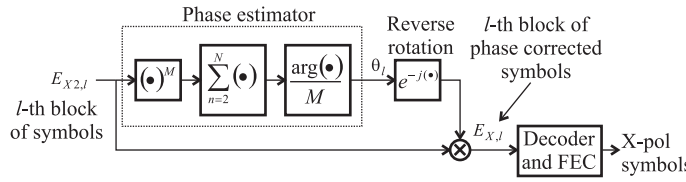


Fig. 20 DSP implementation of feedforward M-th power phase estimation and compensation for M-PSK signals.

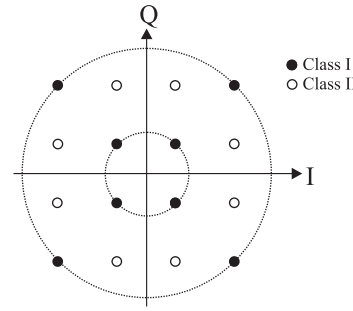


Fig. 21 Square 16-QAM symbol classification.

Schrodinger equation

$$\frac{\partial E_{x,y}}{\partial z} = -\frac{\alpha}{2}E_{x,y} - \frac{j}{2}\beta_2\frac{\partial^2 E_{x,y}}{\partial t^2} + j\gamma\left(|E_x|^2 + |E_y|^2\right)E_{x,y} \tag{58}$$

where α is the attenuation coefficient, γ the nonlinear coefficient and z and t are the propagation distance and time respectively. By reversing the signs of the coefficients in (58), the virtual back-propagation process is described by

$$\frac{\partial E_{x,y}}{\partial z} = \frac{\alpha}{2}E_{x,y} + \frac{j}{2}\beta_2\frac{\partial^2 E_{x,y}}{\partial t^2} - j\gamma\left(|E_x|^2 + |E_y|^2\right)E_{x,y} \tag{59}$$

If the received signal is digitally back-propagated using (59) the effects of deterministic nonlinear impairments can be reversed. If present, optical amplifiers can be included in the back-propagation process as virtual lumped amplifiers having gains equal to the inverse of the link amplifier gains. (59) is usually solved using the split-step Fourier algorithm; however the computational complexity is very high, thereby precluding their use in commercial systems. Therefore there is a critical need for develop improved algorithms that can provide nonlinearity compensation at feasible computational cost. In WDM systems, the signal also suffers deterministic nonlinear effects due to the fields of the neighboring channels, such as cross-phase modulation and four-wave mixing. In amplified systems, Nonlinear Phase Noise (NLPN) results from the interaction between ASE noise and the signal via the Kerr effect. Developing computationally efficient methods for compensation of non-deterministic nonlinear impairments is also of great importance.

Conclusion

Coherent lightwave systems offer the potential to realize ultra-fast communication systems. The combination of coherent detection and the DCR provides new capabilities that are possible without optical signal phase detection. 100 Gb/s DP-QPSK DCR systems are in wide commercial use for core and metropolitan networks. Data rates as high as 400 Gb/s using DP 16-QAM have recently been demonstrated. Higher order format transmission at even higher data rates is an active area of research. Photonic integration technology, including integration with electronic circuits, is undergoing rapid progress enabling more complex optical transmitters and receivers, including LO laser sources, leading to improved performance and reliability. The development of high-speed ASICs, ADCs and DACs underpins the implementation of advanced DCRs that can carry out more complex real-time DSP functions. Concerning the DSP functions, there is a need for flexible and computationally efficient algorithms that can be used to implement adaptive modulation and coding to mitigate transmission link impairments, especially fiber nonlinearities. As coherent lightwave systems technology matures it will also be applied to medium and short reach applications such as data centers and access networks.

See also: Atom Optics. Optical Coherence Tomography and Its Application to Imaging of Skin and Retina

References

Djordjevic, I.B., Arabaci, M., Minkov, L.L., 2009. Next generation FEC for high-capacity communication in optical transport networks. Journal of Lightwave Technology 27, 3518–3530.
Galtarossa, A., Menyuk, C.R. (Eds.), 2006. Polarization Mode Dispersion 1. Berlin Heidelberg: Springer-Verlag.
Gnauck, A.H., Winzer, P.J., 2005. Optical phase-shift-keyed transmission. Journal of Lightwave Technology 23, 115–130.
Ip, E., Lau, A.P.T., Barros, D.J., Kahn, J.M., 2008. Coherent detection in optical fiber systems. Optics Express 16, 753–791.
Karlsson, M., Agrell, E., 2009. Which is the most power-efficient modulation format in optical links? Optics Express 17, 10814–10819.
Kikuchi, K., 2016. Fundamentals of coherent optical fiber communications. Journal of Lightwave Technology 34, 157–179.
Krongold, B., Pfau, T., Kaneda, N., Lee, S.C.J., 2012. Comparison between PS-QPSK and PDM-QPSK with equal rate and bandwidth. IEEE Photonics Technology Letters 24, 203–205.

Nakazawa, M., Kikuchi, K., Miyazaki, T. (Eds.), 2010. High Spectral Density Optical Communication Technologies 6. Berlin Heidelberg: Springer-Verlag.
Roberts, K., O'Sullivan, M., Wu, K.T., *et al*., 2009. Performance of dual-polarization QPSK for optical transport systems. Journal of Lightwave Technology 27, 3546–3559.
Seimetz, M., 2010. High-Order Modulation for Optical Fiber Transmission. Berlin Heidelberg: Springer-Verlag.
Winzer, P.J., 2012. High-spectral-efficiency optical modulation formats. Journal of Lightwave Technology 30, 3824–3835.
Zhou, X., 2014. Efficient Clock and Carrier Recovery Algorithms for Single-Carrier Coherent Optical Systems, 31. IEEE Signal Processing Magazine. pp. 35–45.

Measuring Fiber Characteristics

A Girard, EXFO, Quebec, Canada

Introduction

The optical fiber is divided in two types: multimode and singlemode. Each type is used in different applications and wavelength ranges and is consequently characterized differently. Furthermore, the corresponding test methods also vary.

The optical fiber characteristics may be divided into four categories:

- The optical characteristics (transmission related);
- The dimensional characteristics;
- The mechanical characteristics; and
- The environmental characteristics.

These categories will be reviewed in the following sections, together with their corresponding test methods.

Fiber Optical Characteristics and Corresponding Tests Methods

The following sections will describe the following optical characteristics:

- attenuation;
- macrobending sensitivity;
- microbending sensitivity;
- cut-off wavelength;
- multimode fiber bandwidth;
- differential mode delay for multimode fibers;
- chromatic dispersion;
- polarization mode dispersion;
- polarization crosstalk; and
- nonlinear effects.

Attenuation

The spectral attenuation of an optical fiber follows exponential power decay from the power level at a cross-section 1 to the power level at cross-section 2, over a fiber length L as follows:

$$P_2(\lambda) = P_1(\lambda) \cdot e^{-\gamma(\lambda)L} \tag{1}$$

$P_1(\lambda)$ is the optical power transmitted through the fiber core cross-section 1, expressed in mW; $P_2(\lambda)$ is the optical power transmitted through the fiber core cross-section 2 away from cross-section 1, expressed in mW; $\gamma(\lambda)$ is the spectral attenuation coefficient in linear units, expressed in km^{-1}; and L is the fiber length expressed in km.

Attenuation may be characterized at one or more specific wavelengths or as a function of wavelength. In the later case, attenuation is referred to spectral attenuation. Fig. 1 illustrates such power decay.

Eq. (1) may be expressed in relative units as follows:

$$\log_{10} P_2 = (\log_{10} P_1) \cdot -\gamma L \cdot \log_{10} e \tag{2}$$

P is expressed in dBm units using the following definition.

The power in dBm is equal to 10 times the logarithm in base 10 of the power in mW; or

$$0\ \text{dBm} = 10\log_{10}(1\ \text{mW}) \tag{3}$$

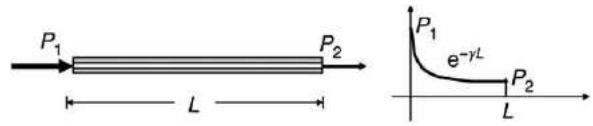


Fig. 1 Power decay in an optical fiber due to attenuation.

 doi:10.1016/B978-0-12-809283-5.00654-6

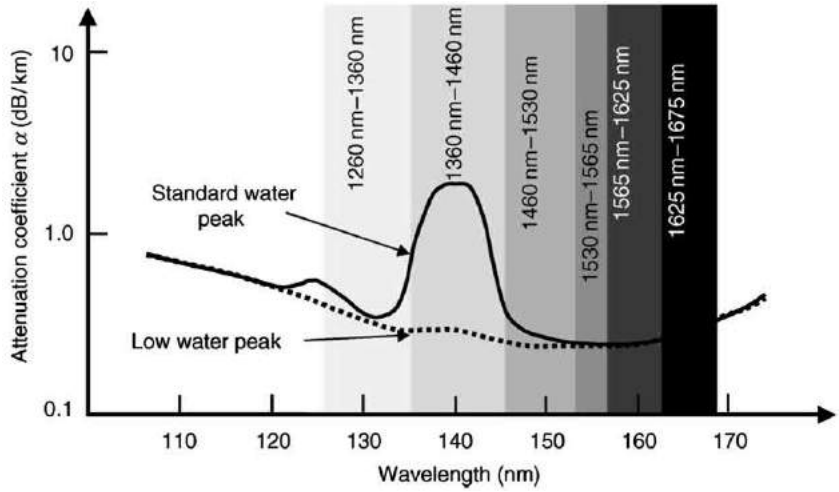


Fig. 2 Typical spectral attenuation of a singlemode fiber.

Then

$$\gamma(\lambda)[\mathrm{km}^{-1}] = [(10\log_{10}P_1) - (10\log_{10}P_2)] \div L\log_{10} e \tag{4}$$

$$\alpha(\lambda)[\mathrm{dB/km}] = [P_1(\mathrm{dBm}) - P_2(\mathrm{dBm})]/L \tag{5}$$

A new relative attenuation coefficient $\alpha(\lambda)$ with units of dB/km has been introduced, which is related to the linear coefficient $\gamma(\lambda)$ with units of km^{-1} as follows:

$$\alpha(\lambda) = (\log_{10} e) \cdot \gamma(\lambda) \approx 4,343\gamma(\lambda) \tag{6}$$

$\alpha(\lambda)$ is the spectral attenuation coefficient of a fiber and is illustrated in Fig. 2.

Test methods for attenuation

The attenuation may be measured by the following methods:

- cut-back method;
- backscattering method; and
- insertion loss method.

Cut-back method

The cut-back method is a direct application of the definition in which the power levels P_1 and P_2 are measured at two points of the fiber change of input conditions. P_2 is the power emerging from the far end of the fiber and P_1 is the power emerging from a point near the input after cutting the fiber.

The output power P_2 is recorded from the fiber under test (FUT) placed in the measurement setup. Keeping the launching conditions fixed, the FUT is cut to the cut-back length, for example 2 m from the launching point. The FUT attenuation, between the points where P_1 and P_2 have been measured, is calculated using P_1 and P_2, from the definition equations provided above.

Backscattering method

The attenuation coefficient of a singlemode fiber is characterized using bidirectional backscattering measurements. This method is also used for:

- attenuation uniformity;
- optical continuity;
- physical discontinuities;
- splice losses; and
- fiber length.

An optical time domain reflectometer (OTDR) is used for performing such characterization. Adjustment of laser pulsewidth and power is used to obtain a compromise between resolution (a shorter pulsewidth provides a better resolution but at lower power) and dynamic range/fiber length (higher power provides better dynamic range but with longer pulsewidth). An example of such an instrument is shown in Fig. 3.

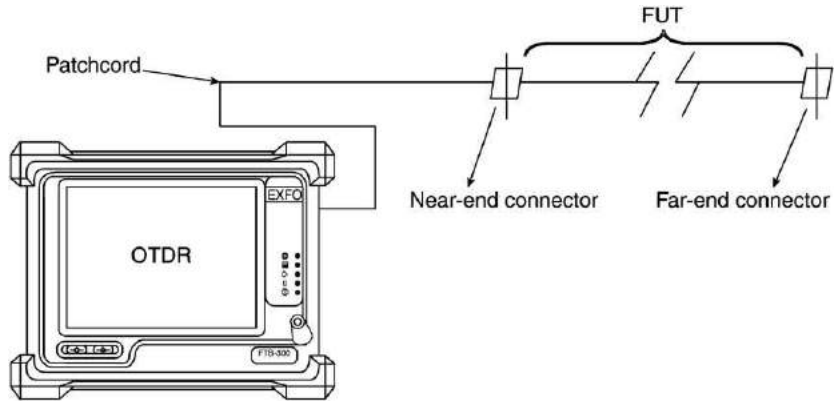


Fig. 3 The backscattering method (OTDR).

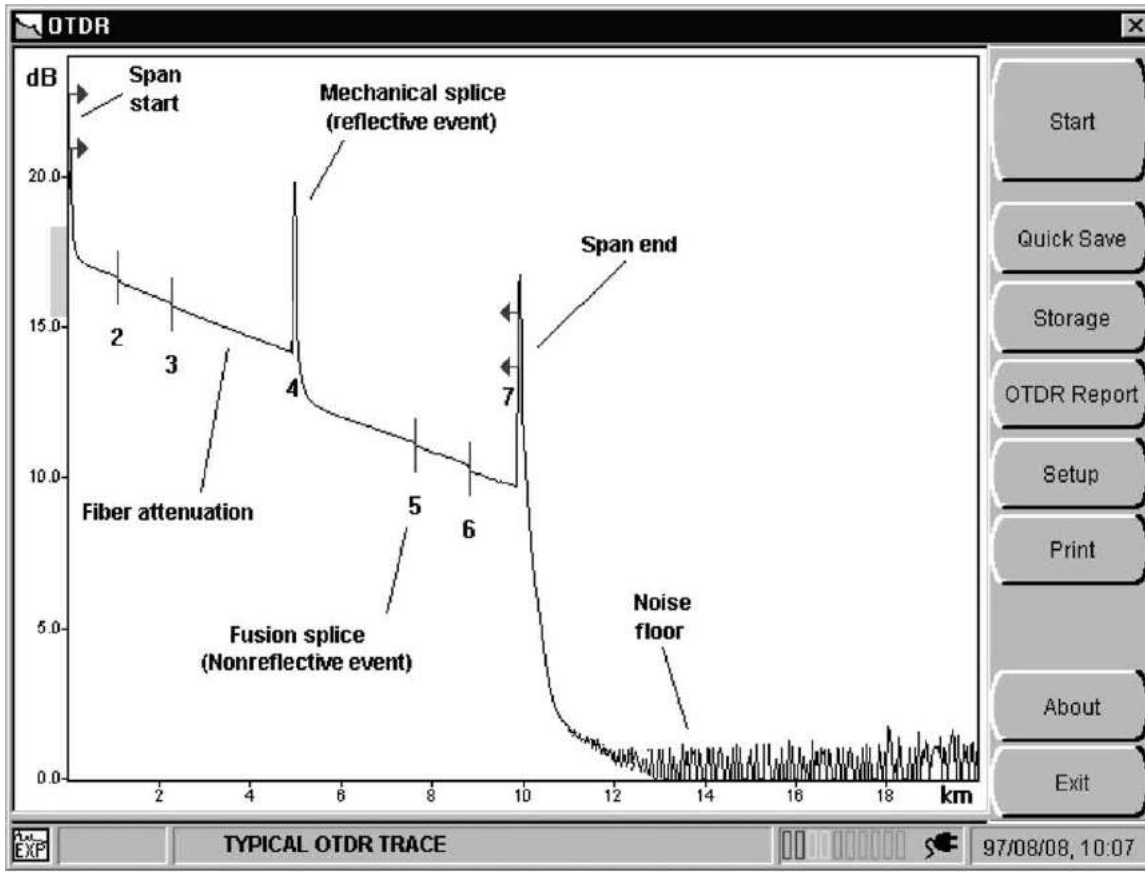


Fig. 4 Unidirectional OTDR backscattering loss measurement.

The measurement is applicable to various FUT configurations (e.g., cabled fiber in production or deployed in the field, fiber on a spool, etc.). Two unidirectional backscattering curves are obtained, one from each end of the fiber (Fig. 4).

Each backscattering curve is recorded on a logarithmic scale, avoiding the parts at the two ends of the curves, due to parasitic reflections.

The FUT length is found from the time interval between the two ends of the backscattering loss curve together with the FUT group index of refraction, n_g, as:

$$L_f = c_{fs} \cdot \frac{\Delta T_f}{n_g} \tag{7}$$

c_{fs} is the velocity of light in free space.

The bidirectional backscattering curve is obtained using two unidirectional backscattering curves and calculating the average loss between them. The end-to-end FUT attenuation coefficient is obtained from the difference between two losses divided by the difference of their corresponding distances.

Insertion loss method

The insertion loss method consists of the measurement of the power loss due to the FUT insertion between a launching and a receiving system, previously interconnected (reference condition). The powers P_1 and P_2 are evaluated in a less straightforward way than in the cut-back method. Therefore, this method is not intended for use in manufacturing.

The insertion loss technique is less accurate than the cut-back, but has the advantage of being non-destructive for the FUT. Therefore, it is particularly suitable in the field.

Macrobending Sensitivity

Macrobending sensitivity is the property by which there is a certain amount of light leaking (loss) into the cladding when the fiber is bent and the bending angle is such that the condition of total internal reflection is no longer met at the core-cladding interface. **Fig. 5** illustrates such a case.

Macrobending sensitivity is a direct function of the wavelength: the longer the wavelength and/or the smaller the bending diameter, the more loss the fiber experiences. It is recognized that 1625 nm is a wavelength that is very sensitive to macrobending (see **Fig. 6**).

The macrobending loss is measured by the power monitoring method (OTDR, especially for field assessment, see **Fig. 6**) or the cut-back method.

Microbending Sensitivity

Microbending is a fiber property by which the core-cladding concentricity randomly changed along the fiber length. It causes the core to wobble inside the cladding and along the fiber length.

Four methods are available for characterizing microbending sensitivity in optical fibers:

- expandable drum for singlemode fibers and optical fiber ribbons over a wide range of applied linear pressure or loads;
- fixed diameter drum for step-index multimode, singlemode, and ribbon fibers for a fixed linear pressure;
- wire mesh and applied loads for step-index multimode and singlemode fibers over a wide range of applied linear pressure or loads; and
- 'basketweave' wrap on a fixed diameter drum for singlemode fibers.

The results from the four methods can only be compared qualitatively. The test is nonroutine for general evaluation of optical fiber.

Cut-off Wavelength

The cut-off wavelength is the shortest wavelength at which a single mode can propagate in a singlemode fiber. This parameter can be computed from the fiber refractive index profile (RIP). At wavelengths below the cut-off wavelength, several modes propagate and the fiber is no longer singlemode, but multimode.

In optical fibers, the change from multimode to singlemode behavior does not occur at a specific wavelength, but rather over a smooth transition as a function of wavelengths. Consequently, from a fiber-optic network standpoint, the actual threshold wavelength for singlemode performance is more critical. Thus an effective cut-off wavelength is described below.

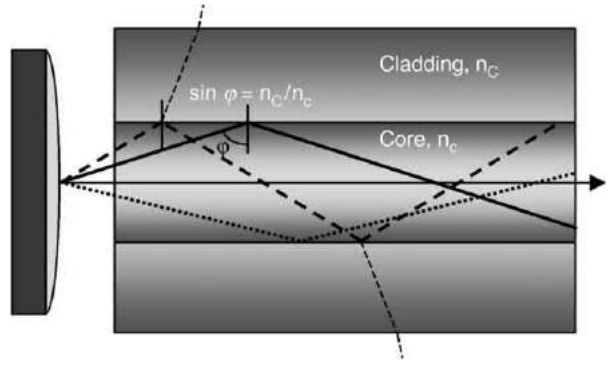


Fig. 5 Total internal reflection and macrobending effect on the light rays.

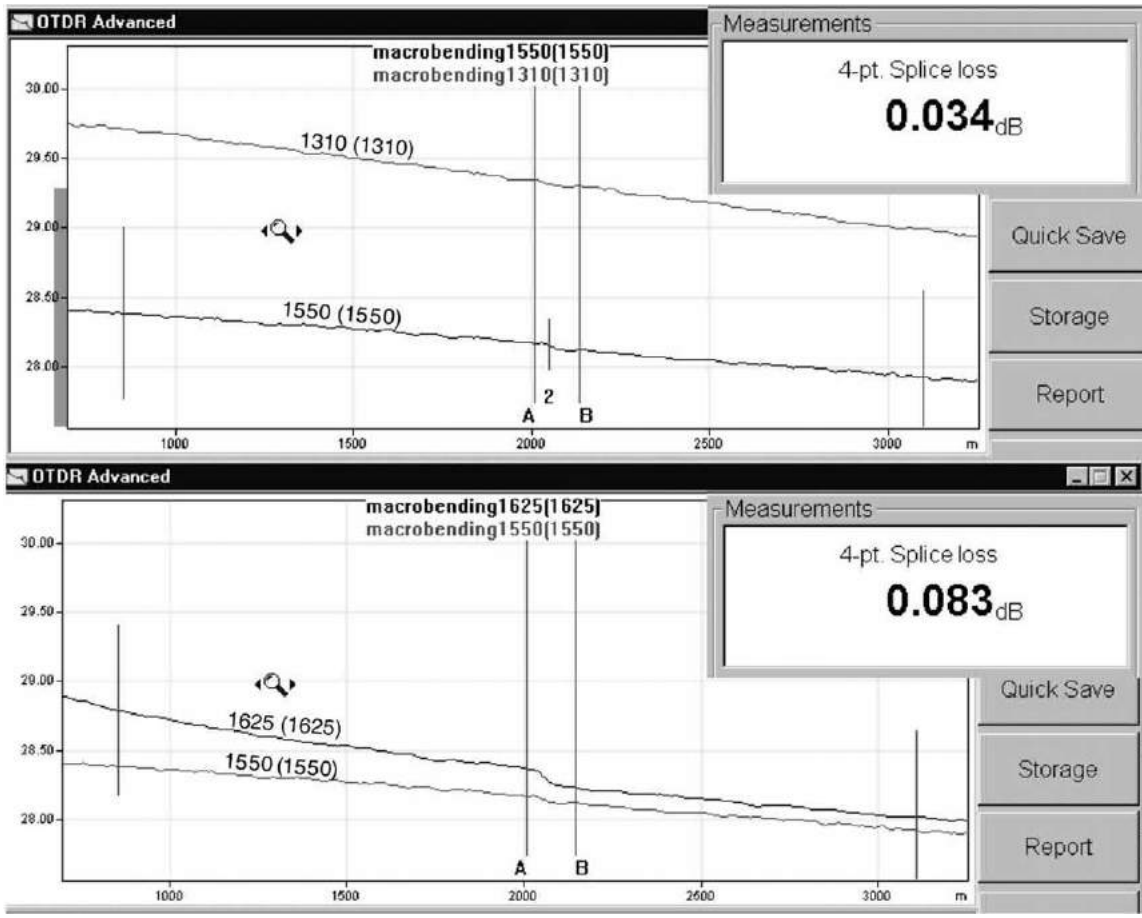


Fig. 6 Macrobending sensitivity as a function of wavelength.

The cut-off wavelength is defined as the wavelength greater than the ratio between the total power, in the higher-order modes and the fundamental mode, which has decreased to less than 0.1 dB. According to this definition, the second-order mode LP_{11} undergoes 19.3 dB more attenuation than the fundamental LP_{01} mode when the modes are equally excited.

Because the cut-off wavelength depends on the fiber length, bends, and strain, it is defined on the basis of the following three cases:

- fiber cut-off wavelength;
- cable cut-off wavelength; and
- jumper cable cut-off wavelength.

Fiber cut-off wavelength: Fiber cut-off wavelength λ_{fc} is defined for uncabled primary-coated fiber and is measured over 2 m with one loop of 140 mm radius loosely constrained, with the rest of the fiber kept essentially straight. The presence of a primary coating on the fiber usually will not affect λ_{fc}. However, the presence of a secondary coating may result in λ_{fc} being significantly shorter than that of the primary coated fiber.

Cable cut-off wavelength: Cable cut-off wavelength is measured prior to installation on a substantially straight 22 m cable length prepared by exposing 1 m of primary-coated fiber at both ends, the exposed ends each incorporating a 40 mm radius loop. Alternatively, this parameter may be measured on 22 m primary-coated uncabled fiber in the same configuration as for the λ_{fc} measurement.

Jumper cable cut-off wavelength: Jumper cable cut-off wavelength is measured over 2 m with one loop of 76 mm radius, or equivalent (e.g., split mandrel), with the rest of the jumper cable kept essentially straight.

Multimode Fiber Bandwidth for Multimode Fibers

The -3 dB bandwidth of a multimode optical fiber (or modal bandwidth) is defined as the lowest frequency where the magnitude of the baseband frequency response in optical power has decreased by 3 dB relative to the power at zero frequency. Modal bandwidth is also called intermodal dispersion as it takes into account the dispersion between the modes of propagation of the transmitted signal into the multimode fiber.

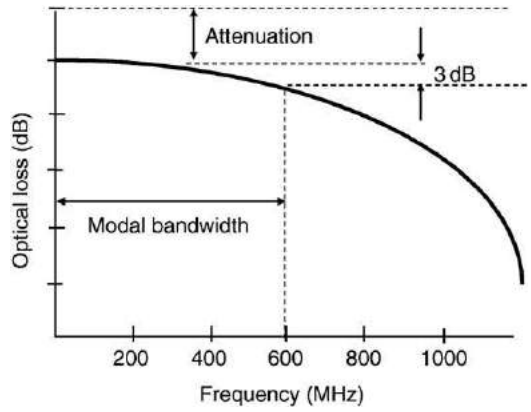


Fig. 7 Determination of modal bandwidth.

Various methods of reporting the results are available, but the results are typically expressed in terms of the −3 dB (optical power) frequency. Fig. 7 illustrates modal bandwidth.

The bandwidth or pulse broadening may be normalized to a unit length, such as GHz · km, or ns/km.

Two methods are available for determining transmission capacity of multimode fibers:

- the frequency domain measurement method in which the baseband frequency response is directly measured in the frequency domain by determining the fiber response to a sinusoidally modulated light source; or
- the optical time domain measurement method (pulse distortion) in which the baseband response is measured by observing the broadening of a narrow pulse of light.

Differential Mode Delay for Multimode Fibers

Differential mode delay (DMD) characterizes the modal structure of a graded-index glass-core multimode fiber. DMD is useful for assessing the bandwidth performance of a fiber when used with short-pulse, narrow spectral-width laser sources.

The output from a singlemode probe fiber excites the multimode FUT at the test wavelength. The probe spot is scanned across the FUT endface, and the optical pulse delay is determined at specified radial offset positions between an inner and an outer limit. DMD is the difference in optical pulse delay time between the FUT fastest and slowest modes excited for all radial offset positions between and including the inner and the outer limits.

The related critical issues influencing DMD are the temporal width of the optical pulse, jitter in the timing, the finite bandwidth of the optical detector, and the mode broadening due to the source spectral width and the FUT chromatic dispersion.

The test method is commonly used in production and research facilities, but is not easily accomplished in the field. DMD may be a good predictor of the source launching conditions. DMD may be normalized to a unit length, such as ps/m.

Chromatic Dispersion

Chromatic dispersion in a singlemode fiber is a combination of material dispersion and waveguide dispersion (see Fig. 8), and it contributes to pulse broadening and distortion in a digital signal.

Material dispersion is produced by the dopants used in glass and is important in all fiber types. Waveguide dispersion is produced by the wavelength dependence of the index of refraction and is critical in singlemode fibers only.

From the point of view of the transmitter, this is due to two causes:

- The presence of wavelengths in the source optical spectrum. Each wavelength has a different phase delay and group delay (different group velocities) along the fiber, because they travel under different index of refraction (or phase) as the index varies as a function of wavelengths, as shown in Fig. 9.
- The other cause is the modulation of the source, which itself has two effects:
  - As bit-rates increase, the spectral width of the modulated signal increases and can become comparable to or exceed the spectral width of the source.
  - Chirp occurs when the source wavelength spectrum varies during the pulse. By convention, positive chirp at the transmitter occurs when the spectrum during the rise/fall of the pulse shifts towards shorter/longer wavelengths respectively. For a positive fiber dispersion coefficient, longer wavelengths are delayed relative to shorter wavelengths. Hence if the sign of the product of chirp and dispersion is positive, the two processes combine to produce pulse broadening. If the product is negative, pulse compression can occur over an initial fiber length until the pulse reaches a minimum width and then broadens again with increasing dispersion.

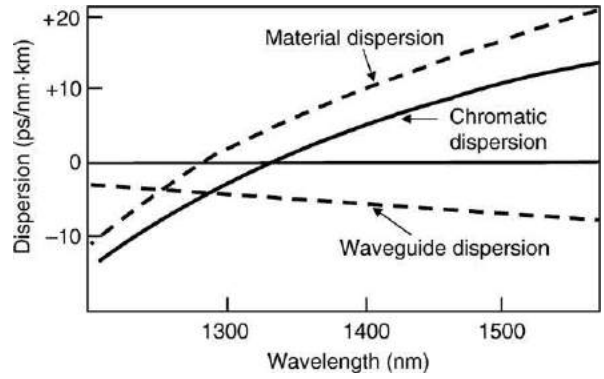


Fig. 8 Contribution of the material and waveguide dispersions to the chromatic dispersion.

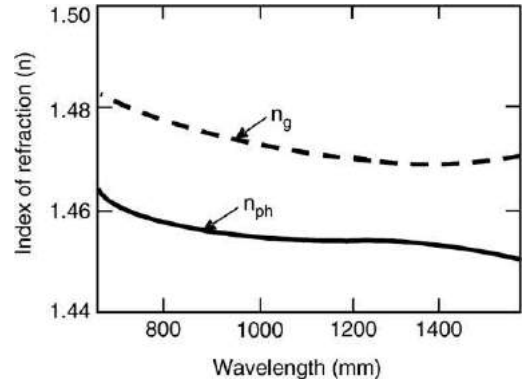


Fig. 9 Difference between the phase and the group index of refraction.

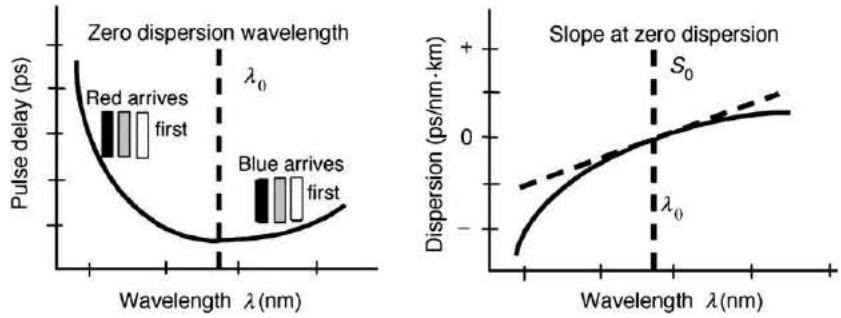


Fig. 10 Relation between the pulse (group) delay and the (chromatic) dispersion.

The electric field propagating into the FUT may be simply described as follows:

$$E(t,z) = E_0 \sin(\omega t - \beta z) \tag{8}$$

$\omega = 2\ \pi c/\lambda$ [rad/s] is the angular frequency; $\beta = kn = (\omega/c)n$ is the effective index; as β has units of m^{-1} it is often referred to as wavenumber or even sometimes propagation constant; k is the propagation constant.

The group delay τ_g is then given by:

$$d\beta/d\omega = \beta_1 = \tau_g \tag{9}$$

An example of the group delay spectral distribution is shown in Fig. 10. Assuming n is not complex; β_1 is the first-order derivative of β.

The group velocity v_g is given by

$$v_g = (d\beta/d\omega)^{-1} = -(\lambda^2/2\pi c)(d\beta/d\lambda)^{-1} \tag{10}$$

The FUT input–output delay is given by

$$\tau_d = L/v_g \tag{11}$$

L is the FUT length.

The group index of refraction n_g is given by

$$n_g = c/v_g = n - \lambda(dn/d\lambda) \tag{12}$$

The dispersion parameter or dispersion coefficient D (ps/nm · km) is given by

$$\begin{aligned} D &= -(\omega/\lambda)(d\tau_g/d\omega) = -(2\pi c/\lambda^2)(d^2\beta/d\omega^2) \\ &= -(\lambda/c)(d^2n/d\lambda^2) \end{aligned} \tag{13}$$

$$d^2\beta/d\omega^2 = \beta_2 \tag{14}$$

An example of the spectral distribution of D obtained from the group delay is shown in Fig. 10.

β_2 (ps^2/km) is the group velocity dispersion parameter, so D may be related to β_2 as follows:

$$D = -(\omega/\lambda)\beta_2 \tag{15}$$

When β_2 is positive then D is negative and vice-versa. The region where β_2 is positive is called normal dispersion while the negative-β_2 region is called anomalous dispersion.

At λ_0, β_1 is minimum, and $\beta_2 = 0$, then $D = 0$.

The dispersion slope S (ps/nm^2 · km), also called the differential dispersion parameter or second-order dispersion, is given by

$$S = dD/d\lambda = (\omega/\lambda)\beta_3 + (2\omega/\lambda^2)\beta_2 \tag{16}$$

$$\beta_3 = d\beta_2/d\omega = d^3\beta/d\omega^3$$

At λ_0, β_1 is minimum, $\beta_2 = 0$, then $D = 0$; but S is not zero and depends on β_3. An example of the spectral distribution of S and S_0 is illustrated in Fig. 10.

Overall, the general expression of β is given by

$$\begin{aligned} \beta(\omega) &= \beta_0 + (\omega - \omega_0)\beta_1 + (1/2)(\omega - \omega_0)^2\beta_2 \\ &\quad + (1/12)(\omega - \omega_0)^3\beta_3 + \cdots \end{aligned} \tag{17}$$

Fig. 11 illustrates the difference between dispersion unshifted fiber (ITU-T Rec. G.652), dispersion shifted fiber (ITU-T Rec. G.653) and nonzero dispersion shifted fiber (ITU-T Rec. G.655).

Test methods for the determination of chromatic dispersion

All methods measure the group delay at a specific wavelength over a range and use agreed fitting functions to evaluate λ_0 and S_0.

In the phase shift method, the group delay is measured in the frequency domain, by detecting, recording, and processing the phase shift of a sinusoidal modulating signal between a reference, a secondary fiber path, and the channel signal.

Setup variances exist and some do not require the secondary reference path. For instance, by using a reference optical filter at the FUT output it is possible to completely decouple the source from the phasemeter. With such an approach, chromatic dispersion may now be measured in the field over very long links using optical amplifiers (see Fig. 12).

In the differential phase shift method, two detection systems are used together with two wavelength sources at the same time. In this case the chromatic dispersion may be determined directly from the two group delays. This technique usually offers faster and more reliable results but costs much more than the phase-shift technique which is usually preferred.

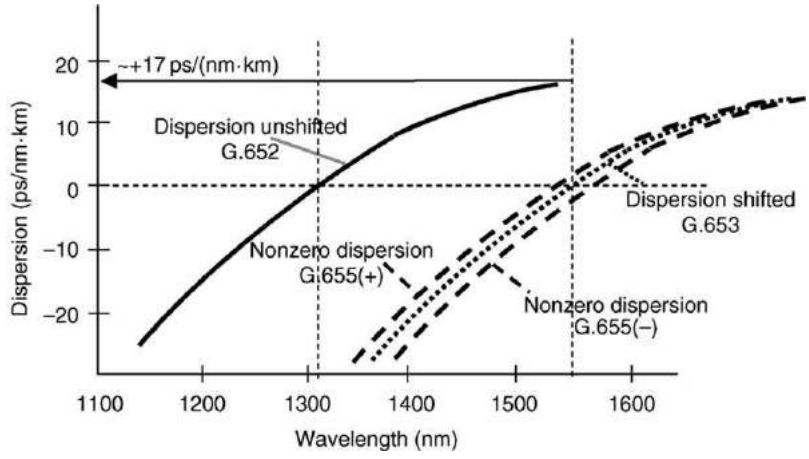


Fig. 11 Chromatic dispersion for various types of fiber.

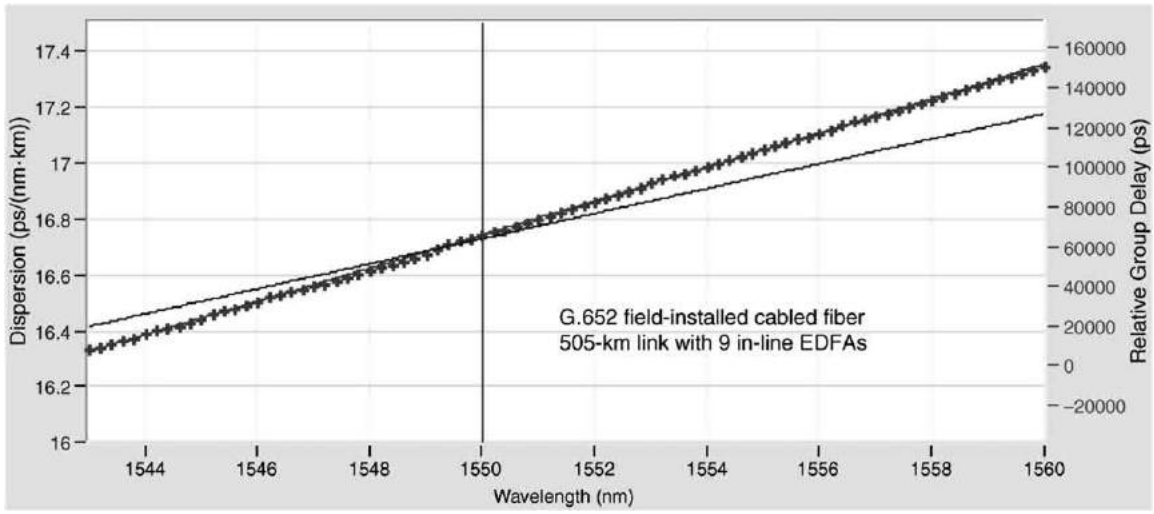


Fig. 12 Test results for field-related chromatic dispersion using the phase-shift technique.

In the interferometric method, the group delay between the FUT and a reference path is measured by a Mach–Zehnder interferometer. The reference delay line may be an air path or a singlemode fiber standard reference material (SRM). The method can be used to determine the following characteristics:

- longitudinal chromatic dispersion homogeneity; and
- effect of overall or local influences, such as temperature changes and macrobending losses.

In the pulse delay method, the group delay is measured in the time domain, by detecting, recording, and processing the delay experienced by pulses at various wavelengths.

Polarization Mode Dispersion

Polarization mode dispersion (PMD) causes an optical pulse to spread in the time domain and may impair the performance of a telecommunications system. The effect can be related to differential phase and group velocities and corresponding arrival time of different polarization components of the pulse signal. For a sufficiently narrowband source, the effect can be related to a differential group delay (DGD), $\Delta\tau$, between a pair of orthogonally polarized principal states of polarization (PSP) at a given wavelength or optical frequency (see **Fig. 13(a)**). In an ideal circular symmetric fiber, the two PSPs propagate with the same velocity. However:

- a real fiber is not perfectly circular;
- the core is not perfectly concentric with the cladding;
- the core may be subjected to microbending;
- the core may present localized clusters of dopants; and
- the environmental conditions may stress the deployed cable and affect the fiber.

Each time the fiber undergoes local stresses and consequently birefringence. These asymmetry characteristics vary randomly along the fiber and in time, lead to a statistical behavior of PMD.

For a deployed cabled fiber at a given time and optical frequency, there always exist two PSPs such that the pulse spreading due to PMD vanishes, if only one PSP is excited. On the contrary, the maximum pulse spread due to PMD occurs when both PSPs are equally excited, and is related to the difference in their group delays, the DGD associated with the two PSPs. For broadband transmission, the DGD statistically varies as a function of wavelengths or frequencies and result in an output pulse that is spread out in the time domain (see **Figs. 13(a)–(c)**). In this case, the spreading can be related to the RMS (root mean square) of DGD values $\langle\Delta\tau^2\rangle^{1/2}$. However, if a known distribution such as the Maxwell distribution may be fit to the DGD distribution probability, then a mean (or average) value of the DGD $\langle\Delta\tau\rangle$ may be correlated to the RMS value and used as a system performance predictor in particular with a maximum value of the DGD distribution associated to a low probability of occurrence. This maximum DGD may then be used to define the quality of service that would tolerate values lower than this maximum DGD.

Test methods for polarization mode dispersion

Three methods are generically used for measuring PMD. Other methods or analyses may exist but they are generally not standardized or are limited in their applications.

- Stokes parameter evaluation (SPE)

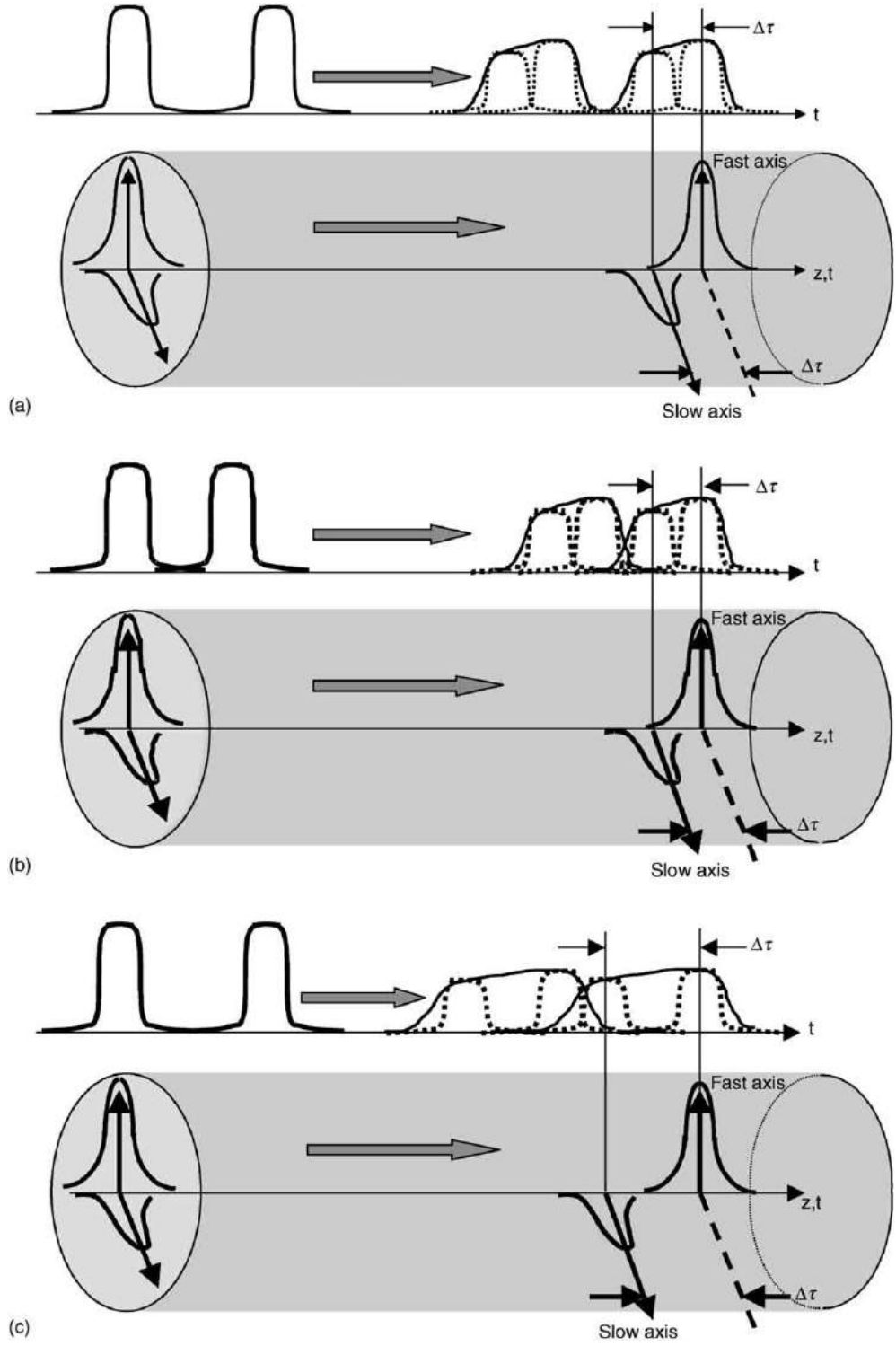


Fig. 13 PMD effect on the pulse broadening and its possible pulse impairment. (a) The pulse is spread by the DGD $\Delta\tau$ but the bit rate is too low to create an impairment. (b) The pulse is spread by the DGD $\Delta\tau$ but the bit rate is high enough to create an impairment. (c) The DGD $\Delta\tau$ is large enough even at low bit rate to make the pulse spreading and creating impairment.

- Jones matrix eigenanalysis (JME)
- Poincaré sphere analysis (PSA)
- Fixed analyzer (FA)

- Extrema counting (EC)
- Fourier transform (FT)
- Interferometry (INTY)
  - Traditional analysis (TINTY)
  - General analysis (GINTY).

All methods use a linearly polarized source at the FUT input and are suitable for laboratory measurements of factory lengths of fiber and cable. However, the interferometric method is the only method appropriate for measurements of cabled fiber that may be moving or vibrating such as is found in the field.

Stokes parameter evaluation

SPE determines PMD by measuring a response to a change of narrowband light (from a tuneable light source with broadband detector – JME, or a broadband source with a filtered detector such as an interferometer – PSA) across a wavelength range. The Stokes vector of the output light is measured for each wavelength. The change of these Stokes vectors with angular optical frequency (wavelength), ω and with the change in input SOP (state of polarization), yields the DGD as a function of wavelength.

For both JME and PSA analyses, three distinct and known linear SOPs (orthogonal on the Poincaré sphere) must be launched for each wavelength. **Fig. 14** illustrates the test setup and examples of test results.

The JME and PSA method are mathematically equivalent.

Fixed analyzer

FA determines PMD by measuring a response to a change of narrowband light across a wavelength range. For each SOP, the change in output power that is filtered through a fixed polarization analyzer, relative to the power detected without the analyzer, is measured as a function of wavelength. **Fig. 15** illustrates a test setup and examples of test results.

The resulting measured function can be analyzed in one of two ways:

- by counting the number of peaks and valleys (EC) of the curve and application of a formula. This analysis is considered as a frequency domain approach; and
- by taking the Fourier transform (FT) of the measured function. This FT is equivalent to the pulse spreading obtained by TINTY.

Interferometry

INTY uses a broadband light source and an interferometer. The fringe pattern containing the source auto-correlation, together with the PMD related cross-correlation of the emerging electromagnetic field, is determined by the interference pattern of the output light, i.e., the interferogram. The PMD determination for the wavelength range associated with the source spectrum is based on the envelope of the fringe pattern interferogram. Two analyses are available to obtain the PMD:

- TINTY uses a set of specific operating conditions for its successful applications and a basic setup; and
- GINTY uses no limiting operating conditions, but in addition to the same basic setup, also using a modified setup compared to TINTY.

Fig. 16 illustrates the test setup for both approaches and examples of test results.

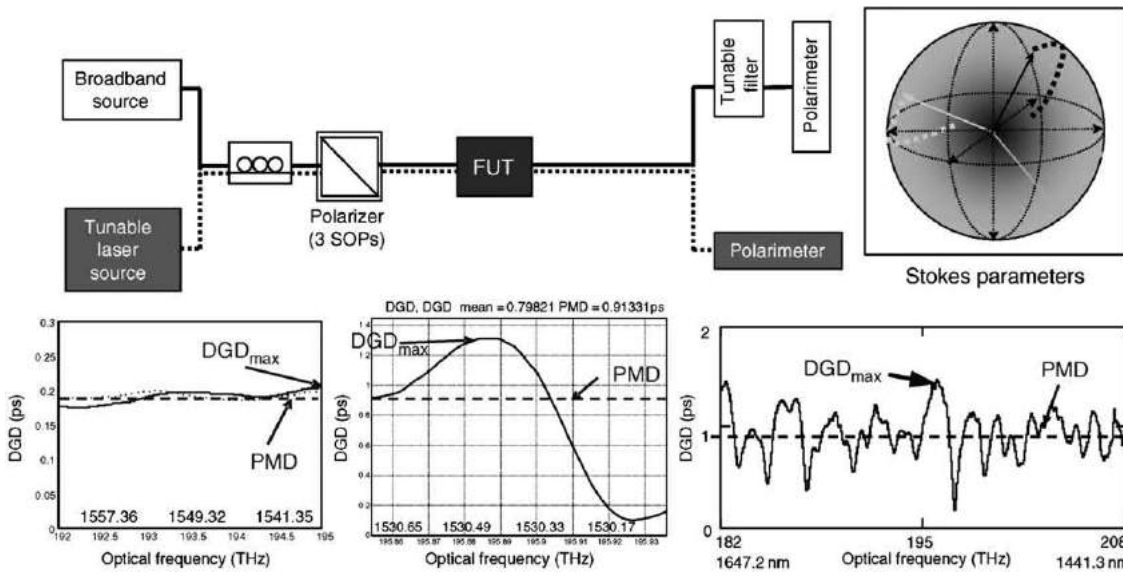


Fig. 14 PMD by Stokes parameter evaluation method.

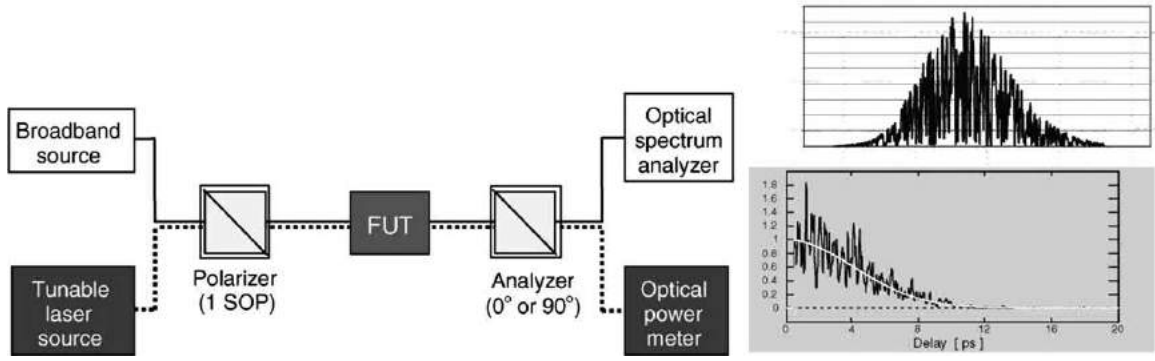


Fig. 15 PMD by fixed analyzer method.

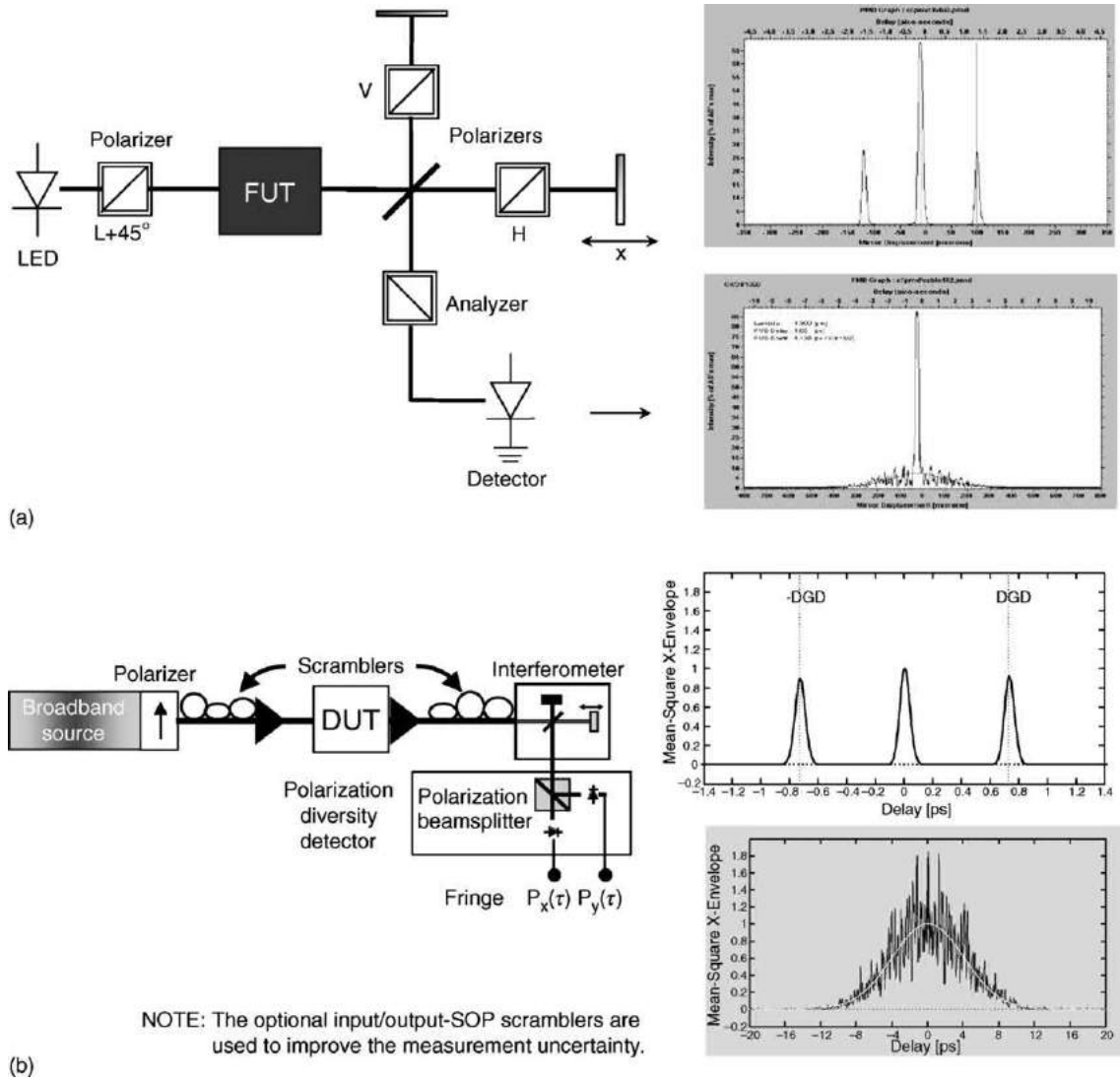


Fig. 16 PMD by the interferometric method. (a) TINTY, (b) GINTY.

Polarization Crosstalk

Polarization crosstalk is a characteristic of energy mixing/transfer/coupling between the two PSPs in a PMF (polarization maintaining fiber) when their isolation is imperfect. It is the measure of the strength of mode coupling or output power ratio between the PSPs within a PMF.

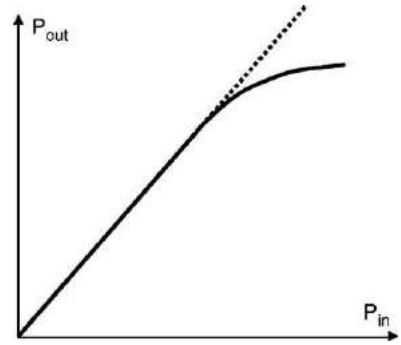


Fig. 17 Power output-to-power input relationship for production of nonlinear effects.

A PMF is an optical fiber capable of transmitting, under external perturbations, such as bending or lateral pressures, both HE_{11}^{x} and HE_{11}^{y} polarization modes whose electric field vector directions are orthogonally to each other and which have different propagation constants β_x and β_y.

Two methods are available for measuring the polarization crosstalk of PMF:

- *power ratio method*, which uses the maximum and minimum values of output power at a specified wavelength, and is applicable to fibers and connectors jointed to a PMF, and to two or more PMFs joined in series; and
- *in-line method*, which uses an analysis of the Poincaré sphere, and is applicable to single or cascaded sections of PMF, and to PMF interconnected with optical devices.

Nonlinear Effects

When the power of the transmission signal is increased to achieve longer span lengths at high bit rates the interaction between the signal and the fiber medium creates nonlinearity. The output signal from the fiber does not increase linearly anymore (as shown in Fig. 17) and part of the signal is lost and will appear in another part of the spectrum.

This is a key issue both in high capacity systems and in long unregenerated links. These nonlinear effects can be generally categorized as follows:

- Scattering phenomena:
  - Stimulated Brillouin scattering (SBS); and
  - Stimulated Raman scattering (SRS).
- Kerr-effect related phenomena:
  - Self-phase modulation (SPM);
  - Cross-phase modulation (XPM);
  - Modulation instability;
  - Soliton formation; and
  - Four wave mixing (FWM).

The Kerr effect is related to the intensity dependence of the index of refraction.

The nonlinear effects are influenced by a number of parameters such as:

- Fiber total dispersion;
- Fiber effective area;
- Channel spacing in WDM systems;
- Unregenerated link distance;
- Degree of longitudinal uniformity of fiber characteristics; and
- Signal intensity and source linewidth.

Nonlinear coefficient (n_2/A_{eff})

Starting from a particularly intense field, the fiber index of refraction becomes dependent on optical intensity inside the fiber. The new index can be expressed as follows:

$$n = n_0 + n_2 I \tag{18}$$

n is the nonlinearity dependent index; n_0 is the linear part of the index; n_2 is the nonlinear index, also called the Kerr nonlinear index (2.2 to 3.4×10^{-16} cm^2/W); and I is the optical intensity inside the fiber.

The field propagation at a distance L into the fiber is described by the following equation:

$$E_{\text{out}}(L) = E_{\text{in}}(0)\exp[-a/2 + \mathrm{i}\beta + \gamma P(L,t)/2]L \tag{19}$$

$a/2$ is the attenuation; iβ is the phase of the wave; and $\gamma P(L, t)/2$ is the nonlinearity term;

$$\gamma = 2\pi n_2/\lambda A_{\text{eff}} \tag{20}$$

γ is the nonlinearity coefficient and may be a complex number; A_{eff} is the fiber core effective area; $P(L, t)$ is the total power; and λ is the signal wavelength and t the time variable.

The nonlinear coefficient is defined as n_2/A_{eff}. This coefficient plays a critical role in the fiber and is closely related to system performance degradation due to nonlinearities when very high power is used.

Methods for measuring the nonlinear coefficient

Two methods are available for measuring the nonlinear coefficient:

- Continuous-wave dual-frequency (CWDF); and
- Pulsed single-frequency (PSF).

In the CWDF method, light from two wavelengths is injected into the fiber. At higher power, the light from the two wavelengths beat due to the nonlinearity and produce an output spectrum that is spread. The relationship of the power level to a particular spreading is used to calculate the nonlinear coefficient.

In the PSF method, the pulsed light from a single wavelength is injected into the fiber. Very short pulses (<1 ns) and their input peak power must be measured and related to the nonlinear spreading of the output spectrum.

Stimulated Brillouin scattering

In an intensity modulated system using a source with a narrow linewidth, significant optical power is transferred from the forward-propagating signal to a backward-propagating signal when the SBS power threshold is exceeded. At that point periodic regions of index of refraction produce a grating traveling at speed of sound away from the source. The grating reflects backward part of the incident light. The reflected sound waves (acoustic phonons) scatter light back to the source. Phase matching (or momentum conservation) dictates that the scattered light preferentially travels in the backward direction. The scattered light will be Doppler-shifted (downshifted or Brillouin-shifted) by approximately 11 GHz (at 1550 nm, for G.652 fiber). The scattered light has a very narrow spectrum (very coherent) and very close to the carrier signal and may be very detrimental.

Stimulated Raman scattering

SRS is an interaction between light and the fiber molecular vibrations as adjacent atoms vibrate in opposite directions (an 'optical phonon'). Some of the energy of the main carrier (optical pump wave) is transferred to the molecules, thereby further increasing the amplitude of their vibrations. If the vibrations become large enough, a threshold is reached at which the local index of refraction changes. These local changes then scatter light in all directions similar to Rayleigh scattering. However, unlike Rayleigh scattering, the wavelength of the Raman scattered light is shifted to longer wavelengths by an amount that corresponds to the molecular vibration frequencies and the Raman signal spreads over a large spectrum.

Self-phase modulation

SPM is the effect that a powerful pulse has on its own phase, considering that in eqn (18), $I(t)$ varies in time:

- $I(t) \rightarrow n(t) = n_0 + n_2 I(t) \rightarrow$ modulates the phase $\beta(t)$ of the pulse; and
- $\mathrm{d}l/\mathrm{d}t \rightarrow \mathrm{d}n/\mathrm{d}t \rightarrow \mathrm{d}\beta/\mathrm{d}t(\text{chirp}) \rightarrow$ broadening in the frequency domain $\rightarrow$ broadening in the time domain.

$I(t)$ peaks at the center of the pulse (peak power) and consequently increases the index of refraction. A higher index causes the wavelengths in the center of the pulse to accumulate phase more quickly than at the wings of the pulse:

- this causes wavelength stretching (shift to longer wavelengths) at pulse leading edge (risetime); and
- this causes wavelength compression (shift to shorter wavelengths) at pulse trailing edge (falltime).

The pulse will then broaden with negative (normal) dispersion and shorten with positive (anomalous) dispersion. SPM may then be used for dispersion compensation considering that self-phase modulation imposes $C > 0$ (positive chirp). It can cancel the dispersion if properly manage as a function of the sign of the dispersion. SPM is one of the most critical nonlinear effects for the propagation of soliton or very short pulses over very long distance.

Cross-phase modulation

XPM is the effect that a powerful pulse has on the phase of an adjacent pulse from another WDM system channel traveling in phase or at slightly the same group velocity. It concerns spectral interference between two WDM channels:

- the increasing $I(t)$ at the leading edge of the interfering pulse shifts the other pulse to longer wavelength; and
- decreasing $I(t)$ at the trailing edge of the interfering pulse shifts the other pulse to shorter wavelengths.

This produces spectral broadening, which dispersion converts to temporal broadening depending on the sign of the dispersion. XPM effect is similar to SPM except it depends on the channel count.

Table 1 Fiber dimensional characteristics

Attribute	*Measured parameter*
Fiber geometry	Core/cladding diameter Core/cladding noncircularity Core-cladding concentricity error
Numerical aperture	
Mode field diameter	
Coating geometry	
Length	

Four-wave (four-photon) mixing

FWM is the by-product production effect from two or more WDM channels. For two channels $I(t)$ modulates the phase of each signal (ω_1 and ω_2). An intensity modulation appears at the beat frequency $\omega_1 - \omega_2$.

Two sideband frequencies are created in a similar way as harmonics generation. New wavelengths are created in a number equal to $N^2(N-1)/2$, where N=number of original wavelengths.

Fiber Dimension Characteristics and Corresponding Tests Methods

Table 1 provides a list of the various fiber dimensional characteristics and their corresponding test methods.

Fiber Geometry Characteristics

The fiber geometry is related to the core and cladding characteristics.

Core

The core center is the center of a circle which best fits the points at a constant level in the near-field intensity profile emitted from the central region of the fiber, using wavelengths above and/or below the cut-off wavelength.

The RIP can be measured by refracted near field (RNF) or transverse interferometry techniques and transmitted near field (TNF).

The core concentricity error is the distance between the core center and the cladding center. This definition applies very well for multimode fibers. The distance between the center of the near field profile and the center of the cladding is also used for singlemode fibers.

The mode field diameter (MFD) represents a measure of the transverse electromagnetic field intensity of the mode in a fiber cross-section and it is defined from the far-field intensity distribution.

The MF is the singlemode field distribution of the LP_{01} mode, giving rise to a spatial intensity distribution in the fiber.

The MF concentricity error is the distance between the MF center and the cladding center.

The core noncircularity is a measure of the core ellipticity. This parameter is one of the causes for creating birefringence in the fiber and consequently PMD.

Cladding

The cladding is the outermost region of constant refractive index in the fiber cross-section.

The cladding center is the center of a circle best fitting the outer limit (boundary) of the cladding.

The cladding diameter is the diameter of the circle defining the cladding center.

The cladding noncircularity is a measure of the difference between the diameters of the two circles defined by the cladding tolerance field divided by the nominal cladding diameter.

Coating

The primary coating is one or more layers of protective material applied to the cladding during or after the drawing process to protect the cladding surface (e.g., a 250 μm protective coating). The secondary coating is one or more layers of protective material applied over the primary coating in order to give additional protection or to provide a particular structure.

Measurement of the fiber geometrical attributes

The fiber geometry is measured by the following methods:

- TNF;
- RNF;
- Side-view technique/transverse interference;
- TNF image technique; and
- Mechanical diameter.

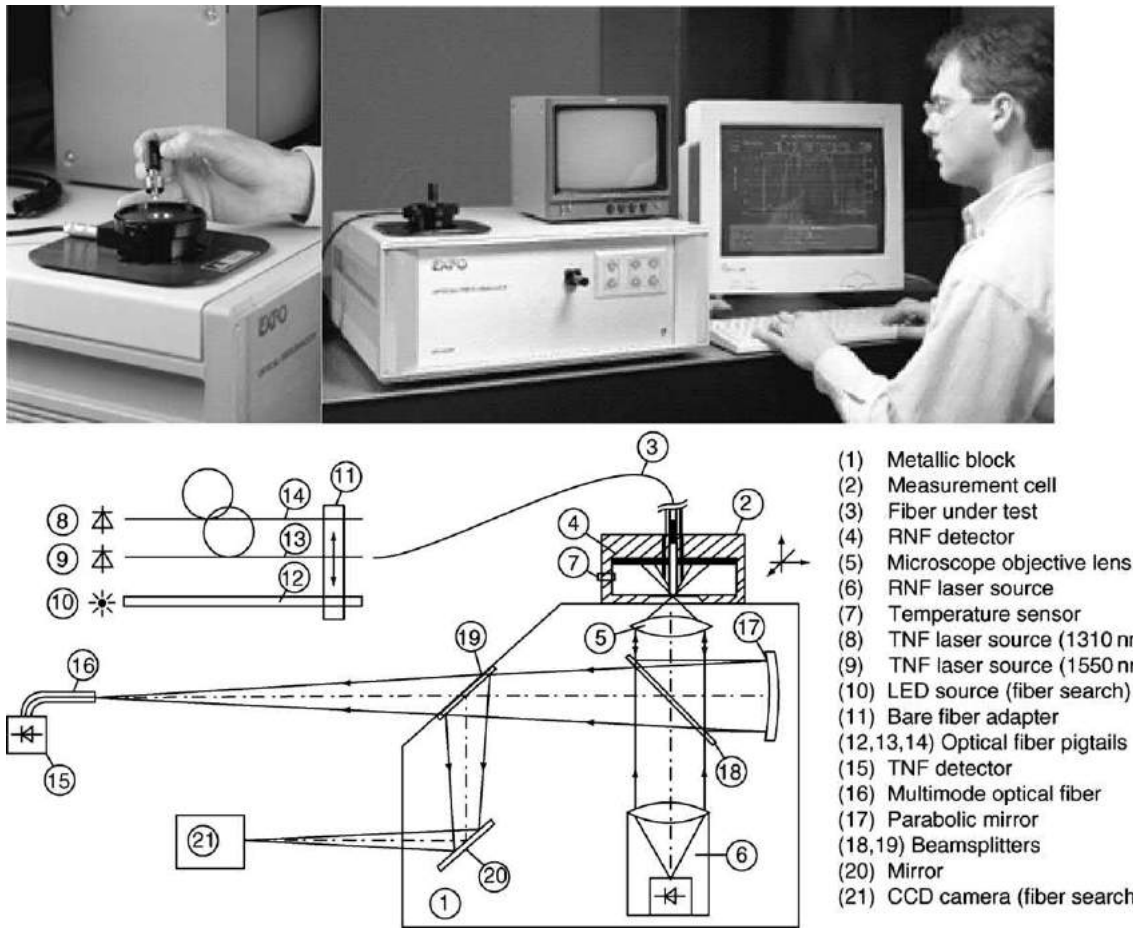


Fig. 18 RNF/TNF combined instrumentation.

Test instrumentation may incorporate two or more methods such as the one shown in Fig. 18.

Transmitted near-field technique

The cladding diameter, core concentricity error, and cladding noncircularity are determined from the near-field intensity distribution. Fig. 19 provides a series of examples of test results from TNF measurements.

Refracted near-field technique

The RIP across the entire fiber (core and cladding) can be directly obtained from the RNF measurement, as shown in Fig. 20.

The geometrical characteristics of the fiber can be obtained from the refractive index distribution using suitable algorithms:

- core/cladding diameter;
- core/cladding concentricity error;
- core/cladding noncircularity;
- maximum numerical aperture (NA); and
- index and relative index of refraction difference.

Fig. 21 illustrates the core geometry.

Side-view technique/transverse interference

The side-view method is applied to singlemode fibers to determine the core concentricity error, cladding diameter and cladding noncircularity by measuring the intensity distribution of light that is refracted inside the fiber. The method is based on an interference microscope focused on the side view of an FUT illuminated perpendicular to the FUT axis. The fringe pattern is used to determine the RIP.

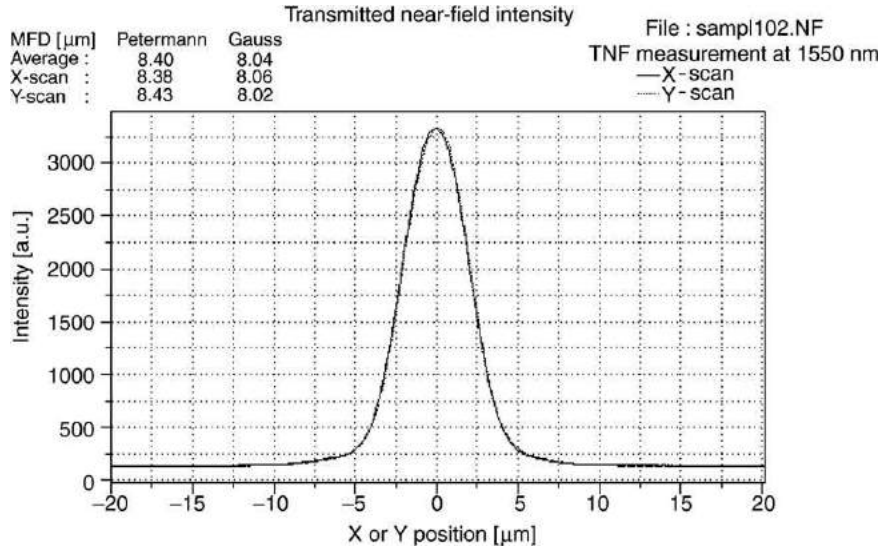


Fig. 19 MFD measurement by TNF.

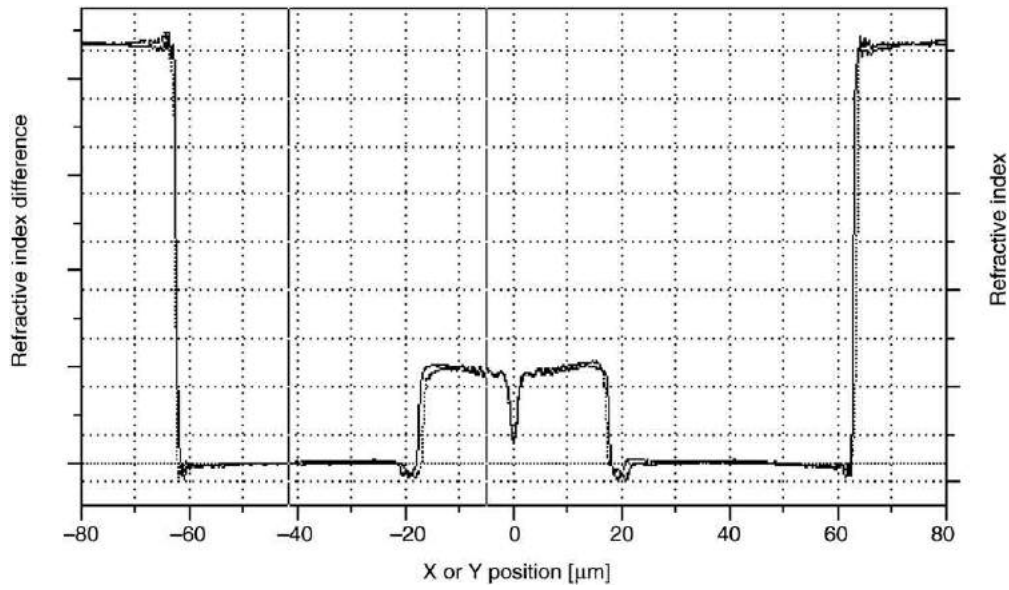


Fig. 20 RIP from RNF measurement.

TNF image technique

The TNF image technique, also called near-field light distribution, is used for the measurement of the geometrical characteristics of singlemode fibers. The measurement is based on analysis of magnified images at the FUT output. Two subsets of the method are available:

- grey-scale technique which performs an x–y near-field scan using a video system; and
- Single near-field scanning technique performing a one-dimensional scan.

Mechanical diameter

This is a precision mechanical diameter measurement technique used to accurately determine the cladding diameter of silica fibers. The technique uses an electronic micrometer such as based on a double-pass Michelson interferometer. The technique is used for providing calibrated fibers to the industry as SRM.

Numerical Aperture

The NA is an important attribute for multimode fibers in order to predict their launching efficiency, joint loss at splices and micro/macrobending characteristics.

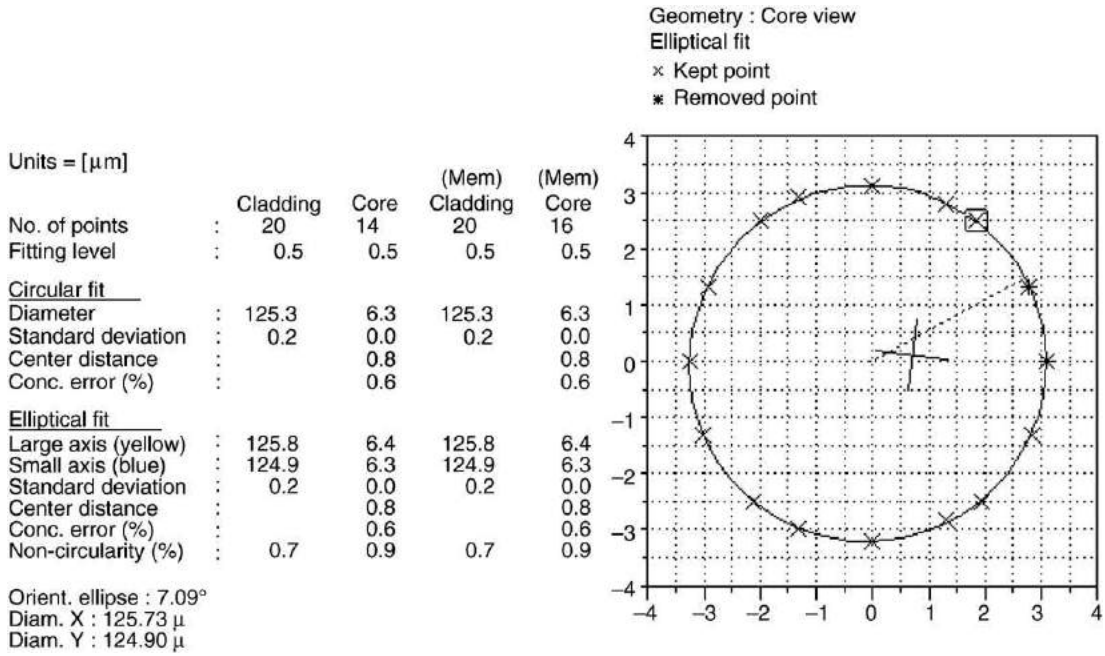


Fig. 21 Core geometry by RNF measurement.

A method is available for the measurement of the angular radiant intensity distribution (far-field) distribution or the RIP at the output of an FUT. NA can be determined from the analysis of the test results.

Mode Field Diameter

The definition of the MFD is given in the section describing the core center above. Four measurement methods are available:

- direct far-field scan determines MFD from the far-field intensity distribution;
- variable aperture in far field determines MFD from the complementary aperture transmission function $a(x)$, $x = d \tan \theta$ being the aperture radius and d the distance between the aperture and the FUT:

$$\mathrm{MFD} = \frac{\lambda}{\pi d}\left[\int_0^\infty a(x)\frac{x}{(x^2+d^2)^2}\,\mathrm{d}x\right]^{-1/2} \tag{21}$$

Eq. (21) is valid for small-θ approximation.

- near-field scan determines MFD from the near-field intensity distribution I_{NF}, r being the radial coordinate:

$$\mathrm{MFD} = 2\left[2\frac{\int_0^\infty rI_{\mathrm{NF}}(r)\,\mathrm{d}r}{\int_0^\infty r\left[\frac{\mathrm{d}I^{1/2}(r)}{\mathrm{d}r}\right]^2\mathrm{d}r}\right]^{1/2} \tag{22}$$

Eq. (22) is valid for small-θ approximation; and

- bidirectional backscattering uses an OTDR and bidirectional measurements to determine MFD by comparing the FUT results with a reference fiber.

Effective Area

A_{eff} is a critical nonlinearity parameter and is defined as follows:

$$A_{\mathrm{eff}} = \frac{2\pi\left[\int_0^\infty I(r)r\,\mathrm{d}r\right]^2}{\int_0^\infty I(r)^2 r\,\mathrm{d}r} \tag{23}$$

$I(r)$ is the field intensity distribution of the fiber fundamental mode at radius r. The integration in the equation is carried out over the entire fiber cross-section. For a Gaussian approximation:

$$I(r) = \exp - 2\left(\frac{2r}{\text{MED}}\right)^2 \tag{24}$$

which yields:

$$A_{\text{eff}} = \frac{\pi}{4}\text{MFD}^2 \tag{25}$$

Three methods are available for the measurement of A_{eff}:

- direct far-field;
- variable aperture in far-field; and
- near-field.

Mechanical Measurement and Test Methods

Table 2 describes the various mechanical characteristics and their corresponding test methods.

Proof Stressing

The proof stress level is the value of tensile stress or strain applied to a full fiber length over a short period of time. The method for fiber proof stressing is the longitudinal tension which describes procedures for applying tensile loads to a length of fiber. The fiber stress is calculated from the applied tension. The tensile load is applied over short periods of time but not too short in order for the fiber to experience proof stress.

Residual Stress

Residual stress is the stress built up by the thermal expansion difference between core and cladding during the fiber drawing process or splicing. Methods are available for measuring residual stress based on polarization effects. A light beam produced by a rotating polarizer propagates to the x-axis direction while the beam polarization is in the $y - z$ plane and the fiber longitudinal axis in the z-axis. The light experiences different phase shift in the y- and z-axis due to the FUT birefringence. The photoelastic effect gives the relationship between this phase change and residual stresses.

Stress Corrosion Susceptibility

The stress corrosion susceptibility is related to the dependence of crack growth on applied stress. It depends on the environmental conditions and static and dynamic values may be observed.

Environmental Characteristics

Table 3 lists the fiber characteristics related to the effect of the environment.

Hydrogen Aging for Low-Water-Peak Single-Mode Fiber

Hydrogen aging on low-water peak fibers, such as G.652.C, is based on a test performed at 1.0 atmosphere of hydrogen pressure at room temperature over a period of one month. Other proportional combinations are possible.

Table 2 Fiber mechanical characteristics

Attribute
Proof stress
Residual stress
Stress corrosion susceptibility
Tensile strength
Stripability
Fiber curl

Table 3 Fiber environmental characteristics

Attribute
Hydrogen aging
Nuclear gamma irradiation
Damp heat
Dry heat
Temperature cycling
Water immersion

Nuclear Gamma Irradiation

Nuclear radiation is considered on the basis of a steady state response of optical fibers and cables exposed to gamma radiation and to determine the level of related radiation-induced attenuation produced in singlemode or multimode cabled or uncabled fibers.

The fiber attenuation generally increases when exposed to gamma radiation. This is primarily due to the trapping of radiolytic electrons and holes at defect sites in the glass (i.e., the formation of 'color centers'). Two regimes are considered:

- the low dose rate suitable for estimating the effect of environmental background radiation; and
- the high dose rate suitable for estimating the effect of adverse nuclear environments.

The effects of environmental background radiation are measured by the attenuation (cut-back method). The effects of adverse nuclear environments are tested by power monitoring before, during and after FUT exposure.

Further Reading

Agrawal, G.P., 1997. Fiber-Optic Communication Systems, 2nd edn. New York: John Wiley & Sons.
Agrawal, G.P., 2001. Nonlinear Fiber Optics, 3rd edn. San Diego, CA: Academic Press.
Girard, A. (Ed.), 2000. Guide to WDM Technology and Testing. Quebec: EXFO, p. 194.
Hecht, J., 1999. Understanding Fiber Optics, 3rd edn. Upper Saddle River, NJ: Prentice Hall.
Masson, B., Girard, A. (Eds.), 2004. FTTx PON Guide, Testing Passive Optical Networks. Quebec: EXFO, p. 56.
Miller, J.L., Friedman, E. (Eds.), 2003. Optical Communications Rules of Thumb. New York: McGraw-Hill, p. 428.
Neumann, E.-G., 1988. Single-Mode Fibers, Fundamentals. Berlin: Springer-Verlag.

Optical Fiber Cables

G Galliano, Telecom Italia Lab, Torino, Italy

Introduction

In order to be put to practical use in the telecommunication network, optical fibers must be protected from environmental and mechanical stresses which can change their transmission characteristics and reliability.

The fibers, during manufacturing, are individually protected with a thin plastic layer (extruded during the drawing) to preserve directly their characteristics from damage due to the environment and handling. This coating surrounding the fiber cladding (with an external diameter of 125 μm) is known as the primary coating, which is obtained with a double layer of UV resin for a maximum external diameter of about 250 μm.

Primary coated fiber, at the end of the manufacturing process, is subjected to a dynamic traction test (proof-test) to monitor its mechanical strength. This test, carried out on all fibers manufactured, consists of the application of a well-defined strain (from 0.5% to 1%) for a fixed time (normally 1 s) and thus permits the exclusion of, from successive cabling, those fibers with poor mechanical characteristics.

At the end of the manufacturing process the optical fibers are not directly usable. It is necessary to surround the single fiber (or groups of fibers) with a structure which provides protection from mechanical and chemical hazards and allows for stresses that may be applied during cable manufacture, installation, and service.

Suitable protection is provided using fiber conditioning within the cable (secondary coating or loose cabling), while entire protection is obtained with some cable elements, such as strength members, core assembling, filling, and outer protections. Moreover, the cable structure and its design is heavily influenced by the application (e.g., number of fibers), installation (laying and splicing), and environmental conditions of service (underground, aerial, or submarine).

In this article, the main characteristics of optical cables (with both multimode or single mode fibers) are described and analyzed, together with suitable solutions.

Secondary Fiber Coating

The fiber, before cabling, can be protected with a secondary coating (a tight jacketing) to preserve it during cable stranding. However, fiber with only a primary coating may be cabled with a suitable cable protection.

In the first case the fiber can be directly stranded to form a cable core; in the second case, a single group of fibers must be inserted into a loose structure (tubes or grooves) with a proper extra-length. These two solutions are not always distinguishable; in fact, tight fibers inserted into loose tubes or grooves are often used.

The implementation of the secondary coating and the choice of proper materials requires particular attention, in order to avoid microbending effects on the fiber, which can cause degradation of the transmission characteristics. Microbending is caused by the pressure of the fiber on a microrough surface, or by the fiber buckling, due to the contraction of the structure (e.g., secondary coating) containing it. Microbending causes an increase in fiber attenuation at long wavelengths (1300 and 1550 nm) for both single and multimode fibers, but is particularly emphasized in multimode fibers.

Tight Jacket

In a tight protection, the primary coating fiber is surrounded by a second plastic jacket in contact with the fiber. This jacket mainly protects the fiber from lateral stresses that can cause the microbending.

The fiber can be individually protected with a single or double layer of plastic material or in a ribbon structure. In **Fig. 1** four different types of tight jacket are shown. In **Fig. 1(a)** and **(b)** the fibers are singularly protected with a single or double layer up to an external diameter of 0.9 mm. The second type is normally preferred because the external hard plastic gives a good fiber protection and the inner soft plastic avoids microbending effects on the fiber.

In a ribbon structure, from 4 to 16 primary coated fibers are laid close and parallel and covered with plastic. Two types of ribbon are normally classified 'encapsulated' (**Fig. 1(c)**) or 'edge bonded' (**Fig. 1(d)**). In the first case, the fibers are assembled with a common secondary coating (one or two layers), in the second case, an adhesive is used to stick together the fibers with only the primary coating.

Care should be taken in applying tight protection, in particular for a ribbon structure, to avoid an attenuation increase due to the action of the jacket on the fiber (in particular for the thermal contraction of the coating materials). The ribbon structure is designed to permit the simultaneous junction of multiple fibers; for this reason the ribbon geometry must be precision controlled.

 doi:10.1016/B978-0-12-809283-5.00655-8

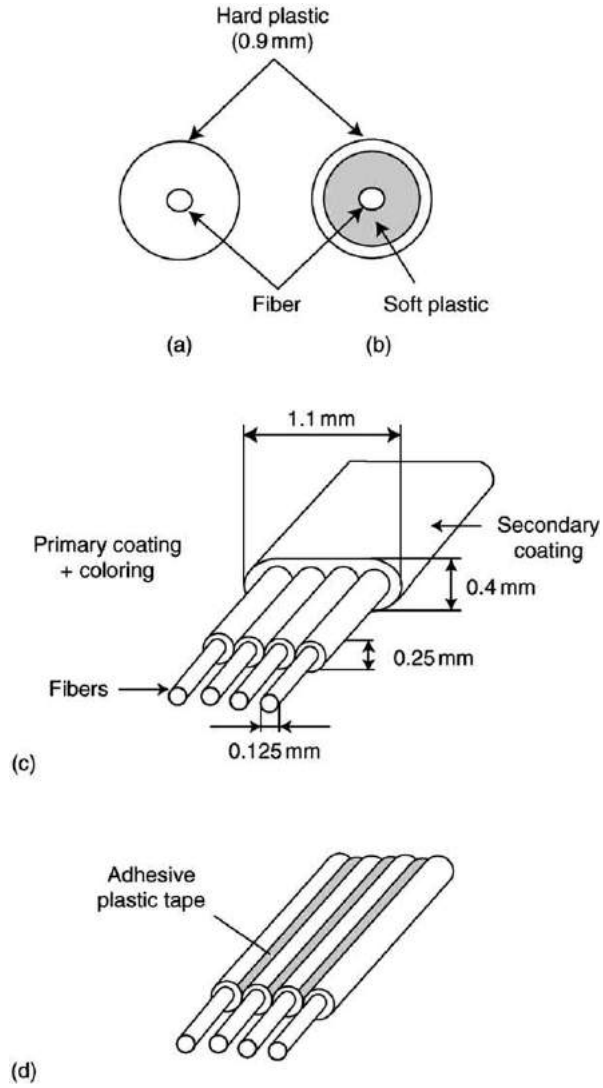


Fig. 1 Different types of tight coatings.

Loose Jacket

In a loose protection the fiber can be inserted into a cable with only the primary coating. A thin layer of colored plastic may be added onto the primary coating to identify the fiber in the cable. The fibers are arranged in a tube or in a groove on a plastic core (slotted core). The diameter of the tube or the dimension of the core must be far larger than the diameter of the fiber so that the fiber is completely free inside the structure. The fibers can be inserted into a tube or into a slotted core, singularly or in groups. The space inside the tube around the fibers can be empty or filled with a suitable jelly. Some examples of loose structures are shown in Fig. 2. In the tube structure, the external diameter of the jacket depends on the number of the fibers: for a single fiber up to 2 mm (Fig. 2(a)), and for a multifiber tube (Fig. 2(b)) up to 3 mm (for 10 fibers).

Examples of slotted core structures are shown in Fig. 2(c) and (d), respectively, for one fiber and for a group of fibers, the dimensions of grooves depending on the number of fibers.

Plastic materials with a high Young modulus are normally used for the tube that is extruded on the single or on the group of fibers. Instead, for the slotted core, plastic materials with good extrusion are used to obtain a precise profile of the grooves.

The loose structure guarantees freedom to the fibers in a defined interval of elongation and compression. Out of this interval, microbending effects can arise which increase fiber attenuation. Axial compression generally occurs when the cable is subjected to low temperatures and is generated from the thermal contraction of the cable materials that is different from that of the fibers. Axial elongation normally occurs when the cable is pulled or by the effects of high temperature.

For a correct design of a cable with a loose structure, the extra length of the fiber with respect to the tube or groove dimension and the stranding pitch must be evaluated, starting from the range of axial compression and elongation. In fact, care must be taken to prevent the fiber from being strained or forced against the tube or groove walls in the designed range.

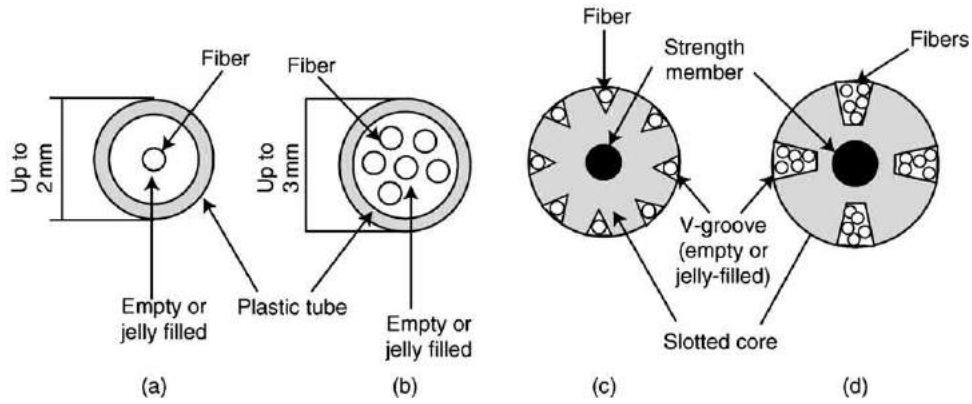


Fig. 2 Different types of loose coatings.

For the tube solution, the extra length of the fibers is controlled during the tube extrusion; for the slotted core solution, the length is controlled during cabling, giving a controlled tension to the slotted core. The loose structure is often used for the cabling of fibers with tight jacketing, joining the advantages of the tight protection and loose cable. This solution is typical for ribbon cables.

Fiber Identification

To identify the fibers in a cable, a proper coloration with defined color codes must be given to the single fibers. In the tight protection, the tight buffer is directly colored during extrusion. In the loose protection, a thin colored layer is added on the primary coating of the fiber; tubes or slotted core may be also colored to identify groups of fibers in high potentiality cables.

Cable Components

The fundamental components of an optical cable are: the cable core with the optical fibers, the strength member (one or more) which provide the mechanical strength of the cable, sheaths which provide the external protection, and filling materials.

Cable Core

The optical fibers with the secondary coating (tight or loose) are rejoined together in a cable core. For tight fibers or loose tubes, the cable core is obtained by stranding the fibers or the tubes around central elements that normally act as strength members too (see below).

The stranding pitch must be long enough to avoid excessive bending of the fibers and short enough to guarantee stress apportionment among the fibers in cable bends and maximum elongation of the cable without fiber strain. Normally, a helical stranding is used, but often SO stranding, that consists in an inversion of stranding direction every three of more turns, is employed. The second stranding type makes the cabling easier but may reduce its performance.

The unit core or the slotted core are made of plastic and normally incorporate a central strength member.

Strength Member

The strength member may be defined as an element that provides the mechanical strength of the cable, to withstand elongation and contraction during the installation and to ensure thermal stability. Fiber elongation, when the cable is pulled or the fiber compressed due to low temperatures, depends greatly on the characteristics of the strength member inserted in the cable.

As the maximum elongation values suggested for the fibers are in the range from 20 to 30% of the proof-test value, the maximum elongation allowed for a tight cable is $\sim$0.1% instead of $\sim$0.3% for a loose cable. To warrant these performances, a material with a high Young modulus is normally used as a strength member. Also, the strength member must be light (to avoid excessive cable weight) and flexible: these properties make it easier when laying the cable in ducts.

The strength member may be metallic or nonmetallic and may be inserted in the core (Fig. 3(a)) or in the outer part of the cable (Fig. 3(b)), normally under the external sheath. Often two strength members are placed in an optical cable: the first one in the core and the second one under the sheath (Fig. 3(c)).

Metallic strength members, generally one or more wires of steel, are inserted in the center or in the outer part of the cable. The steel, for its low thermal expansion coefficient and high Young modulus, behaves as a good strength member, particularly when it is arranged in the central core. Steel is a nonexpensive material, but needs protection against corrosion and electrical-induced voltages, such as by lightning or current flow in the ground.

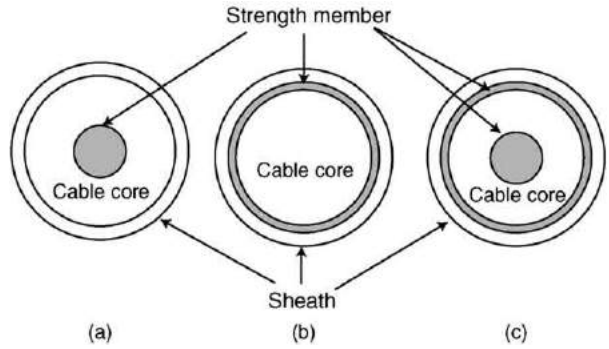


Fig. 3 Different locations of the strength member in an optical cable.

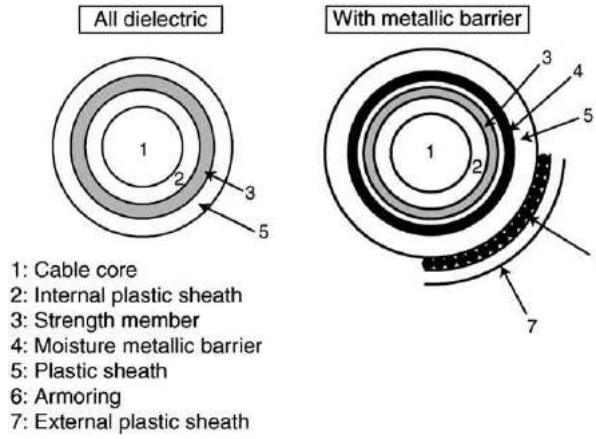


Fig. 4 Examples of optical cable sheaths.

If cables with high flexibility are requested, steel is not suitable and aramid fiber yarn or glass-fiber reinforced plastic (G-FRP) are used, particularly when the strength member is placed in the outer part of the cable. Glass fiber elements are generally used as a compressive strength member in the central cable core to avoid contraction at low temperatures, but they may be used as strength elements when the traction strength is low.

Aramid yarns are not active as resisting compression elements but are normally arranged in the outer part of the optical cable. Aramid yarns have a high Young modulus (as with steel) and low specific weight: in commercial form they consist of a large number of filaments gathered into a layer. Normally two layers stranded in opposite directions are used to avoid cable torsion during the installation.

Metal-free cables are protected by G-FRP and aramid yarns. This solution results in optical cables which are insensitive to electrical interferences.

Sheaths

The functions of the sheaths are to protect an optical cable during its life, in particular from mechanical and environmental stresses. During the installation or when the cable is in service, it may be subjected to mechanical stresses, for example, presence of humidity or water and chemical agents. The sheaths act as a barrier to avoid degradations of the fibers in the internal cable core.

The plastic sheaths (one or more, depending on the cable type and structure) are extruded over the core. This extrusion is a delicate operation in cable manufacturing because the cable core may be subjected to tension, while during the cooling the thermal compression of the plastic produces a stress on all the cable elements. Examples of cable sheaths are shown in Fig. 4.

Different plastic materials, such as polyvinylchloride (PVC), polyethylene, and polyurethane are normally used. PVC has excellent mechanical properties, flexibility, and is flame retardant but is permeable to humidity. On the other hand, high-density polyethylene has a permeability lower than PVC, with good mechanical properties and low coefficient of friction, but it is more flammable than PVC and less flexible. Finally, polyurethane, for its softness and high flexibility, is often used for inner sheaths. When low toxicity is requested (e.g., in indoor cables), halogen-free materials are employed.

Often a metallic barrier is used under the external sheath to prevent moisture entering the cable core. The type most widely used is an aluminum ribbon (of a few tenths of a millimeter) bonded into the external polyethylene sheath, during the sheath extrusion. This structure is normally indicated as a LAP (laminated aluminum polyethylene) sheath.

When the cable is directly buried, a metallic armor (corrugated steel tape or armoring wires) is generally inserted. The metallic armor protects the cable core against radial stresses but it is also used as protection against rodents and insects.

Filling Compounds and Other Components

Fillings are used in optical cables to avoid the presence of moisture and water propagation in the cable core should a failure occur in the cable sheath. Suitable jelly compounds, with constant viscosity at a wide range of temperatures and chemical stabilities, are generally inserted in the cable core and, sometimes, in the outer protection layer. In loose structures, the jelly is in direct contact with the fibers: the compound must be compatible with them and guarantee the fiber freedom. Special compounds may be used to adsorb and permanently fix any hydrogen present in the cable core. The presence of hydrogen gas, that may be generated from the internal materials of the cable (mainly metals), is harmful to the fibers because it causes permanent attenuation increases and thus damage to them.

With the generic terminology 'other components', some other components used in the optical cables are included. They are often similar to those used in conventional cables. The most important are: insulated conductors, plastic fillers, and cushion layers and tapes around the fiber core.

Insulated copper conductors may be incorporated in the cable core, for example, to carry power supply or as service channels, or to detect the presence of moisture in the cable (in this case, the insulation is missing or partial). The conductors may cause problems due to the voltages induced by power lines or lighting.

Plastic fillers are often used to complete the cable core geometry. Cushion layers are used to protect the cable core against radial compression. Normally, they are based on plastic materials wound around the core. Tapes are wound around the cable core to hold the assemblies together and secondly to provide a heat barrier during the sheath extrusion. Mylar tapes are usually employed.

Optical Cable Structures, Types, and Applications

Cable Structures

The structure of an optical cable is strictly dependent on the construction method and the application type that determines the design and the grade of external protection. The main core (or inner) structures of an optical cable can be classified as: stranded structures (tight and loose); slotted core cable; or ribbon cable. In this section, a few examples of cable structures are presented. Below, separate sections are devoted to submarine cables and to special application cables, for their particular structures and features.

In the stranded tight structure, a low number of tight or loose fibers is stranded around a strength member to form a cable unit. Some of this cable unit may be joined to constitute a cable with a medium/high number of fibers (Fig. 5(a)). Alternatively the fibers may be arranged in a multilayer structure (Fig. 5(b)).

The main advantages of a tight stranded structure are the small dimensions of the cable core, even when a large number of fibers are involved, and also the facility of handling. Care must be taken in the design of the strength member: in fact, the fibers are immediately subjected to tension when the cable is pulled into the ducts or compressed when subjected to low temperatures.

In the stranded loose structures, two categories of cables are defined, one fiber per tube and a group of fibers (up to 20) per tube (Fig. 5(c)). As already mentioned, in the loose structures, the fibers are free inside the tube and for this reason the cable is less critical than the tight cable. The external dimensions of loose cables are larger than tight cables, but the solution with multiform tubes allows for cables with a high number of fibers in relatively small dimensions.

Slotted core cables (Fig. 2(c) and (d)) contain fibers with only the primary coating and may be divided, as in the stranded loose structure, into one fiber per groove or into a group of fibers per groove. The cable design and the choice of a suitable plastic

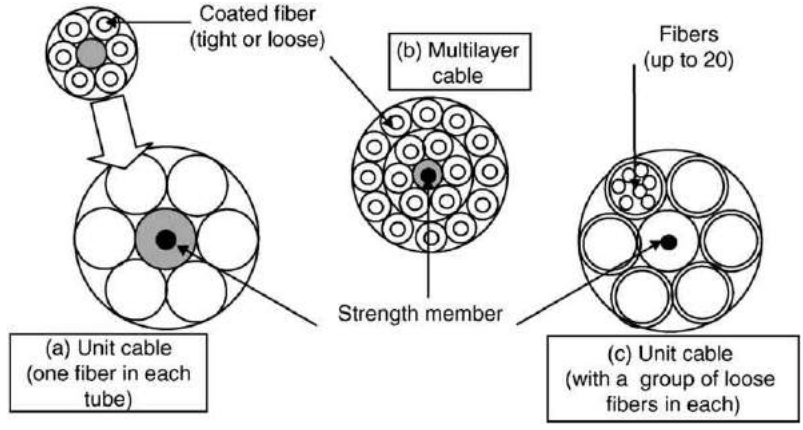


Fig. 5 Structures of stranded core cables.

material for the slotted core joined to a proper strength member, make it possible to obtain cables with fiber stability at a wide range of temperature. As in the loose stranded solutions, the slotted core solution is adopted to obtain cables with a high number of fibers in small dimensions.

In the ribbon solution, the fibers (from 4 to 16) are placed in a linear ribbon array. Many ribbons are stacked together to constitute the optical unit. The ribbon cable normally has a loose structure (loose tube or slotted core). Three structures are shown in Fig. 6. In the first example (classical AT & T cable), 12 ribbons of 12 fibers are directly stacked and stranded in a tube (Fig. 6(a)). In the second and third examples, ribbons are inserted into tubes or grooves (Fig. 6(b) and (c)). Several tubes or slotted cores can be assembled together to give a cable with very high fiber density. The ribbon approach may be efficient when a large number of fibers must be simultaneously handled, spliced, and terminated at connectors.

Types and Applications

The choice of an optical cable is heavily established by many factors, but mainly on the type of installation. On this basis, some different typology of cables may be identified: aerial cables, duct cables, direct buried cables, indoor cables, underwater cables, etc.

Aerial cables can be clasped onto a metal strand or can be self supporting. In the first case, the cable is not subjected to a particular tension but requires good thermal and mechanical performances. This solution is chosen when the cable must withstand strong ice and wind and when long distances between the poles are required. All the structures mentioned above may be adopted. In the second case, the cable is normally exposed to high mechanical stresses during its life and high tensile strength must be guaranted to maintain the fiber elongation at a safety level. Generally, loose structures are preferred for their greater fiber freedom. Often, in self-supporting cables, the strength member is external to the cable core. An example is shown in Fig. 7.

Cables pulled in ducts must be resistant to pulling and torsion forces, and light and flexible to permit the installation of long sections. The cable must also protect the fibers against water and moisture that may be present in ducts or manholes. Usually the cable is filled with a jelly compound, a metallic barrier is normally used and, often, an armoring is added when protection against rodents is necessary. All the structures described above may be employed in duct cables.

The characteristics of direct buried cables are very similar to those of duct cables, but additional protections, such as metallic armoring, are required to avoid the risk of damages from digging and other earthmoving work.

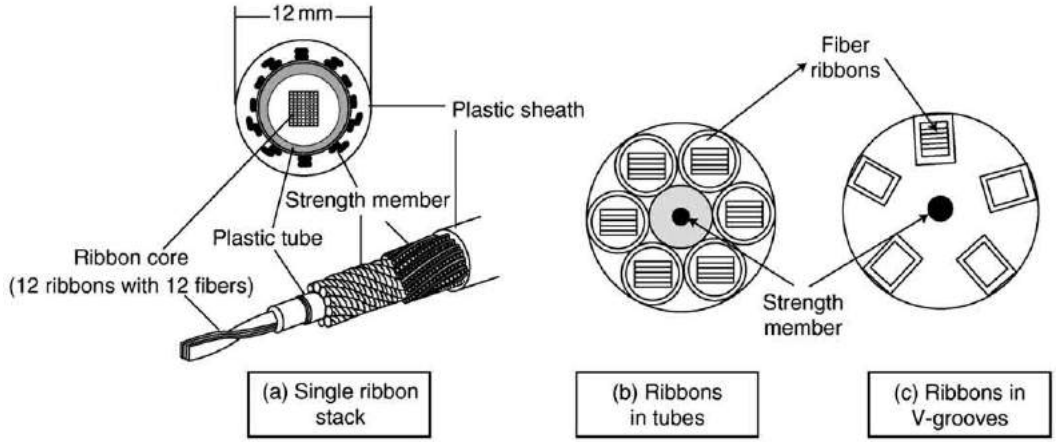


Fig. 6 Structures of ribbon cables.

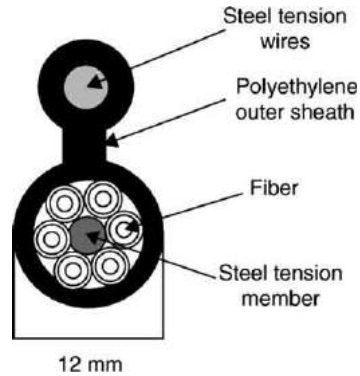


Fig. 7 Self-supporting aerial cable.

Cables for indoor applications normally require a smaller number of fibers. Their main characteristics are: small dimensions, flexibility, small curvature radius, and ease in handling and jointing. For indoor applications, the sheath must not permit flame propagation and must have a low emission of toxic gases and dark smoke. Tight cables are preferably used for indoor applications due to their compactness and small dimensions.

Underwater cables are employed in the crossing of rivers, in lakes and lagoons, and for shallow water applications in general. Such cables must have stringent requirements such as resistance to moisture and water penetration and propagation, pulling resistance (during the installation, and recovery in event of failure), and high resistance to static pressure and core structure compatible with land cables (when it is part of a ground link). In addition, due to the presence of metallic structures, attention should be paid to hydrogen effects.

Submarine Cables

Submarine cable is manufactured for laying in deep-sea conditions. In designing a submarine cable, it is necessary to provide high reliability that the mechanical and transmission characteristics of the optical fibers will be stable over a long period of time. Different cables are used for submarine applications, but generally follow the basic scheme of Fig. 8.

The central core (with a few fibers – up to 12 – in a slotted core or in a tight structure) is surrounded by a double layer of steel wires that act as strength members against tensile action and water pressure. Metal pipes at the inner or outer sides of the strength member, act as a water barrier and as power supply conductors (for the supply of the undersea regenerators).

The outer polyethylene insulating sheath is the ultimate protection for the ordinary deep-sea cable. For cables requiring special protection, steel wire armoring (anti-attack from fishes), and further polyethylene are added.

The cable must have stringent mechanical requirements, such as resistance to traction, torsion, crushing, impact, and to shark attack. The cable must be suitable for installation using standard cable-laying ships. Furthermore the cable materials must have low content of hydrogen and emissions. If necessary, a hydrogen absorber may be included in the cable core.

Special Cables

This category comprises cables not specifically used in ordinary telecommunication networks, but cables used for specific applications and according to specified requirements. Cables for military use are normally employed for temporary plant restoration. They must be light but crush resistant and with good mechanical characteristics. Cables for mobile applications (robots, elevators, aircrafts, submarines, etc.) generally require high flexibility and tensile strength. Special cables may be installed in particular environments, where they must be insensitive, for instance, to nuclear radiation, chemical agents, and high temperatures.

Optical cables can be installed in power lines, together with the power conductors, or into the ground wires of high-voltage overhead power lines. In this last type of cable, the central stranded steel wire, making up the ground wire, is replaced with a metal tube containing the optical unit (normally a small groups of optical fibers inserted in a tube or in a slotted core).

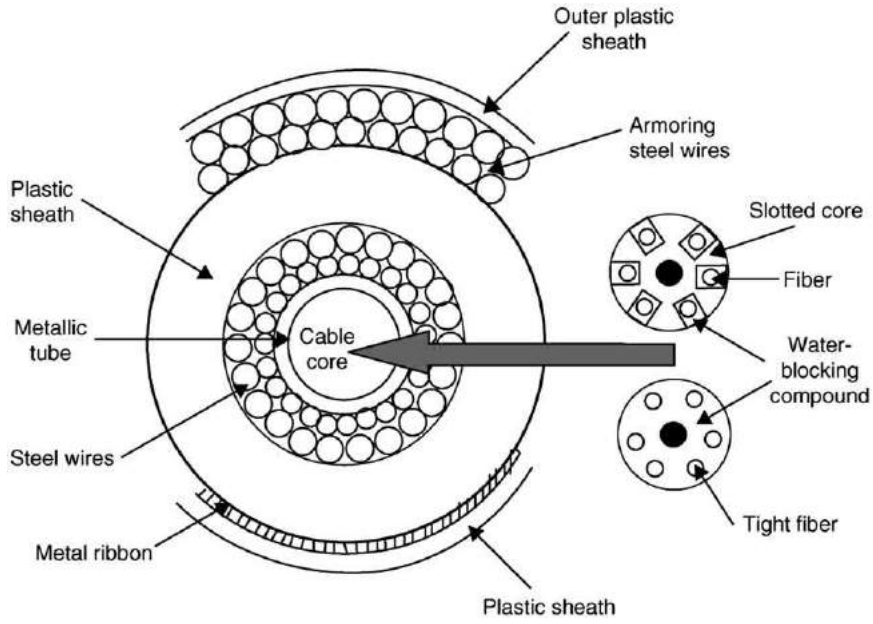


Fig. 8 Basic structure of a submarine cable.

Blown Fiber

This technology consists of the installation of a single fiber or of small fiber bundles (or cables) into pre-installed tubes or ducts using a flux of compressed air. Initially, cables with empty tubes are installed and successively, at the time of need, the tubes can be filled by blowing fibers (singly, or in bundles) or cables. The installation time is normally fast; some compressors can blow tens of meters in a minute. The typical maximum blowing distance is about 1 km for horizontal planes, but it is reduced along winding roads. The advantage of this technique is to produce variable and flexible structures, increasing the number of fibers to fit the network need. Besides, substitutions or change of fibers can be quickly carried out without substitution of the cable.

See also: Fabrication of Optical Fiber

Further Reading

Griffioen, W., 1989. The installation of conventional fiber-optic cables in conduits using the viscous flow of air. Journal of Lightwave Technology 7 (2), 297–302.
Hughes, H., 1997. Telecommunications Cables. Design, Manufacture and Installation. John Wiley and Sons.
Keiser, G., 1986. Optical Fiber Communications. McGraw-Hill.
Murata, H., 1988. Handbook of Optical Fibers and Cables. Marcel Dekker.
Technical Staff of CSELT, 1990. Fiber Optic Communications Handbook, second ed. TAB Books.

Passive Optical Components

D Suino, Telecom Italia Lab, Torino, Italy

Introduction

Modern telecommunications networks are based on fiber optics as signal transporters. In these networks the signals travel in the form of light confined within the fibers. As light power travels through the fibers from the source apparatus to the receiver, a number of components are employed to manage the optical signal. All other components that are not used to manage the signal are called 'passive optical components'. They serve several functions, such as to join two different fibers to distribute the signal onto more optical branches to limit the optical power to select particular wavelengths, and so on.

These components can be located anywhere in the network, depending on their function and on the network architecture. However, their main location is near the apparatus in which it is necessary to manage the signal before it is sent through the network or at the junction in the network where it is sent on to the receiver.

The main types of passive components that are found in a network can be grouped as follows:

- Joint devices:
  - Optical connectors;
  - Mechanical splices.
- Branching devices:
  - Nonwavelength-selective branching devices;
  - WDM.
- Attenuators;
- Filters;
- Isolators;
- Circulators.

Optical fibers are also divided into several groups (multimode 50/125, multimode 62.5/125, singlemode, dispersion shifted, NZD, and so on). There are also components of different typology that match up with each particular fiber. However, the general concept in terms of functionality, technology, and characteristic parameters, are the same among passive optical components of the same type (connectors, branching devices, or attenuators) but of different groups (single-mode, multimode, etc.).

Optical Joint Devices

The function of an optical joint is to perform a junction between two fibers, giving optical continuity to the line. Two different classes of joint can be considered: the connectors and the splices. The joints made by using connectors, are flexible points in the network, which means they can be opened and reconnected several times in order, for example, to reconfigure the network plan. However, a splice is a fixed joint and usually cannot be opened and re-mated. There are two type of splice: mechanical splices, holding joints to fibers by glue or mechanical crimp, and fusion splices fusing the two fiber heads with a suitable fusion splicer.

General Considerations

Considering that fibers can be described as pipes, in the core of which flows light, continuity between two fibers means that the two cores must be aligned in a stable and accurate way. Obviously this alignment must be carried out with minimum power loss in the junction. The main factors that affect the power loss in a joint point between two fibers are:

- x = Lateral offset;
- z = Longitudinal offset;
- θ = Angular misalignment; and
- w_T/w_R = Mode field diameter ratio (in the single mode fiber) or numerical aperture ratio (in the multimode fiber).

These geometrical mismatching between the two fibers is represented in **Fig. 1**.

The function that describes the joint attenuation is different for single-mode and multimode connectors, but the parameters that contribute to it are the same. The general formulation for these functions are complicated and, in particular, for the multimode connector it involves an integral on the shape distribution of the light in the fiber core. These functions become simpler in particular conditions: for example, if we consider a single-mode joint without a gap between the two fibers (that is the more usual condition in the modern connectors) the z parameter is nil, and in this case the function that describes the attenuation

 doi:10.1016/B978-0-12-809283-5.01256-8

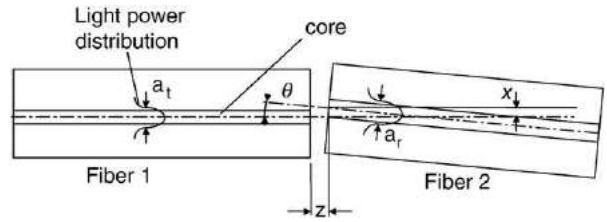


Fig. 1 Main parameter that affects the joint power loss.

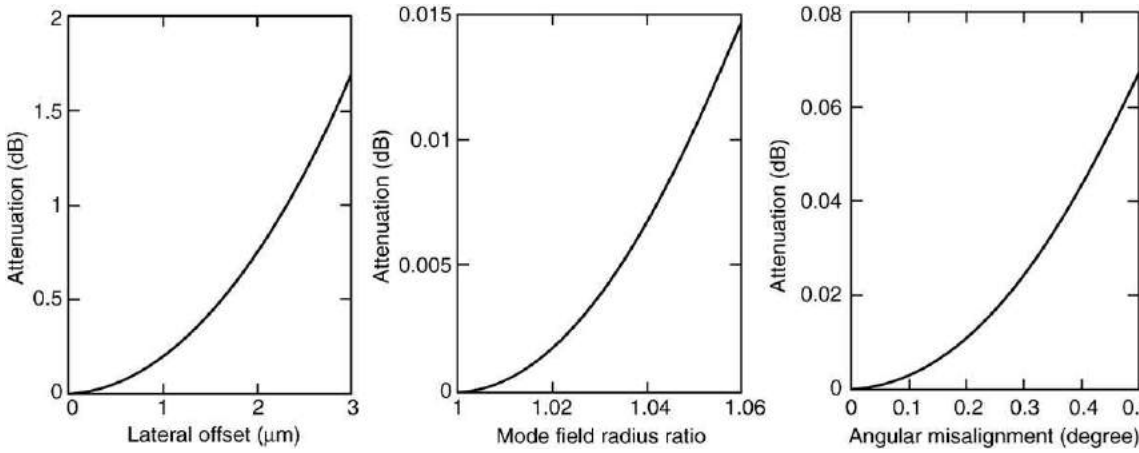


Fig. 2 Single mode connector attenuation as a function of the geometrical parameters. Each curve is plotted keeping the other parameter constant.

of the connector is

$$A(\mathrm{dB}) = -10\log\left\{\left(\frac{2w_T w_R}{w_T^2 + w_R^2}\right)^2 \times \exp\left[-2\frac{(x\lambda)^2 + (\pi \mathrm{n} w_T w_R \sin(\vartheta))^2}{\lambda^2(w_T^2 + w_R^2)}\right]\right\} \tag{1}$$

where n is the fiber core refractive index and λ is the wavelength of the light.

Fig. 2 shows the trend of Eq. (1), where the parameter ranges in the graphs are typical for the most common products on the market. It is evident that the main contribution to attenuation arises from the lateral offset.

For connectors of good quality, both for single-mode and multimode fibers, the attenuation values are in the range of a few tenths of dB (typically less than 0.3 dB). Another important factor for a joint, in particular for connectors used in networks for high speed or analog signals, is the return loss (RL). This is a parameter that gives a measure of the quantity of light back-reflected onto the optical discontinuity of the connector, due to the Fresnel effect. The RL is defined as the ratio in dB between the incident light power (P_{inc}) and the reflected light power (P_{ref}):

$$\mathrm{RL} = -10\log\left(\frac{P_{\mathrm{ref}}}{P_{\mathrm{inc}}}\right) \tag{2}$$

Fresnel reflection arises when the light passes between two media with different refractive indices. In the case of connectors in which the two fibers are not in contact, the light passes from the silica of the fiber core to the air in the gap. This causes the reflection of 4% of the incident light that, as RL, is about 14 dB.

To improve the RL performances of an optical joint, it is necessary to reduce the refractive index difference. This can be achieved either by index matching material between the two fibers that reduces the differences between the refractive index of the fiber and the gap, or by performing a physical contact between the fiber heads. The first technique is typically used in mechanical splices but, due to the problem related with the pollution contamination of the index matching, it is not a good solution for connectors that are to be opened and reconnected several times.

The physical contact solution is the most frequently adopted for these connectors. Physical contact is obtained by polishing the ferrule and the fiber end-faces in a convex shape, with the fiber core positioned in the apex (Fig. 3). With this technique, the typical values of return loss for a single mode connector are in the range from 35 dB to 55 dB, depending on the polishing grade. The residual back-reflected light is generated in the thin layer of the fiber end surface, because of slight changes in the refractive index due to the polishing process.

Higher return loss values (small quantity of reflected light) are achieved with the angled physical contact (APC) technique. This is obtained by polishing the convex ferrule end face angled with respect to the fiber axis. In this way, the light is not back reflected

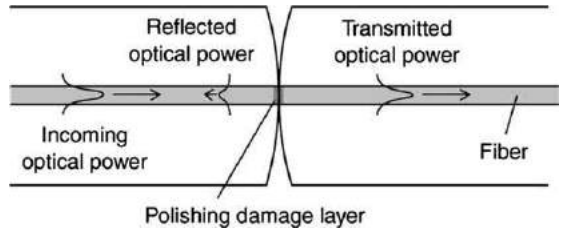


Fig. 3 Back reflection effects in PC connectors.

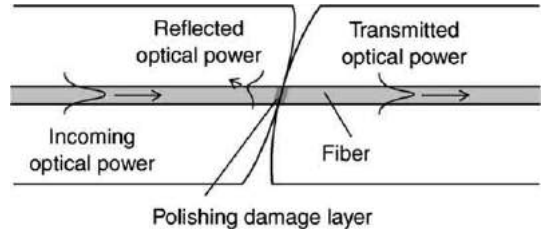


Fig. 4 Back reflection effects in APC connectors.

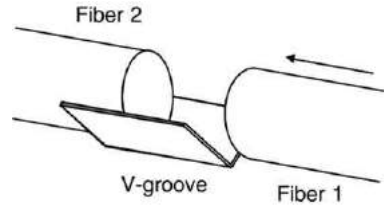


Fig. 5 Scheme of the bare fibers alignment on a V-groove.

into the fiber core, but into the cladding, and is thus eliminated (Fig. 4). In APC connectors, typical return loss values are greater than 60 dB.

Optical Connectors

On the market are several types of optical connectors. It is possible to divide them into three main groups:

- Connectors with direct alignment of bare fiber;
- Connectors with alignment by means of a ferrule; and
- Multifiber connectors.

Connectors with direct alignment of bare fiber

In this type of connector the bare fibers are aligned on an external structure that is usually a V-groove (Fig. 5). Two configurations can be performed: plug and socket or plug–adapter–plug. In the first case, the fiber in the socket is fixed and the one in the plug aligns to it. In the plug–adapter–plug configuration, two identical plugs containing the bare fibers are locked onto an adapter in which the fibers are placed on the aligned structure. These connectors are cheaper than the connectors with a ferrule, (described below), but this is unreliable. In fact, the bare fibers are delicate and brittle and they can be affected by the external mechanical or environmental stresses.

Connectors with alignment by means ferrule

In this type of connector, the fibers are fixed in a secondary alignment structure, usually of cylindrical shape, called a ferrule. The fiber alignment is obtained by inserting the two ferrules containing the fibers into a sleeve. Usually this type of connector is in the plug–adapter–plug configuration (Fig. 6). With this technique, the precision of the alignment is moved from the V-groove precision to the dimensional tolerance of the ferrule. Typical tolerance values on the parameters for a cylindrical ferrule with a diameter of 2.5 mm, are in the range of a tenth of a micrometer, and must be verified on the diameter, eccentricity between the hole (through which the fiber is fixed) and cylindricity of the external surface.

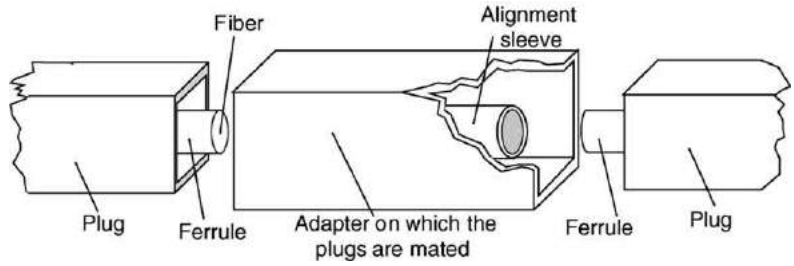


Fig. 6 Scheme of a plug–adapter–plug connector with alignment by ferrules and sleeve.

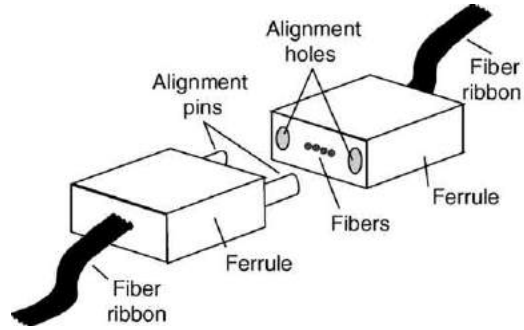


Fig. 7 Multifiber connector.

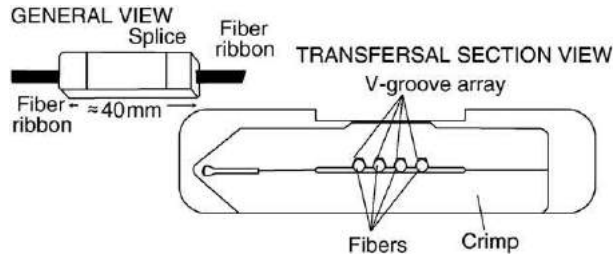


Fig. 8 Example of a multifiber mechanical splice.

Multifiber connectors

In some types of optical cable, the fibers are organized in multiple structures, in which many fibers are glued together to form a ribbon. The ribbon can have 2, 4, 8, 12, or 16 fibers, for example. When these cables are used, it is convenient to maintain the ribbon fiber structure in the connectors (until the point in which it is necessary to use the fibers singularly). Multifiber connectors are typically based on a secondary alignment on a plastic ferrule of rectangular shape in which the fibers maintain the same geometry that they have in the ribbon. The rectangular ferrules of the two plugs are normally aligned by means of two metallic pins placed in two hole in line with the fiber. In Fig. 7, a scheme of a multifiber connector is shown.

Mechanical Splices

Mechanical splices are yet another way to join together two fibers. All the main considerations for the connectors are true for the mechanical splices. The difference is that the mechanical splice is a fixed joint and, normally, it cannot be opened and re-mated. Mechanical splices, both for single fiber and for multifiber structures, are the most usually manufactured. In a mechanical splice, the fibers are aligned directly in a V-groove and fixed by glue or a mechanical holder.

These components are cheaper than fusion splices but their optical, mechanical, and environmental functions are lower than the fusion splices. Fig. 8, shows a mechanical splice for multifibers in which the fibers are crimped on a V-groove array. Usually, to avoid high optical reflection in the junction point, this type of joint is filled with index matching material.

Branching Devices

Branching devices are passive components devoted to distributing the optical power from an input fiber to two or more fibers. These components can be classified in two broad classes, on the basis of the wavelength dependence. The nonwavelength-selective

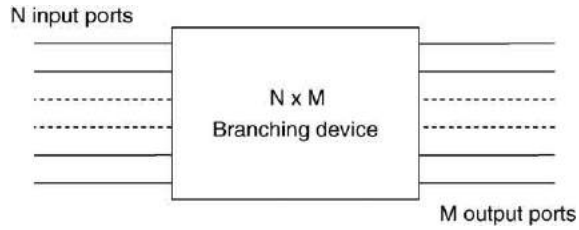


Fig. 9 Scheme of a generic N × M branching device.

branching devices, called also couplers or splitters, share the input power to two or more fibers independently from the light wavelength. The components that perform the light distribution on more fibers on the basis of the different wavelengths of the input light are called a WDM (wavelength division (de)multiplexer).

Branching devices have three or more ports, bare fiber or connectors, for the input and/or output of optical power, and share optical power among these ports. A generic branching device can be represented as a multiport device having N input ports and M output ports, as shown in the Fig. 9.

The optical characteristics of a branching device can be represented by a square transfer matrix of $n \times n$ coefficients, where n is the total number of the component ports. Each coefficient t_{ij} in the matrix is the fractional optical power transferred among designated ports, that is the ratio of the optical power P_{ij} transferred out of port j with respect to input power P_i into port n:

$$T = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & & & \\ \vdots & & t_{ij} & \vdots \\ & & & \\ t_{n1} & & & t_{nn} \end{bmatrix} \tag{3}$$

where:

$$t_{ij} = \frac{P_{ij}}{P_i} \tag{4}$$

So, according to the relationship between two ports, the coefficients represent several parameters. For example, if the port i is an input port and j is an output port, t_{ij} is the transferred power signal through the component, while if they exist as two isolated ports the coefficients represent the crosstalk between the two ports.The three main techniques to obtain this type of component are:

(i) All fiber components;
(ii) Micro-optics; and
(iii) Integrated optics.

Nonwavelength-Selective Branching Devices

A nonwavelength-selective branching device is a component having three or more ports, which shares an input signal among the output ports in a predetermined fashion. Usually these components are completely reversible and so they also work as a combiner of signals from more input ports onto a single output port.There are several types of couplers for different applications:

- Y-coupler with one input port and two output ports;
- X-coupler with two input ports and two output ports; and
- Star coupler with more than two ports in input and/or in output.

Moreover the couplers can be symmetrical, and in this case they divide the input power light onto n equal output flows, or asymmetrical devices in which the quantity of power light in the output ports is shared in a predefined nonuniform way.

The techniques used to carry out this type of device are mainly the fusion technique or planar technique. The first one is based on the phenomenon of evanescent wave coupling: when the core of two optical guides (as fibers) are located in close proximity, a power exchange from a guide to another is possible. This configuration can be obtained by stretching the fibers under controlled fusion (see Fig. 10).

The planar technique, that is the simpler integrated optic application, is based on the optical guides with appropriate shapes being placed directly onto a silicon wafer by means of lithographic processes (Fig. 11).

The fusion technique has the advantage of being 'all in fiber' so the component is ready to be inserted into the optical networks by means of usual joint techniques. But for a high number of ports, the coupling ratio among the branches cannot be controlled in

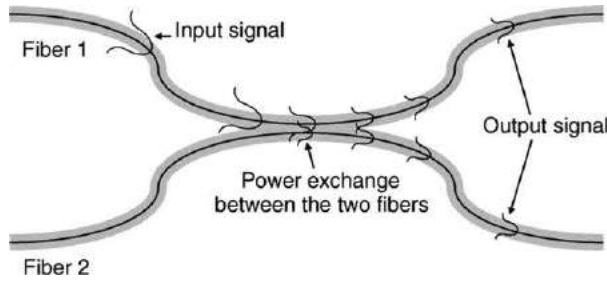


Fig. 10 Scheme of a fusion 2 × 2 coupler.

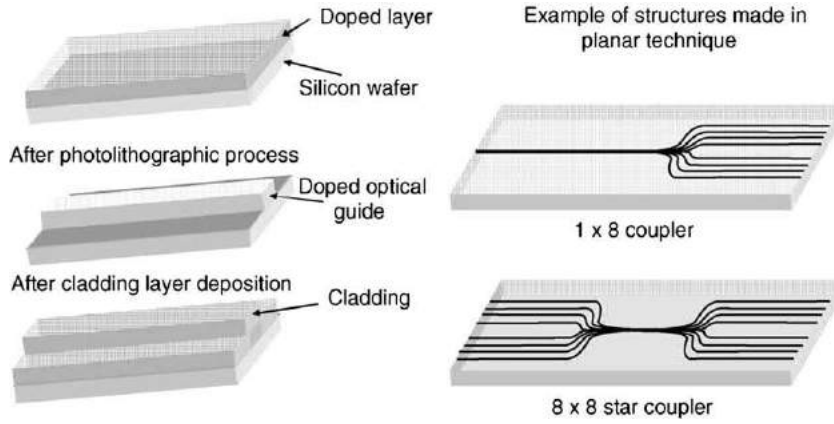


Fig. 11 Scheme of the process to obtain a planar coupler and an example of a device.

a precise way and the cost of the fusion device is proportional to the number of the branches. On the contrary, a device obtained with the planar technology must be coupled with fibers normally used and this operation may be difficult due to the different geometry of the planar guide (rectangular) and the fiber (circular). On the other hand, this technology allows a high precision in the optical characteristics of the component and the cost of the device is independent of the branch number.

WDM

A wavelength-selective branching device, usually called WDM (Wavelength Depending Multiplexer), is a component having three or more ports which shares an input signal among the output ports in a predetermined fashion, depending on the wavelength, so that at least two different wavelength ranges are nominally transferred between two different couples of ports.

WDM devices may be divided in three categories on the basis of the bandwidth of the channels (the spacing between two adjacent wavelength discriminated by the device):

(i) DWDM (Dense WDM) device operates for channel spacing equal or less than 1000 GHz (about 6–8 nm in the range of typical optical wavelength used in telecommunications systems, 1310 or 1550 nm);
(ii) CWDM (Coarse WDM) device operates for channel spacing less than 50 nm and greater than 1000 GHz; and
(iii) WWDM (Wide WDM) device operates for channel spacing equal or greater than 50 nm.

These components are used to combine, at the input side of an optical link in only one fiber, more optical signals with different wavelengths and separate them at the receiving end. This technology allows to send along a fiber, more communication channels, thus increasing the total bit rate transmitted. These types of components can be obtained using filters to select single a wavelength (as a diffractive grating that resolves the light into its monochromatic components) or using the phenomena of evanescent field coupling occurring between two adjacent fiber cores (as described for coupler devices), controlling in a precise way the coupling zone; this is, in fact, dependent on the wavelength (Fig. 12).

Using wavelength filters, as they select a singular wavelength, it is necessary to perform a cascade filter structure for separating more than two wavelengths.

Micro-optic technologies are generally used for multimode fiber due to the critical collimation problem. For the single-mode fiber the evanescent field coupling technique is applied directly to fiber or to planar optical guides.

Another efficient technique to create WDM devices is with a fiber Bragg grating (FBG) wavelength filter directly applied to the fiber. FBGs consist in periodic modulation (see Fig. 13) of the refractive index along the core of the fiber, and they act as reflection filters.

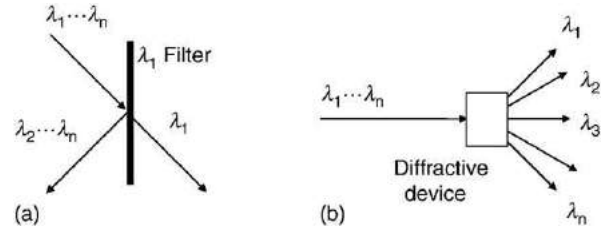


Fig. 12 (a) Scheme of the filter technique and (b) diffractive device technique.

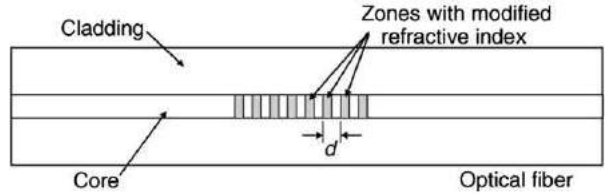


Fig. 13 Scheme of FBG.

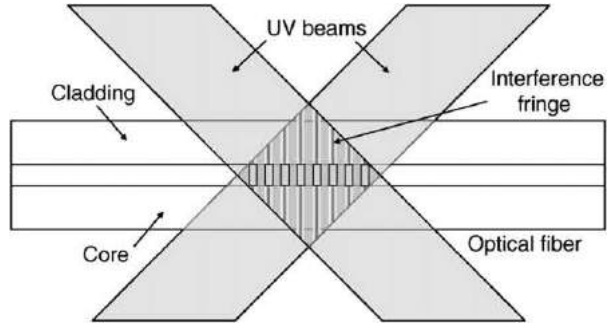


Fig. 14 Scheme of FBG production.

During the fabrication process (see Fig. 14) the core of the optical fiber is exposed to UV radiation, and this exposure increases the refractive index of the core in defined zones, with a periodic shape obtained by the interference fringe created by crossing two coherent beams.

Fig. 15 shows a representative spectral curve of a FBG filter centered at 1625 nm. The central wavelength, the width, and the value of reflection depend on the grating period (d) and on the deposited UV energy in the fiber core.

Other Passive Optical Components

In an optical network, besides the devices described above, there are a number of other passive optical components that perform particular functions on the optical signal; the main ones are:

Attenuators These are devices used to attenuate the optical power, for example, in front of a receiver when the optical line is short and the power of the received signal is still too high for the detector, or when it is necessary to equalize the signals arriving from different optical lines. It exists both in tunable and fixed versions. The first ones allow adjustment of the attenuation value, however, the second ones have a fixed and predefined value. The fixed optical attenuators can be obtained by inserting a filter or stressing a fiber in a controlled and permanent way. The tunable attenuators are based on variable optical parameters as, for example, the distance or the lateral offset in an optical connection. They exist both as connect lengths of fiber (attenuated optical jumpers) or as compact components similar to a double connector (plug stile attenuator).

Filters These are components with two ports employed to select or to filter some fixed wavelengths. They can be divided into the following categories:

- short-wave pass (only wavelengths lower than or equal to a specified value are passed);
- long-wave pass (only wavelengths greater than or equal to a specified value are passed);
- band-pass (only an optical window is allowed); and
- notch (only an optical window is inhibited).

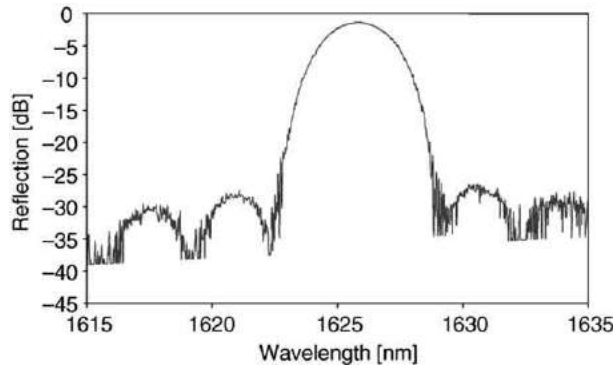


Fig. 15 Transmission characteristic of FBG.

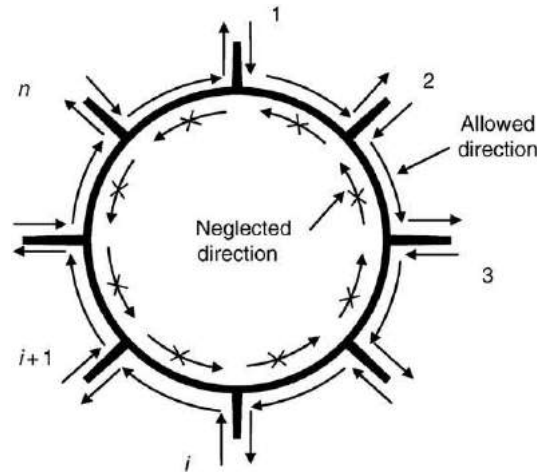


Fig. 16 Principle scheme of an optical circulator.

It is also possible to have a combination of the above categories. The optical filtering can be obtained by inserting one or more filtering devices (as interferential films) into a fiber or creating a filter directly on the fiber by the FBG described above.

Isolators These are nonreciprocal two-port optical device in which the light flows in the forward direction, and not in the reverse one. The isolators are used to suppress backward reflections along an optical fiber transmission line, while having minimum insertion loss in the forward direction. Fiber optic isolators are commonly used to avoid back reflections into laser diodes and optical amplifiers, which can make the laser and amplifiers oscillate unstably, and cause noise in the fiber optic transmission system.

Circulators These are passive components having three or more ports numbered sequentially $(1, 2, \ldots, n)$ through which optical power is transmitted nonreciprocally from port 1 to port 2,..., from port (i) to port $(i + 1)$, ... and from port n to port 1 (**Fig. 16**).

See also: Broadband Passive Optical Access Networks

Further Reading

Adam, M.J., 1981. An Introduction to Optical Waveguides. New York: John Wiley & Sons.
Blonder, G.E., *et al.*, 1989. Glass waveguides on silicon for hybrid optical packaging. IEEE Journal of Lightwave Technology 7, 1530.
Digonnet, M., Shaw, H.J., 1983. Wavelength multiplexing in single-mode fiber couplers. Applied Optics 22, 484.
Jeunhomme, L.B., 1983. Single-Mode Fiber Optics – Principles and Applications. New York: Marcel Dekker.
Kashima N (1995) *Passive Optical Components for Optical Fiber Transmission*:.
Kashima, N., 1993. Optical Transmission for the Subscriber Loop. Norwood, MA: Artech House.
Miller, C.M., 1986. Optical Fiber Splices and Connectors – Theory and Methods. New York: Marcel Dekker.
SC86B IEC, 1999. IEC 60874-1: Connectors for Optical Fibres and Cables – Part 1: Generic Specification, 4th edn. Geneva: International Electrotechnical Commission.
SC86B IEC, 1999. IEC 61073-1: Optical Fibres and Cables – Mechanical Splices and Fusion Splice Protectors – Part 1: Generic Specification, 3rd edn. Geneva: International Electrotechnical Commission.

SC86B IEC, 1999. IEC 60869-1: Fibre Optic Attenuators – Part 1: Generic Specification, 3rd edn. Geneva: International Electrotechnical Commission.
SC86B IEC, 2000. IEC 60875-1: Non-wavelength Selective Fibre Optic Branching Devices – Part 1:Generic Specification, 4th edn. Geneva: International Electrotechnical Commission.
SC86B IEC, 2000. IEC 62077-1: Fibre Optic Circulators – Part 1: Generic Specification, 1st edn. Geneva: International Electrotechnical Commission.
SC86B IEC, 2000. IEC 61977-1 Ed. 1.0: Fibre Optic Filters – Generic Specification, 1st edn. Geneva: International Electrotechnical Commission.
SC86B IEC, 2000. IEC 61202-1: Fibre Optic Isolators – Part 1: Generic Specification, 1st edn. Geneva: International Electrotechnical Commission.
Technical Staff of CSELT, 1990. Fiber Optic Communications Handbook, 2nd edn. New York: McGraw-Hill.
Yokohama, I., *et al*., 1988. Novel mass-fabrication of fiber couplers using arrayed fiber ribbons. Electronic Letters 24, 1147.

Nonlinear Effects (Basics)

G Millot and P Tchofo-Dinda, Université de Bourgogne, Dijon, France

Nomenclature

n_2 Nonlinear refractive index [$m^2\ W^{-1}$]
I Optical intensity [GW $cm^{-2} = 10^9$ W cm^{-2}]
CPM cross phase modulation
FWM four wave mixing
MI modulational instability
SPM self-phase modulation

Introduction

Many physical systems in various areas such as condensed matter or plasma physics, biological sciences, or optics, give rise to localized large-amplitude excitations having a relatively long lifetime. Such excitations lead to a host of phenomena referred to as nonlinear phenomena. Of the many disciplines of physics, the optics field is probably the one in which practical applications of nonlinear phenomena have been the most fruitful, in particular, since the discovery of the laser in 1960. This discovery has thus led to the advent of a new branch in optics, referred to as *nonlinear optics*. The applications of nonlinear phenomena in optics include the design of various kinds of laser sources, optical amplifiers, light converters, light-wave communication systems for data transmission purposes, to name a few.

In this article, we present an overview of some basic principles of nonlinear phenomena that result from the interaction of light waves with dielectric waveguides such as optical fibers. These nonlinear phenomena can be broadly divided into two main categories, namely, parametric effects and scattering phenomena. Parametric interactions arise whenever the state of the dielectric matter is left unchanged by the interaction, whereas scattering phenomena imply transitions between energy levels in the medium. More fundamentally, parametric interactions originate from the electron motion under the electric field of a light wave, whereas scattering phenomena originate from the motion of heavy ions (or molecules).

Linear and Nonlinear Signatures

The macroscopic properties of a physical system can be obtained by analyzing the response of the system under an external excitation. For example, consider at time t the response of a system, such as an amplifier, to an input signal $E_1 = A\sin(\omega t)$. In the low-amplitude limit of the output signal, the response R_1 of the system is proportional to the excitation:

$$R_1 = \alpha_1 E_1 \tag{1}$$

where α_1 is a constant. This type of behavior corresponds to the so-called linear response. In general, a physical system executes a linear response when the superposition of two (or more) input signals E_1 and E_2 yields a response which is a superposition of the output signals, as schematically represented in **Fig. 1**:

$$R = \alpha_1 R_1 + \alpha_2 R_2 \tag{2}$$

Now, in almost all real physical systems, if the amplitude of an excitation E_1 becomes sufficiently large, distortions will occur in the output signals. In other words, the response of the system will no longer be proportional to the excitation, and consequently,

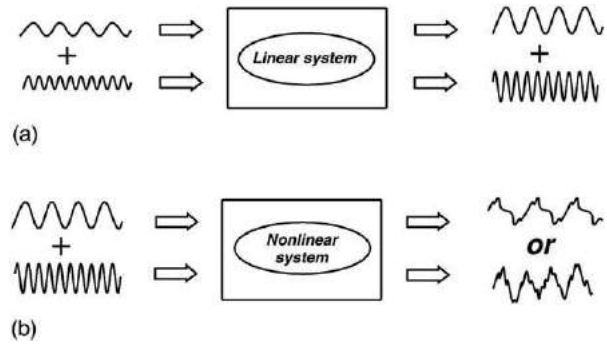


Fig. 1 Schematic representation of linear and nonlinear responses of a system, to two input signals.

the law of superposition of states will no longer be observed. In this case, the response of the system may take the following form:

$$R = \alpha_1 E_1 + \alpha_2 E_1^2 + \alpha_3 E_1^3 + \ldots \tag{3}$$

which involves not only a signal at the input frequency ω, but also signals at frequencies 2ω, 3ω, and so on. Thus, harmonics of the input signal are generated. This behavior, called *nonlinear response*, is at the origin of a host of important phenomena in many branches of sciences, such as in condensed matter physics, biological sciences, or in optics.

Physical Origin of Optical Nonlinearity

Optical nonlinearity originates fundamentally from the action of the electric field of a light wave on the charged particles of a dielectric waveguide. In contrast to conductors where charges can move throughout the material, dielectric media consist of *bound* charges (ions, electrons) that can execute only relatively limited displacements around their equilibrium positions. The electric field of an incoming light wave will cause the positive charges to move in the polarization direction of the electric field whereas negative charges will move in the opposite direction. In other words, the electric field will cause the charged particles to become dipoles, as schematically represented in **Fig. 2**.

Then each dipole will vibrate under the influence of the incoming light, thus becoming a source of radiation. The global light radiated by all the dipoles represents the scattered light. When the charge displacements are proportional to the excitation (incoming light), i.e., in the low-amplitude limit of scattered radiation, the output light vibrates at the same frequency as that of the excitation. This process corresponds to Rayleigh scattering. On the other hand, if the intensity of the excitation is sufficiently large to induce displacements that are not negligible with respect to atomic distances, then the charge displacements will no longer be proportional to the excitation. In other words, the response of the medium, which is no longer proportional to the excitation, becomes nonlinear. In this case, the scattered waves will be generated not only at the excitation frequency (the Kerr effect), say ω, but also at frequencies that differ from ω (e.g., 2ω, 3ω). On the other hand, it is worth noting that when all induced dipoles vibrate coherently (that is, their relative phase does not vary randomly), their individual radiation may, under certain conditions, interfere constructively and lead to a global field of high intensity. The condition of constructive interference is the phase-matching condition.

In practice, the macroscopic response of a dielectric is given by the polarization, which corresponds to the total amount of dipole moment per unit volume of the dielectric. As the mass of an ion is much larger than that of an electron, the amplitude of the ion motion is generally negligible with respect to that of the electrons. As a consequence, the electron motion generally leads to the dominant contribution in the macroscopic properties of the medium. The behavior of an electron under an optical electric field is similar to that of a particle embedded in an anharmonic potential. A very simple model (called the Lorentz model) that provides a deep insight into the dielectric response consists of an electron of mass m and charge $-e$ connected to an ion by an elastic spring (see **Fig. 2**). Under the electric field $E(t)$, the electron executes a displacement $x(t)$ with respect to its equilibrium position, which is governed by the following equation:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + 2\lambda\frac{\mathrm{d}x}{\mathrm{d}t} + \omega_0^2 x + (a^{(2)}x^2 + a^{(3)}x^3 + \ldots) = -\frac{e}{m}E(t) \tag{4}$$

where $a^{(2)}$, $a^{(3)}$, and so on, are constant parameters, ω_0 is the resonance angular frequency of the electron, and λ is the damping coefficient resulting from the dipolar radiation. When the amplitude of the electric field is sufficiently large, then the restoring force on the electrons becomes a nonlinear function of x; hence the presence of terms such as $a^{(2)}x^2$, $a^{(3)}x^2$, and so on, in **Eq. (4)**. In this situation, the macroscopic response of the dielectric is the polarization

$$P = -\sum ex(\omega, 2\omega, 3\omega, \ldots) \tag{5}$$

where $x(\omega, 2\omega, 3\omega,\ldots)$ is the solution of **Eq. (4)** in the frequency domain, and the summation extends over all the dipole moments per unit volume. In terms of the electric field E the polarization may be written as

$$P = \varepsilon_0\left(\chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \ldots\right) \tag{6}$$

where $\chi^{(1)}$, $\chi^{(2)}$, $\chi^{(3)}$, and so on, represent the susceptibility coefficients. **Fig. 3** (top left) illustrates schematically the polarization as a function of the electric field.

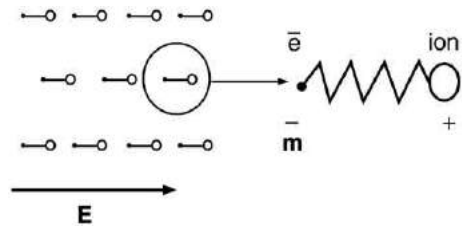


Fig. 2 Electric dipoles in a dielectric medium under an external electric field.

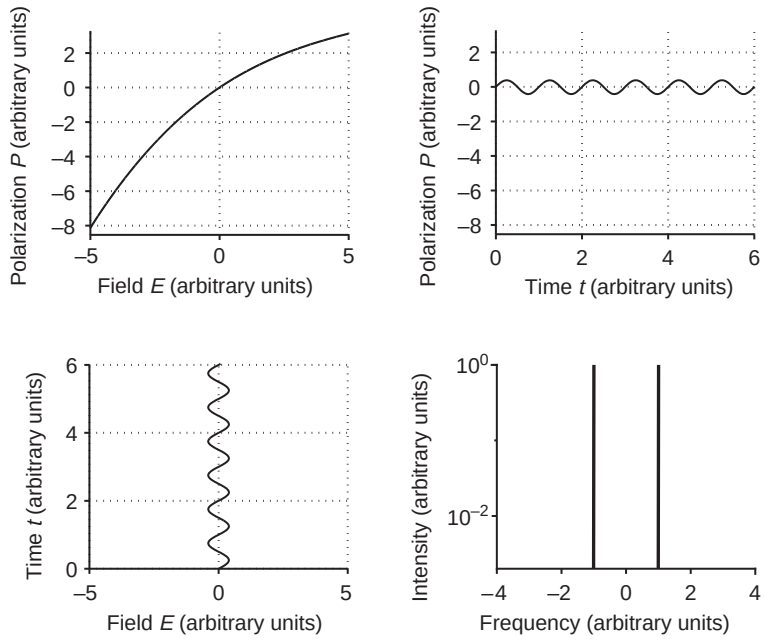


Fig. 3 Polarization induced by an electric field of small amplitude. Nonlinear dependence of the polarization as a function of field amplitude (top left) and time dependence of the input electric field (bottom left). Time dependence of the induced polarization (top right) and corresponding intensity spectrum (bottom right).

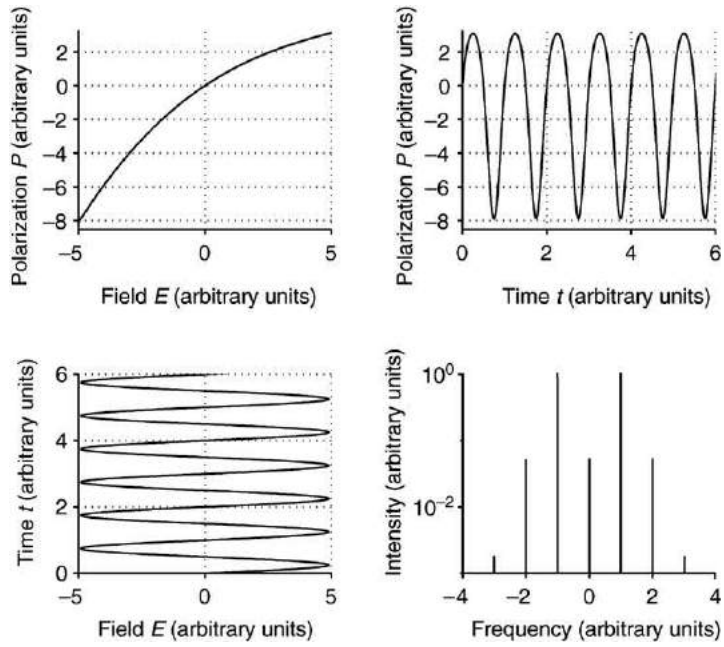


Fig. 4 Polarization induced by an incoming electric field of large amplitude.

In particular, one can clearly observe that when the amplitude of the incoming electric field is sufficiently small (bottom left in Fig. 3), the polarization (top right) is proportional to the electric field (bottom left), thus implying that the electric dipole radiates a wave having the same frequency as that of the incoming light (bottom right). On the other hand, Fig. 4 shows that for an electric field of large amplitude, the polarization is no longer proportional to the electric field, leading to radiation at harmonic frequencies (see Fig. 4, bottom right).

This nonlinear behavior leads to a host of important phenomena in optical fibers, which are useful for many optical systems but detrimental for other systems.

Parametric Phenomena in Optical Fibers

In anisotropic materials, the leading nonlinear term in the polarization, i.e., the $\chi^{(2)}$ term, leads to phenomena such as the harmonic generation or optical rectification. This $\chi^{(2)}$ term vanishes in homogeneous isotropic materials such as cylindrical optical fibers, and there the leading nonlinear term becomes the $\chi^{(3)}$ term. Thus, most of outstanding nonlinear phenomena in optical fibers originate from the third-order nonlinear susceptibility $\chi^{(3)}$. Some of those phenomena are described below.

Optical Kerr Effect

The optical Kerr effect is probably the most important nonlinear effect in optical fibers. This effect induces an intensity dependence of the refractive index, which leads to a vast wealth of fascinating phenomena such as self-phase modulation (SPM), cross-phase modulation (CPM), four-wave mixing (FWM), modulational instability (MI) or optical solitons. The Kerr effect can be conveniently described in the frequency domain through a direct analysis of the polarization, which takes the following form:

$$P_{\mathrm{NL}}(\omega) = \frac{3}{4}\varepsilon_0\chi^{(3)}(\omega)|E(\omega)|^2E(\omega) \tag{7}$$

The constant 3/4 comes from the symmetry properties of the tensor $\chi^{(3)}$. Setting $P_{\mathrm{NL}}(\omega)=\varepsilon_0\varepsilon_{\mathrm{NL}}E(\omega)$, where $\varepsilon_{\mathrm{NL}} = \frac{3}{4}\chi^{(3)}|E|^2$ is the nonlinear contribution to the dielectric constant, the total polarization takes the form

$$P(\omega) = P_{\mathrm{L}} + P_{\mathrm{NL}} = \varepsilon_0\left[\chi^{(1)}(\omega) + \varepsilon_{\mathrm{NL}}\right]E(\omega) \tag{8}$$

As Eq. (8) shows, the refractive index n, at a given frequency ω, is given by

$$n^2 = 1 + \chi^{(1)} + \varepsilon_{\mathrm{NL}} = (n_0 + \Delta n_{\mathrm{NL}})^2 \tag{9}$$

with $n_0^2 = 1 + \chi^{(1)}$. In practice $\Delta n_{\mathrm{NL}} \ll n_0$, and then, the refractive index is given by

$$n(\omega,\ |E|^2) = n_0(\omega) + n_2^e|E|^2 \tag{10}$$

where n_2^e is the nonlinear refractive index defined by $n_2^e = 3\chi^{(3)}/(8n_0)$. The linear polarization P_{L} is responsible for the frequency dependence of the refractive index, whereas the nonlinear polarization P_{NL} causes an intensity dependence of the refractive index, which is referred to as the optical Kerr effect. Knowing that the wave intensity I is given by $I=\alpha|E|^2$, with $\alpha = \frac{1}{2}\varepsilon_0 c n_0$, the refractive index can be then rewritten as

$$n(\omega, I) = n_0(\omega) + n_2 I \tag{11}$$

with $n_2 = n_2^e/\alpha = 2n_2^e/(\varepsilon_0 c n_0)$. For fused silica fibers one has typically: $n_2 = 2.66\times10^{-20}$ m^2 W^{-1}. For example, an intensity of $I=1$ GW cm^{-2} leads to $\Delta n_{\mathrm{NL}} = 2.66\times10^{-7}$, which is much smaller than $n_0 \approx 1.45$.

Four-Wave Mixing

The four-wave mixing (FWM) process is a third-order nonlinear effect in which four waves interact through an energy exchange process. Let us consider two intense waves, $E_1(\omega_1)$ and $E_2(\omega_2)$, with $\omega_2 > \omega_1$, called pump waves, propagating in an optical fiber. Hereafter we consider the simplest case when waves propagate with the same polarization. In this situation the total electric field is given by

$$E_{\mathrm{tot}}(\mathbf{r},t) = E_1 + E_2 = A_1(\omega_1)\exp[\mathrm{i}(\mathbf{k}_1\cdot\mathbf{r} - \omega_1 t)] + A_2(\omega_2)\exp[\mathrm{i}\mathbf{k}_2\cdot\mathbf{r} - \omega_2 t)] \tag{12}$$

where $\mathbf{k}_1$ and $\mathbf{k}_2$ are the wavevectors of the fields E_1 and E_2, respectively. Eq. (7), which gives the nonlinear polarization induced by a single monochromatic wave, remains valid provided that the frequency spacing between the two waves is relatively small, i.e., $|\Delta\omega| = |\omega_2 - \omega_1| \ll \omega_0 = (\omega_1+\omega_2)/2$. In this context, Eq. (7) leads to

$$P_{\mathrm{NL}} \approx \frac{3}{4}\varepsilon_0\chi^{(3)}(\omega_0)|E_{\mathrm{tot}}|^2E_{\mathrm{tot}} \tag{13}$$

Substituting Eq. (12) in Eq. (13) yields

$$P_{\mathrm{NL}} = 2n_0n_2\varepsilon_0\left[(|E_1|^2 + 2|E_2|^2)E_1 + (|E_2|^2 + 2|E_1|^2)E_2 + E_1^2E_2^* + E_1^*E_2^2\right] \tag{14}$$

The term $|E_1|^2E_1$ in the right-hand side of Eq. (14) represents the self-induced Kerr effect on the wave ω_1. The term $|E_2|^2E_1$ which corresponds to a modification of the refractive index seen by the wave ω_1, due to the presence of the wave ω_2, represents the cross-Kerr effect. Thus, the refractive index at frequency ω_1 depends simultaneously on the intensities of the two pumps, i.e., $n = n(\omega_1, |E_1|^2, |E_2|^2)$ Similarly the third and fourth terms in the right-hand side of Eq. (14) illustrate the self-induced and cross-Kerr effects on the wave ω_2, respectively. Note that the self-induced Kerr effect is responsible for a self-phase modulation, which induces a spectral broadening of a pulse propagating through the optical fiber, whereas the cross-Kerr effect leads to a cross-phase modulation that induces a spectral broadening of one wave in the presence of a second wave. The two last terms in the right-hand side of Eq. (14) correspond to the generation of new waves at frequencies $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$, respectively. The wave with the smallest frequency $2\omega_1 - \omega_2 = \omega_{\mathrm{s}}$ is called the Stokes wave, whereas the wave with the highest frequency $2\omega_2 - \omega_1 = \omega_{\mathrm{as}}$ is called the anti-Stokes wave. These two newly generated waves interact with the two pumps through an energy exchange process. This interaction is commonly referred to as a four-wave mixing process.

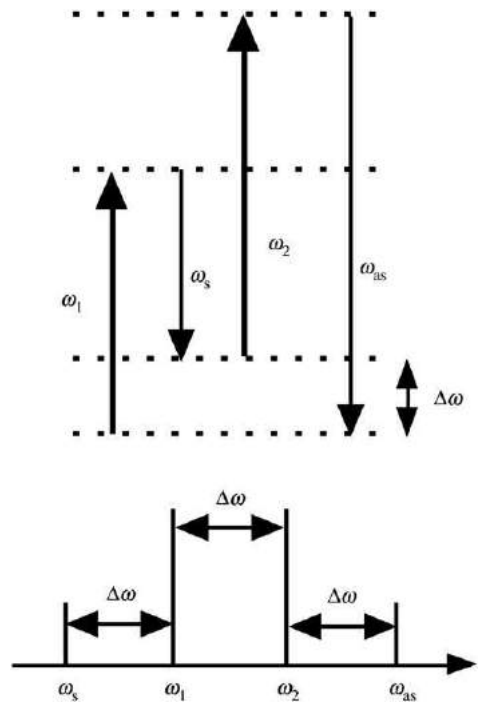


Fig. 5 Schematic diagram representing a non-degenerate four-wave mixing process (top) and the corresponding frequency spectrum (bottom).

From a quantum mechanical point of view, four-wave mixing corresponds to a process in which two photons with frequencies ω_1 and ω_2 are annihilated with simultaneous creation of two photons at frequencies ω_s and ω_{as} respectively. The new waves are generated at frequencies ω_s and ω_{as} such that

$$\omega_1 + \omega_2 = \omega_s + \omega_{as} \tag{15}$$

which states that the total energy is conserved during the interaction. But the condition for this FWM process to occur is that the total momentum is conserved, that is,

$$\Delta\mathbf{k} = \mathbf{k}_s + \mathbf{k}_{as} - \mathbf{k}_1 - \mathbf{k}_2 = 0 \tag{16}$$

or equivalently,

$$n(\omega_s)\omega_s + n(\omega_{as})\omega_{as} - n(\omega_1)\omega_1 - n(\omega_2)\omega_2 = 0 \tag{17}$$

Eq. (16) is known as the phase-matching condition. **Fig. 5** displays a schematic diagram of a four-wave mixing process with the corresponding frequency spectrum.

FWM with different pump frequencies $\omega_1 \neq \omega_2$ is called 'nondegenerate FWM', whereas the case $\omega_1 = \omega_2$ is referred to as 'degenerate FWM', or more simply, as three-wave mixing.

Conclusion

Optical fibers constitute a key device for many areas of optical sciences, and in particular for ultrafast optical communications. The nonlinear phenomena that arise through the nonlinear refractive index change (induced mainly by the Kerr effect) have been largely investigated these last two decades for applications such as all-optical wavelength conversion, parametric amplification, generation of new optical frequencies or ultrahigh repetition rate pulse trains. However, in many other optical systems such as optical communication systems, these nonlinear phenomena become harmful effects, and in this case control processes are being developed in order to suppress or at least reduce their impact in the system.

See also: Nonlinear Optics. Nonlinear Optics in Photonic Crystal Fibers

Further Reading

Agrawal, G.P., 2001. Nonlinear Fiber Optics. San Diego, CA: Academic Press.
Bloembergen, N., 1977. Nonlinear Optics. Reading, MA: Benjamin.

Boyd, R.W., 1992. Nonlinear Optics. San Diego, CA: Academic Press.
Butcher, P.N., Cotter, D.N., 1990. The Elements of Nonlinear Optics. Cambridge, UK: Cambridge University Press.
Dianov, E.M., Mamyshev, P.V., Prokhorov, A.M., Serkin, V.N., 1989. Nonlinear Effects in Optical Fibres. Switzerland: Harwood Academic Publishers.
Hellwarth, R., 1977. Third-order optical susceptibilities of liquids and solids. Progress in Quantum Electronics 5, 1–68.
Pocholle, J.P., Papuchon, M., Raffy, J., Desurvire, E., 1990. Non linearities and optical amplification in single mode fibres. Revue Technique Thomson-CSF 22, 187–268.
Shen, Y.R., 1984. Principles of Nonlinear Optics. New York: Wiley.

Basic Concepts of Optical Amplifiers

MFS Ferreira, University of Aveiro, Aveiro, Portugal

Nomenclature

g_0	Peak value of the gain coefficient at peak	m^{-1}
$g(\nu)$	Gain coefficient	m^{-1}
G_0	Unsaturated amplifier gain	
$G(\nu)$	Amplifier gain	
L	Amplifier length	m
n_{sp}	Spontaneous emission factor	
P	Signal power	W
P_{in}	Input signal power	W
P_{in}^{s}	Input saturation power	W
P_{out}	Output signal power	W
P_{out}^{s}	Output saturation power	W
P_{sat}	Saturation power	W
$S_{sp}(\nu)$	ASE noise spectral density	J
Δf	Detector bandwidth	Hz
$\Delta\nu_0$	Bandwidth of the gain coefficient	Hz
ν	Optical frequency	Hz
ν_0	Atomic transition frequency	Hz

Introduction

The transmission distance of a fiber-optic communication system is limited by fiber loss and dispersion. For long-haul lightwave systems, the loss limitations have traditionally been overcome by periodic regeneration of the optical signals at repeaters applying conversion to an intermediate electric signal. Because of the complexity and high cost of such regenerators, the need for optical amplifiers became obvious in the mid-1980s. The optical amplifier is ideally a transparent box that provides gain and is also insensitive to the bit rate, modulation format, power and wavelength of the signal passing through it.

Several means of obtaining optical amplification has been suggested since the 1970s, including direct use of the transmission fiber as gain medium through nonlinear effects, semiconductor amplifiers, or doping optical waveguides with an active material (rare-earth ions), that could provide gain. Due to the spectacular results on erbium-doped fiber amplifiers, which are particularly suitable in the third transmission window (around 1.5 μm), an intense worldwide research activity on optical amplifiers has developed. As a consequence, the development of erbium-doped fiber amplifiers has reached an industrial level, and commercial devices are now available.

Semiconductor amplifiers, on the other hand, have the same technical basis as semiconductor lasers. Although the strong nonlinearity of semiconductor amplifiers degrades the performances of transmission systems, the state-of-art semiconductor devices seem to be the most interesting amplifiers for transmission in the second transmission window (around 1.3 μm).

Amplifier Gain and Bandwidth

In a perfect amplifier, the amplification process would be insensitive to the bit rate, modulation format, power, state of polarization, wavelength, and optical bandwidth of the signal passing through it. On the other hand, no interaction would take place if more than one signal were amplified simultaneously. In practice, however, the optical gain depends not only on the wavelength (or frequency) of the incident signal, but also on the electromagnetic field intensity at any point inside the amplifier. Details of wavelength and intensity dependence of the optical signal depend on the amplifying medium.

We consider a case in which the gain medium is modeled as a homogeneously broadened two-level system. In such medium, the gain coefficient (i.e., the gain per unit length) can be written as:

$$g(\nu, P) = \frac{g_0}{1 + \frac{(\nu - \nu_0)^2}{\Delta\nu_0^2} + \frac{P}{P_{sat}}} \tag{1}$$

where g_0 is the peak value of the gain coefficient determined by the pumping level of the amplifier, ν the optical frequency, ν_0 the atomic transition frequency, $\Delta\nu_0$ the 3 dB local gain bandwidth, P the optical power of the signal, and P_{sat} the saturation power,

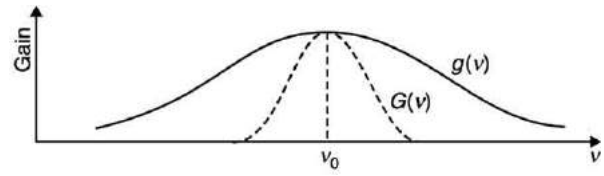


Fig. 1 Gain coefficient profile $g(\nu)$ and the corresponding amplifier gain spectrum $G(\nu)$.

which depends on the gain medium parameters. It must be emphasized that $\Delta\nu_0$ and P_{sat} refer to the local gain. However, from the communication system point of view, it is more desirable to use the related concepts of amplifier bandwidth and amplifier saturation power that will be evaluated below.

The amplifier gain G is defined as

$$G = \frac{P_{out}}{P_{in}} \tag{2}$$

where P_{in} is the input power and P_{out} the output power of a continuous wave (CW) signal being amplified. The amplifier gain G may be found by using the relation:

$$\frac{dP}{dz} = g(\nu, P)P \tag{3}$$

where $P(z)$ is the optical power at a distance z from the amplifier input end.

If the signal power obeys the condition $P \ll P_{sat}$ throughout the amplifier, the gain coefficient given by **Eq. (1)** can be considered independent of the signal power. In such a case, the amplifier is said to be operated in the unsaturated regime and works as a linear device. The gain coefficient presents in this situation a Lorentzian profile that is characteristic of homogeneously broadened two-level systems. However, the gain spectrum of actual amplifiers can deviate significantly from the Lorentzian profile.

The solution of **Eq. (3)** in the unsaturated regime is an exponentially growing signal power, given by

$$P(z) = P_{in}\exp(gz) \tag{4}$$

For an amplifier length L, we then find that the linear amplifier gain is

$$G(\nu) = \exp(gL) = \exp\left[\frac{g_0 L}{1 + (\nu - \nu_0)^2/\Delta\nu_0^2}\right] \tag{5}$$

Both the amplifier gain $G(\nu)$ and the gain coefficient $g(\nu)$ are maximum when $\nu = \nu_0$. However, $G(\nu)$ decreases much faster than $g(\nu)$ with the signal detuning $\nu - \nu_0$, because of the exponential dependence of G on g. As a consequence, the amplifier bandwidth $\Delta\nu_A$, which is defined as the FWHM of $G(\nu)$, is much smaller than the gain bandwidth $\Delta\nu_0$ (**Fig. 1**). This can result in signal distortion in the case where a broadband optical signal is transmitted through the amplifier. From **Eq. (5)** we can obtain the following relation between the amplifier bandwidth and the gain bandwidth:

$$\Delta\nu_A = \Delta\nu_0\sqrt{\frac{\ln 2}{g_0 L - \ln 2}} \tag{6}$$

Gain Saturation

An important limitation of the nonideal amplifier is related with the power dependence of the gain coefficient given by **Eq. (1)**. This property is known as gain saturation and it appears when the signal power ratio P/P_{sat} is non-negligible. Since the gain coefficient is reduced when the signal power P becomes comparable to the saturation power P_{sat}, the amplifier gain G will also decrease.

Assuming that $\nu = \nu_0$ and substituting g from **Eq. (1)** in **Eq. (3)** gives

$$\frac{dP}{dz} = \frac{g_0 P}{1 + P/P_{sat}} \tag{7}$$

Considering the initial condition $P(0) = P_{in}$, we obtain from **Eqs. (2)** and **(7)** the following implicit relation for the amplifier gain:

$$(1 - G)\frac{P_{in}}{P_{sat}} = \ln\left(\frac{G}{G_0}\right) \tag{8}$$

where $G_0 = \exp(g_0 L)$. The input saturation power P_{in}^s is defined as the input power for which the amplifier gain G is reduced by a factor of 2 from its unsaturated value G_0 (**Fig. 2**). Indeed, it is obtained by using $G = G_0/2$ in **Eq. (8)**:

$$P_{in}^s = \frac{2\ln(2)P_{sat}}{(G_0 - 2)} \tag{9}$$

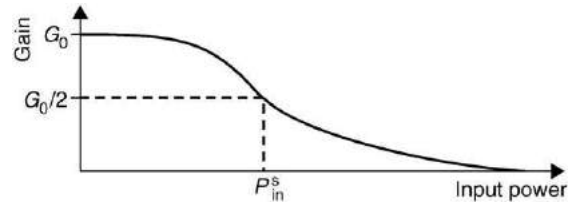


Fig. 2 Saturated amplifier gain as a function of the input power.

As observed from **Eq. (9)**, the input saturation power P^s_{in} does not coincide with P_{sat}. The output saturation power is given by $P^s_{out} = G_0 P^s_{in}/2$. In practice, $G_0 \gg 2$ and P^s_{out} is found to be smaller than P_{sat}, by about 30%.

Gain saturation can be seen as a serious limitation, particularly for multichannel communication systems. However, the self-regulating effect of gain saturated amplifiers can be useful in long-haul lightwave communication systems, including many concatenated amplifiers. In fact, if the signal level in a chain of amplifiers is unexpectedly increased along the chain, the saturation effect causes a lower gain provided by the following amplifiers and vice versa for a sudden signal power decrease.

Amplifier Noise

Besides the bandwidth and gain saturation limitations, another property must be considered concerning practical optical amplifiers. In fact, optical amplifiers always add spontaneously emitted photons to the signal during the amplification process. Those photons are amplified besides the signal photons so that, at the amplifier output, an amplified spontaneous emission (ASE) noise is presented. Since spontaneous emission always takes place, ASE noise is unavoidable and does not depend on the amplifier temperature. This is one of the most important differences between optical and electrical amplifiers, where amplifier noise is of thermal origin and can be reduced by lowering the amplifier temperature.

The ASE determines a degradation of the signal-to-noise ratio (SNR). The SNR degradation is usually characterized by the amplifier noise figure, which is defined as the SNR ratio between input and output:

$$\mathrm{NF} = \frac{\mathrm{SNR}_{in}}{\mathrm{SNR}_{out}} \tag{10}$$

The SNR is usually referred to the electrical power generated when the optical signal is converted to electrical current by using a photodetector. Therefore, the noise figure as defined in **Eq. (10)** would usually depend on several detector parameters, which determine the shot noise and thermal noise associated with the practical detector. We will consider the case of an ideal detector, whose performance is limited by shot noise only.

The SNR of the input signal is simply determined by the detection shot noise and can be written as:

$$\mathrm{SNR}_{in} = \frac{P_{in}}{2h\nu\Delta f} \tag{11}$$

where Δf is the detector bandwidth.

To evaluate the term SNR_{out}, one should add the contribution of spontaneous emission to the receiver noise. The ASE noise spectral density is assumed to be constant and can be written as

$$S_{sp}(\nu) = (G-1)n_{sp}h\nu \tag{12}$$

where G the amplifier gain and

$$n_{sp} = \frac{N_1}{N_1 - N_0} \tag{13}$$

is known as the spontaneous emission factor or the population inversion factor. In **Eq. (13)** N_0 and N_1 are the atomic populations for the ground and excited states, respectively.

Considering a low noise amplifier, the signal power impinging the photodetector is larger than the optical noise power and the shot noise power. As a consequence, the electrical noise, due to the signal-ASE beating, is the dominant contribution and the SNR of the amplified signal is given by:

$$\mathrm{SNR}_{out} \approx \frac{GP_{in}}{4S_{sp}\Delta f} \tag{14}$$

Using **Eqs. (11)–(14)**, the amplifier noise figure defined by **Eq. (10)** becomes:

$$\mathrm{NF} = 2n_{sp}\frac{G-1}{G} \approx 2n_{sp} \tag{15}$$

where the approximation holds when the gain is much higher than one. In the case of an ideal amplifier, $n_{sp} = 1$ and **Eq. (15)** show that the SNR is degraded by 3 dB. For most practical amplifiers, NF can be as large as 6–8 dB.

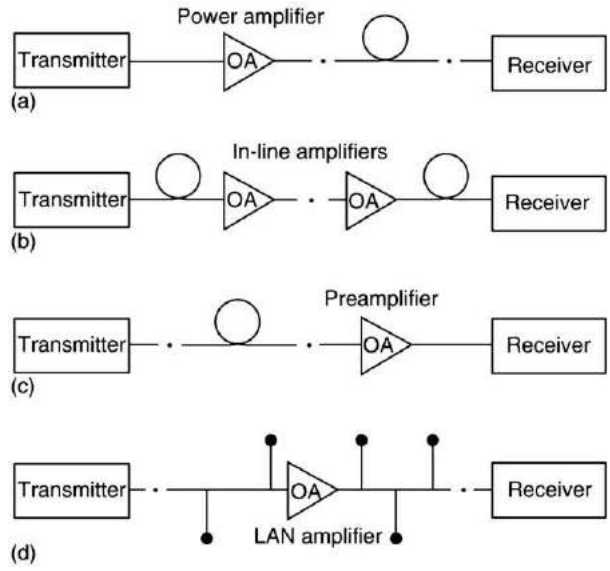


Fig. 3 Four generic configurations for incorporating optical amplifiers into transmission systems: (a) as a power amplifier; (b) as in-line amplifiers; (c) as a preamplifier; (d) for compensation of distribution losses in local-area networks.

Basic Amplifier Configurations

The relative importance of the different limiting factors discussed above depends on the actual amplifier application. **Fig. 3** shows the four basic system configurations envisioned for the incorporation of optical amplifiers. The first configuration is to place the amplifier immediately following the laser transmitter to act as a power amplifier or booster (**Fig. 3(a)**). The main purpose of such amplifiers is to boost the signal power, which can provide an increase of the transmission distance by 100 km or more. Since the signal input power is typically large (0.1–1.0 mW), the key parameter for the power amplifier will be to maximize the saturation output power and not necessarily the absolute gain.

The second configuration is to place the amplifier in-line and perhaps incorporated at one or more places along the transmission path (**Fig. 3(b)**), replacing the electronic regenerators. The in-line amplifier corrects for periodic signal attenuation and may exist in a cascade form. The use of in-line optical amplifiers is particularly attractive for multichannel communication systems, since they can amplify all channels simultaneously.

The third configuration consists of using the amplifier immediately before the receiver, so it functions as a preamplifier (**Fig. 3(c)**). The purpose of such an amplifier is to improve the receiver sensitivity. The main figures of merit are high gain and low amplifier noise, because the entire amplifier output is immediately detected.

In local-area networks (LANs), distribution losses often limit the number of possible nodes. The fourth application of optical amplifiers consists of using them for compensating such distribution losses (**Fig. 3(d)**).

See also: Optical Amplifiers: SOAs

Further Reading

Agrawal, G.P., 1992. Fiber-Optic Communication Systems. New York: Wiley.
Bjarklev, A., 1993. Optical Fiber Amplifiers: Design and System Applications. Boston, MA: Artech House.
Desurvire, E., 1994. Erbium-Doped Fiber Amplifiers: Principles and Applications. New York: Wiley.
Green Jr, P.E., 1993. Fiber Optic Networks. Englewood Cliffs, NJ: Prentice Hall.
Iannone, E., Matera, F., Mecozzi, A., Settembre, M., 1998. Nonlinear Optical Communication Networks. New York: Wiley.
Kazovsky, L., Benedetto, S., Willner, A., 1996. Optical Fiber Communication Systems. Boston: Artech House.
Shimada, S., Ishio, H. (Eds.), 1994. Optical Amplifiers and their Applications. New York: Wiley.

Erbrium Doped Fiber Amplifiers for Lightwave Systems

P Bollond, JDS Uniphase Corporation, Ewing, NJ, USA

Introduction

The telecommunications industry has undergone a revolution since the 1980s, by using glass optical fibers for the transmission of information encoded as pulses of light. A single telecommunications-grade optical fiber has been shown to support the propagation of more than 1 Tbit per second (1×10^{12} pulses per second) over distances comparable to typical city separations (>100 km). From this technology, optical fiber links have evolved to become a network on a planet-wide scale, and form the physical backbone of the information age.

Erbium-doped fiber amplifiers (EDFAs) are an enabling technology for optical fiber communication networks. They have several important properties that make them the amplification component of choice in long-distance commercial data transport networks. Erbium ions deposited in silica-based glass allow amplification in the lowest loss region of commercial-grade optical fiber (<0.25 dB/km from approximately 1530–1620 nm). Erbium-doped fiber (EDF) is manufactured in a form that allows for low-loss fusion splicing to standard communications fiber. Compact semiconductor laser diodes are available to excite the erbium ions into an amplification state. The long lifetime of the excited state of erbium provides the ability to simultaneously amplify multiple wavelength channels without significant cross-channel interference. Multiple channel systems have been deployed with more than 100 optical channels (or wavelengths) through each EDFA, and this has allowed network capacities to be dramatically increased. The low noise properties of the EDFA also allow networks to be constructed with many amplified spans before the optical signal has to be electronically regenerated. Practically unlimited transmission distance has been demonstrated using a small number of optical soliton channels through periodically amplified EDFA lightwave systems.

The development history of the EDFA can be traced back to the first optical amplifier. In 1962, a neodymium-based fiber amplifier was invented that operated at 1064 nm. During the 1980s, the need for an optical amplifier at telecommunications wavelengths initiated research at many locations throughout the world. In 1987, the University of Southampton (UK) was first to demonstrate an EDFA that had optical gain at 1550 nm, and during the following years the design of erbium-doped fiber was optimized for this application. In 1989, a practical semiconductor laser diode became available to pump EDF, and the first compact optical fiber amplifier modules soon appeared for commercial deployment. The traditional method of signal regeneration, prior to 1990, was to use electronics to detect the optical signal after each transmission span, recover the digital signal, and then retransmit using another laser diode. The EDFA allows practical wavelength division multiplexing (WDM) of multiple optical signals with all optical signal amplification, and provides significant performance and cost advantages over electronic regeneration.

EDFAs have emerged from the laboratory to be widely deployed in communication networks. EDFAs are used to boost transmitted power of the signal lasers (booster amplifier), amplify signals in transit to compensate losses sustained in the fiber (line amplifier), or amplify signals before a receiver (pre-amplifier). Typically, the amplifier module is specifically manufactured for particular systems that are mounted on electronic circuit boards. These circuit packs are then housed in central offices (local telephone exchanges), remote 'repeater huts', or even in undersea 'bottles' as part of a transoceanic cable. The high cost of network failure requires that the manufactured EDFA modules comply with stringent reliability criteria, to provide a useful operating lifetime greater than 25 years when subject to extreme environmental conditions.

Amplifiers are constructed for incorporation into either existing fiber links as part of an upgrade, or for newly planned systems. Because of the high cost of installing new fiber into the ground and securing property rights, it has become economically desirable to upgrade many existing fiber links rather than to build new systems. Transmission cables usually have many pairs of individual optical fibers, some of which will not initially be transporting data, and these 'un-lit' or 'dark fibers' can be activated as consumer demand increases over time. Typically, each fiber of a pair is used to carry either 'east'- or 'west'-bound traffic. Around city areas metropolitan area networks can be expanded in this way, but for long-distance links (long haul networks with distances >1000 km) operation is designed for a larger number of channels (40 to 120 wavelengths) at higher data rates (10 or 40 GHz per channel) over specialized transmission cables containing low numbers of fiber pairs. Typically communication systems are designed to meet certain cost targets, expressed as dollars per Gbit/s per km, for total transmission distances. The total transmission distance is limited by the optical signal to noise ratio (OSNR) degradation after each fiber-amplifier link, with a smaller permissible degradation at higher bit rates. The required gain and OSNR performance for the system is then translated to an EDFA module specification.

This article discusses EDFA design and applications, and shows the elements involved in producing reliable modules for commercial lightwave systems.

Components for EDFAs

EDFAs are comprised of passive optical components, erbium-doped fiber, and pump lasers. Passive components are chosen to meet optical and environmental specifications, while the erbium-doped fiber is selected based on the optical power, gain, and noise figure requirements. Pump lasers are a key influence on the price and performance of optical amplifiers.

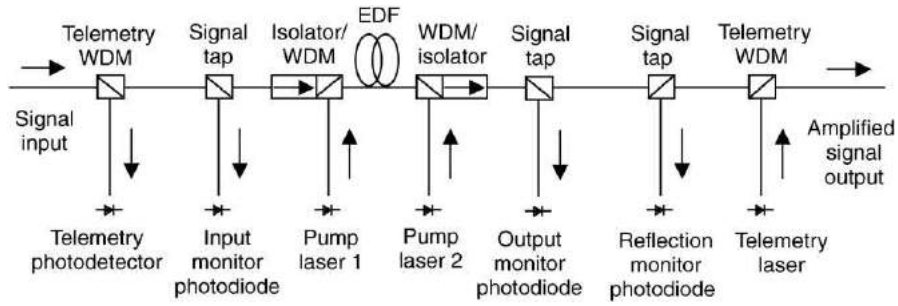


Fig. 1 Single-stage EDFA with features.

The amplifier module is typically connected to the transmission fiber using fiber connectors. These polished fiber connectors have a higher insertion loss and reflectivity than fusion splicing, but allow for easy deployment in the network. Internal optical components are fusion spliced together to provide low loss, low back reflection, high strength, and high reliability joins. Fusion splicing is tailored to particular fiber types, so that optical components with dissimilar fiber types are joined with the lowest loss.

The signal and pump radiation is combined with low loss using optical components called wavelength division multiplexers (see Fig. 1). These components are based on fused fiber or interference filter based technology. Fused fiber WDMs offer the lowest insertion loss (<0.1 dB is commercially available) but is restricted to widely spaced wavelengths (e.g., 980 nm pump and 1550 nm signal). Interference filter based WDMs are available for closely spaced wavelengths (e.g., 1480 nm pump and 1530 nm signal) and have a very low wavelength dependent insertion loss (flatness). Interference filters can be designed to produce sharp low-pass, high-pass, or bandpass type filters suitable for combining closely spaced wavelengths, as well as broader peaks suitable for lowering the gain in particular signal wavelength regions to produce gain flattened amplifiers.

Signal reflections can cause an amplifier to act as a laser, and this detrimental effect is eliminated in EDFAs by using optical isolators. In an isolator the signal light is coupled out of the single-mode fiber through a graded index (GRIN) lens and passes through a nonlinear crystal before being coupled back into the optical fiber. The isolator consists of a birefringent rutile (TiO_2) or Yttrium OrthoVanadate (YVO_4) wedge, followed by a Yttrium Iron Garnet (YIG) Faraday rotator, followed by another birefringent wedge. The YIG crystal is surrounded by a permanent magnet that rotates the light's polarization by 45 degrees. The 45-degree polarization rotation, coupled with the two birefringent wedges, ensures that light is efficiently coupled to the output fiber but not in the reverse direction. Commercial isolators are available with low insertion losses across the signal band, with some samples below 0.35 dB. Note that the YIG crystal works well for the 1480 nm pump and 1530–1620 nm signal bands, but currently there is no suitable isolator material that covers 980 to 1550 nm and this puts some limitations on certain EDFA designs.

The isolator design has been extended to make multiport circulators. A three-port circulator has the input into port 1 and output of port 2, light entering port 2 is directed to port 3, and light entering port 3 is blocked with an isolator. The circulator allows for adding and dropping of individual channels when a narrow bandwidth reflective filter is placed on port 2. Circulators also can be used to separate co-propagating and counter-propagating traffic on a single transmission fiber before amplification is done.

Erbium-doped fiber has two strong absorption bands around 980 nm and 1480 nm that are suitable for achieving a population inversion in the erbium ion. The 980 nm wavelength allows low noise amplification, while 1480 nm lasers provide higher EDFA output. 980 nm pump lasers usually incorporate fiber Bragg gratings to stabilize the laser wavelength to match the narrow EDF absorption peak, and also have lower drive current requirements. A 1480 nm pump can provide more amplification, since there are more photons in each mW of laser power at 1480 nm than 980 nm. High-output power pump lasers incorporate thermo-electric coolers (TECs) within the pump laser package and can provide fiber coupled power greater than 500 mW. Both high-power lasers have been qualified to meet the most stringent reliability standards, with typical mean time before failure beyond 25 years. Low-cost 980 nm pump lasers, that do not have TECs, are also available in smaller packages and can supply up to $\sim$200 mW. By using wavelength division or polarization combining techniques, it is possible to further increase the available pump power in the erbium fiber.

The EDFA module is assembled into a package that may contain passive optical components, several meters of coiled EDF, pump lasers, photo-detectors, and electronic circuit boards. In some cases, it is advantageous to thermally stabilize the components so that the amplifier performance can be maintained when exposed to extreme environmental conditions. This may occur when the central office's environmental control is compromised (e.g., air conditioning failure). In particular, EDF can exhibit undesirable spectral gain changes if the ambient temperature changes by more than 5°C.

Pump lasers with internal TECs can dissipate more than 5 Watts of heat per laser and since cooling fans usually do not have the required reliability, passive cooling is commonly used in the module in the form of a metal heat sink. The amplifier module's size can be compact, limited by the height of optical components or pump laser diodes, or by the size of a built-in heat sink. The module is designed to survive conditions of electrostatic discharge, humidity, temperature, thermal shock, extreme vibration, and other stresses that may be inadvertently present during operation in the field. In addition, all material in the EDFA module is also qualified against problems with out-gassing (e.g., hydrogen release), combustion and chemical or biological exposure.

The Single-Stage Amplifier

A simple single-stage EDFA consists of an erbium-doped fiber spool with signal and pump combining multiplexer. The fiber is optically pumped by 980 nm and/or 1480 nm laser(s), whose light is coupled into the signal fiber by a passive component called an optical multiplexer. The pump wavelengths are readily absorbed by erbium ions embedded in silica raising them to an excited state. Amplification occurs when, stimulated by a nearby signal photon, an excited erbium ion relaxes back to the ground state producing a second, identical signal photon. The erbium ion can be approximated as a three-level atomic system that can be completely inverted by a 980 nm photon, to provide the lowest noise amplification. In contrast, the 1480 nm pump will excite the erbium ion directly into the upper laser level as a two-level system, and because of rapid spontaneous emission from this level, the maximum inversion in this case cannot exceed approximately 75%. Note that as the pump photons are absorbed, the inversion will be nonuniform along the EDF length.

There are two signal wavelength regions commonly amplified by EDFAs, the C-band (conventional band) from approximately 1528 to 1565 nm, and the L-band (long band) from approximately 1570 to 1620 nm. Amplification in the C-band readily occurs when moderate pump power is available, and relies on the erbium ion's spectral absorption and emission wavelength window that is suited to high levels of atomic inversion. L-band amplification is also achieved with moderate pump powers, but because of the lower absorption and emission cross-sections, similar gain is reached using approximately five times more EDF with a lower average inversion. The C-band amplifier is typically less costly because less EDF is used, while high-concentration erbium fibers are available specifically for L-band EDFAs.

An EDFA's most critical performance parameters are its amplified signal output power (typically stated in dBm) and its noise figure (stated in dB). Output power is mainly determined by total pump power and the amplifier internal loss. The noise figure (NF) is defined as the ratio of the signal-to-noise ratio at the input to the signal-to-noise ratio at the output.

A single-stage amplifier typically has 1 or 2 pump lasers but can have more if polarization- or wavelength-pump-combining is implemented for higher power. When the pump radiation propagates in the same direction as the signal, the amplifier is co-pumped, while counter-pumping denotes the case when the pump laser propagates against the signal. For a single pump, a co-pumping 980 nm laser minimizes the NF (suitable for a pre-amplifier) while counter-pumped 1480 nm architecture optimizes output power at some expense to the NF (suitable for a booster amplifier). Recent semiconductor pump laser improvements have enabled 980 nm pump lasers with output powers >500 mW to be commercially available, allowing most single- or dual-stage EDFAs to be energized by one laser.

Single-stage designs can be enhanced, as shown in **Fig. 1**. To control the network an optical signal may be used as a telemetry or network supervisory channel, and this is removed with a filter. Telemetry wavelengths are usually outside of the useful EDF amplification window, and commonly range from 1500–1520 nm and from 1620–1640 nm. An isolator may be used at the input and/or output to prevent pump laser or amplified spontaneous emission from the erbium-doped fiber 'leaking' into the transmission path. Optical taps may be included to provide information about signal spectra at the input and output sides. Their feedback can be used to control pump laser biases for tuning output power, monitoring amplifier performance, or simply to trigger alarms. In addition, a reflection monitor is sometimes placed at the output to observe backwards propagating optical signals arising from reflections. Electronics in the module will continuously monitor the pump lasers and photodetectors and, for example, can place the module in an 'eye-safe' low-output power (<10 mW) mode within milliseconds if a transmission fiber break is detected through increased back-reflected optical power.

The amplifier has a signal input power of -30 to -10 dBm (1–100 μW) for each optical channel, and has a gain of 25 dB to compensate for a typical span loss of 100 km optical fiber link. This signal level allows high powers at the receiver photodetector for high-quality signal detection, yet is low enough to avoid nonlinear propagation effects in the transmission fiber. The typical total output signal power of an 80 channel (wavelength) C-band EDFA is less than 200 mW ($+23$ dBm) to avoid stimulated Raman scattering (SRS) in standard transmission fiber. The use of improved low nonlinearity transmission fibers and longer transmission distances can lead to specified total EDFA output powers to be greater than 400 mW ($+26$ dBm). To provide the best performance, the amplifier usually will allow only a single direction of propagation, with bi-directional communications systems using one transmission fiber and circulators to separate 'east' and 'west' traffic into individual EDFAs.

Fig. 2 shows the results of a numerical simulation for a single-stage amplifier as a function of the input power and EDF length. The amplifier was assumed to have EDF with peak absorption of 5.5 dB/m near 1530 nm, and was pumped with 100 mW at 980 nm. The amplifiers signal loss before and after the EDF was taken to be 0.8 and 1.2 dB, respectively. As amplifier input power increases there is a decrease in gain provided by the medium since there is a fixed amount of pump power available. The maximum gain is usually achieved near 1530 nm where the difference between the EDF's emission and absorption cross-sections is greatest. Selecting the length of EDF is critical to achieving the desired performance, and this can be examined using numerical simulation for a wide range of design options.

Undersea systems, with their long distances and large costs of network failure, place stringent design requirements on EDFAs. These systems mostly use single-stage designs with emphasis on low noise operation. This can be achieved by eliminating most of the optical components prior to the EDF, and also by co-propagating a strong 980 nm pump using a low loss fused fiber WDM. Undersea repeaters are spaced by 30 to 80 km, shorter than terrestrial networks, with each channel operated at higher power to achieve multi-thousand kilometer distances. For example, the trans-Pacific TPC-5J cable spans 8600 km from Coos Bay, Oregon (USA) to Ninomiya (Japan), using EDFAs spaced every 33 km. The tight electrical power budgets available to each repeater (powered from land) necessitate using pump lasers without TECs. During installation, the EDFA will experience large mechanical

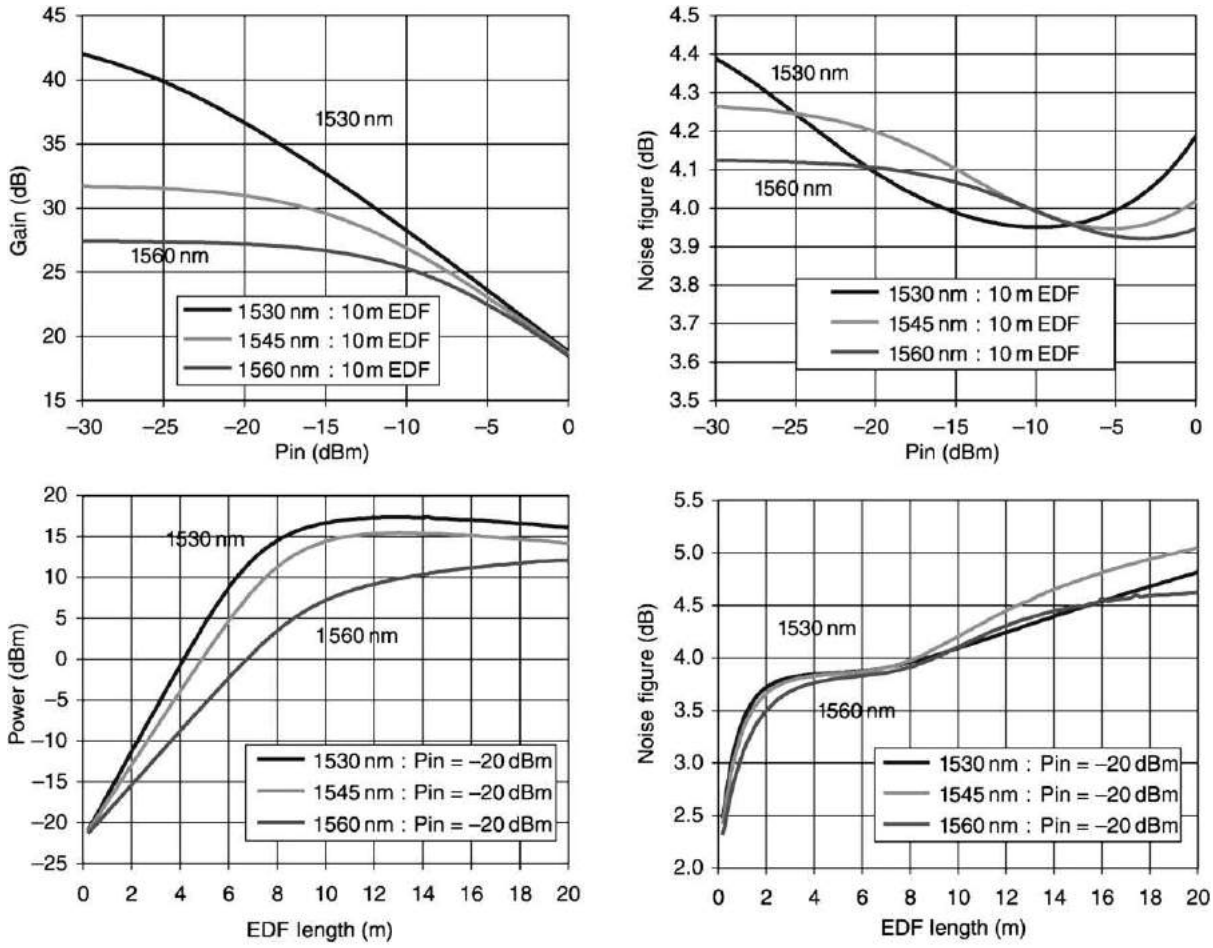


Fig. 2 Typical single-channel performance of a single-stage EDFA.

shock, as the cable and metal repeater bottles are unwound aboard ship and dropped into the ocean. Other design considerations are done for operation at the constant ambient temperature of the ocean bottom ($\sim +2$ to 4°C) or for seasonal temperature changes on the continental shelves, where the optics and EDF will operate at 15 to 30 degrees above ambient due to heating from the electronics.

Single-stage optical amplifiers are suited for a wide range of applications such as single-channel amplifiers, simple WDM amplifiers, and low-cost amplifiers. Using high-power pump lasers or combined pump laser schemes allows such amplifiers to deliver output powers >20 dBm. However, single-stage amplifiers cannot meet the requirements of all telecommunications architectures, leading to increasing demand for multiple-stage EDFAs that are discussed below.

Multiple-Stage EDFAs

The most common implementation of a multiple-stage EDFA is to improve the noise and output power characteristics by imbedding an optical isolator between two sections of EDF. The isolator blocks backward-traveling amplified spontaneous emission to improve the efficiency of the amplifier. This is shown in Fig. 3 for two common single-pump designs. The first is the 'pump by-pass', design where the residual pump radiation from the first stage is redirected around the mid-stage isolator. This is often used when pumping with 980 nm since signal band optical isolators will not transmit 980 nm radiation. This design can introduce problems when using low-quality components as small signal levels can propagate around the pump by-pass fiber to create multiple path interference (MPI) effects. Although MPI effects can be eliminated by using a pump splitting coupler to directly pump each EDF section, the pump by-pass design makes the most efficient use of available pump power. The second design is called the 'pump through' design, and is used when pumping at 1480 nm since both the pump and signal band (e.g., 1550 nm) photons will transmit through commercial isolators with only small loss. Note that for cost and space constraints within the module, it is sometimes advantageous to use hybrid optical components, e.g., an isolator and WDM can be combined into one compact package.

An EDFA with multiple input signals will have a very nonuniform output gain profile. This is a consequence of the erbium ion's wavelength-dependent emission and absorption cross-sections in the host glass material. Fig. 4 shows gain spectra for both C-band

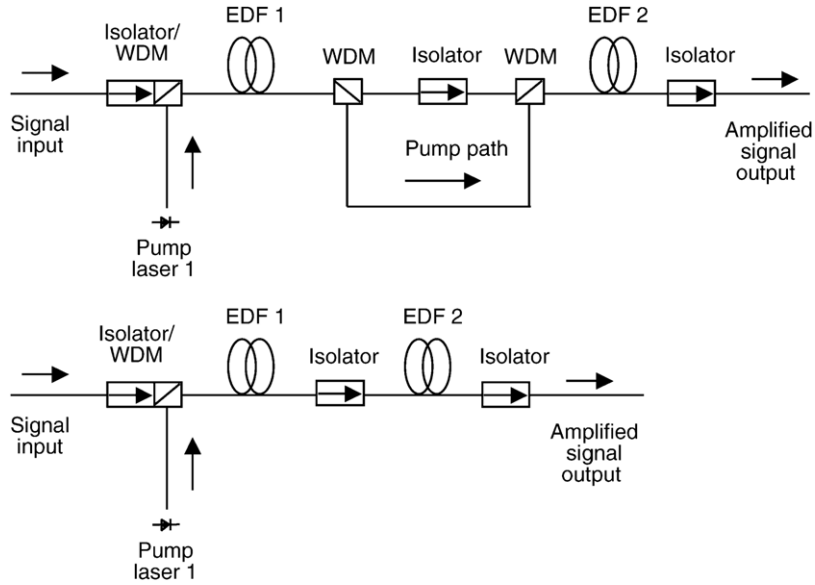


Fig. 3 Imbedded Isolator EDFA designs.

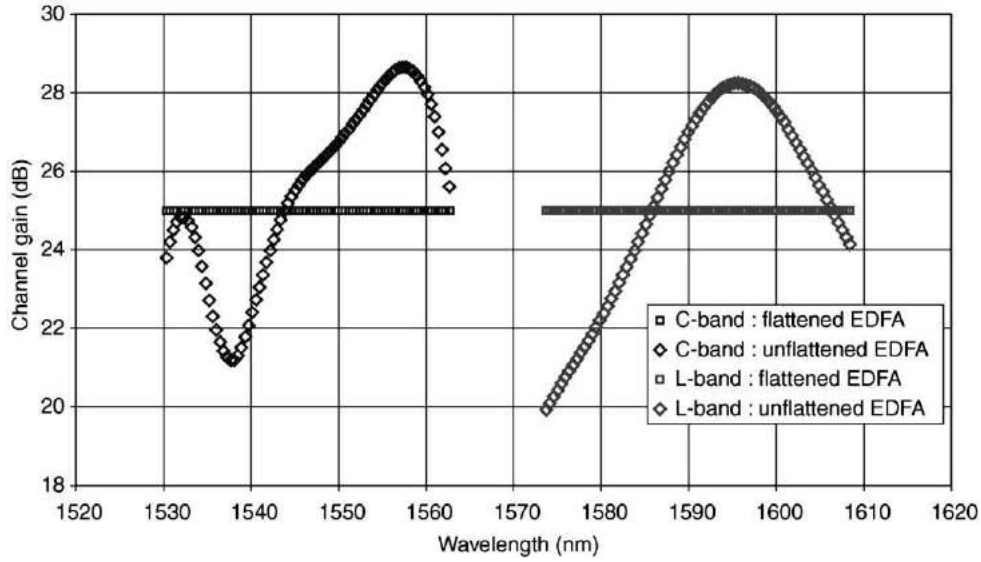


Fig. 4 Gain spectrum of a multi-staged optical amplifier with and without gain flattening filters.

and L-band EDFAs with and without gain-flattening filters (GFFs). The gain spectrum is dependent on the EDF's chemical composition and EDFA design features. For an amplifier with 25 dB gain, a GFF with peak loss less than 10 dB is usually needed to correct for these gain deviations. Although many technologies are available for GFFs (e.g., thin-filter interference filters, Bragg gratings, tapered fiber filters, etc.) the basis function of these technologies is usually not identical to the EDF's gain profile, and this mismatch results in gain flatness error. However, when using these technologies, manufactured EDFAs can have a gain error (flatness) less than 1.0 dB over bandwidths greater than 35 nm.

Early EDFAs for WDM systems used 16 optical channels spaced by 200 GHz in the 1540 to 1560 nm part of the C-band. In this case, the EDF gain profile is relatively smooth and no GFFs are needed to achieve a 1.0 dB gain flatness specification. It is important to note that for a particular level of signal and pump power, a longer length of EDF in the EDFA will produce more gain at the long wavelength end of the spectrum, hence by shortening the EDF length in Fig. 4 the result will be relatively flat gain from approximately 1540 to 1560 nm. As EDFA technology has matured, the EDF can be the bandwidth-limiting component in a system, and it has become necessary to use GFFs to realize useable bandwidths of up to 55 nm when using regular silica-based EDF. The most economic EDFAs operate in the C-band where relative to the L-band, pump laser efficiency is highest, EDF lengths are shortest, and the chromatic dispersion of most installed transmission fiber limits nonlinear optical effects.

In single-stage amplifiers, the gain-flattening filter is at the amplifier output, and this results in lost output power. Wideband gain-flattened amplifiers usually have a multiple EDF stage design, with the GFF component between two EDF stages. The EDF sections that follow the attenuating GFF provide additional amplification, to give high power out of the EDFA module that has large internal losses.

In addition to GFF components inside the EDFA, an attenuator can be used to compensate for input signal power changes, which will produce a spectral gain rotation or tilt. The attenuator reduces the power on all optical channels equally and allows the amplifier to operate in a 'fixed gain' mode. In WDM systems, optical amplifiers may also need mid-stage access where the signal is fed into an external device between the amplifier stages. Reasons for doing this include monitoring, adding or dropping of individual channels, and dispersion compensation. Since additional losses up to 10 dB are introduced in the amplification path, the amplifier design has to be optimized for those losses.

Typical two-stage gain-flattened optical amplifier architecture is shown in **Fig. 5**. Input and output couplers and telemetry WDMs can be included if required. A single pump can be split and shared between stages to save cost. When multiple pumps are needed, the most common configuration is a 980 nm pump laser co-pumping the first stage (EDF1) for low noise and a 1480 nm laser counter-pumping the second spool (EDF2) for high gain. A significant loss element, e.g., a gain-flattening filter, add/drop module or dispersion compensation module, is situated between the stages. Note that mid-stage access can be located before or after the gain flattening or between additional EDF stages. The gain spectrum of multistage gain-flattened EDFAs is shown in **Fig. 6**, showing the gain equalization possible using a multistage amplifier design and gain-flattening filters. Using deeper GFFs can increase the usable bandwidth of the amplifiers, but this requires higher-power pump lasers to maintain the same gain and optical signal-to-noise ratio (OSNR) performance.

The two-stage EDFA shown in **Fig. 5** highlights a common problem for amplifier design. Given high-grade optical components and pump lasers, what length of erbium-doped fiber should be used to give the best amplifier performance? This question is best answered by using the results of extensive numerical simulations of the optical amplifier.

Fig. 7 shows the results of numerical simulation for a two-stage amplifier as a function of the first and second EDF stage lengths. The amplifier was assumed to have EDF with a peak absorption of 5.5 dB/m near 1530 nm, and was pumped with 130 mW at 980 nm in the first stage and 160 mW at 1480 nm in the second stage. The total input power was assumed to be -2.5 dBm for 80 channels distributed from 1530 nm to 1563 nm. Note that each of the 80 channels is separated by 100 GHz, and will generally support a long-distance communications system with each wavelength modulated at 10 GHz.

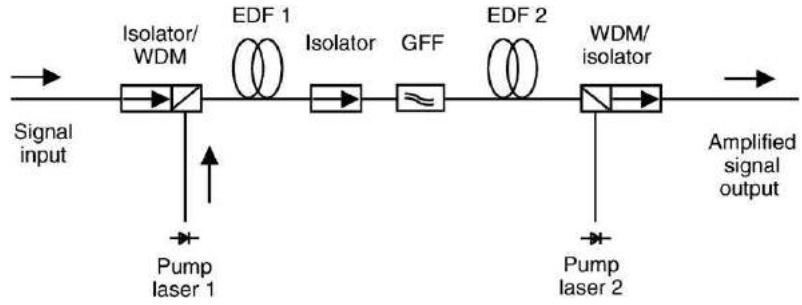


Fig. 5 Typical multistage gain-flattened optical amplifier architecture.

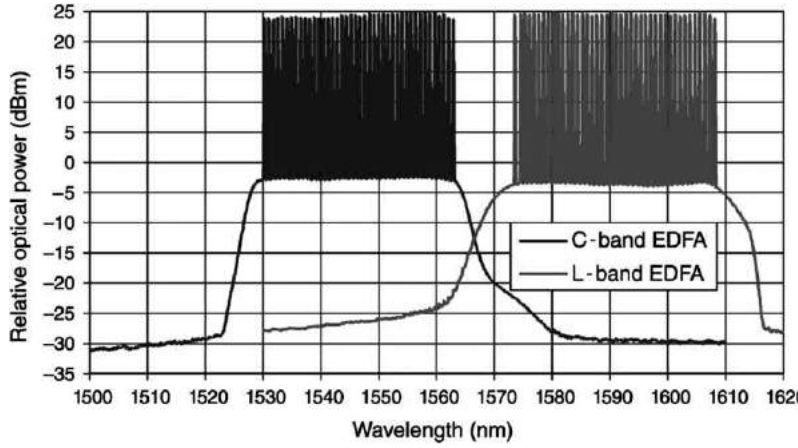


Fig. 6 Measured gain and noise spectra of multistaged gain-flattened C-band and L-band EDFAs.

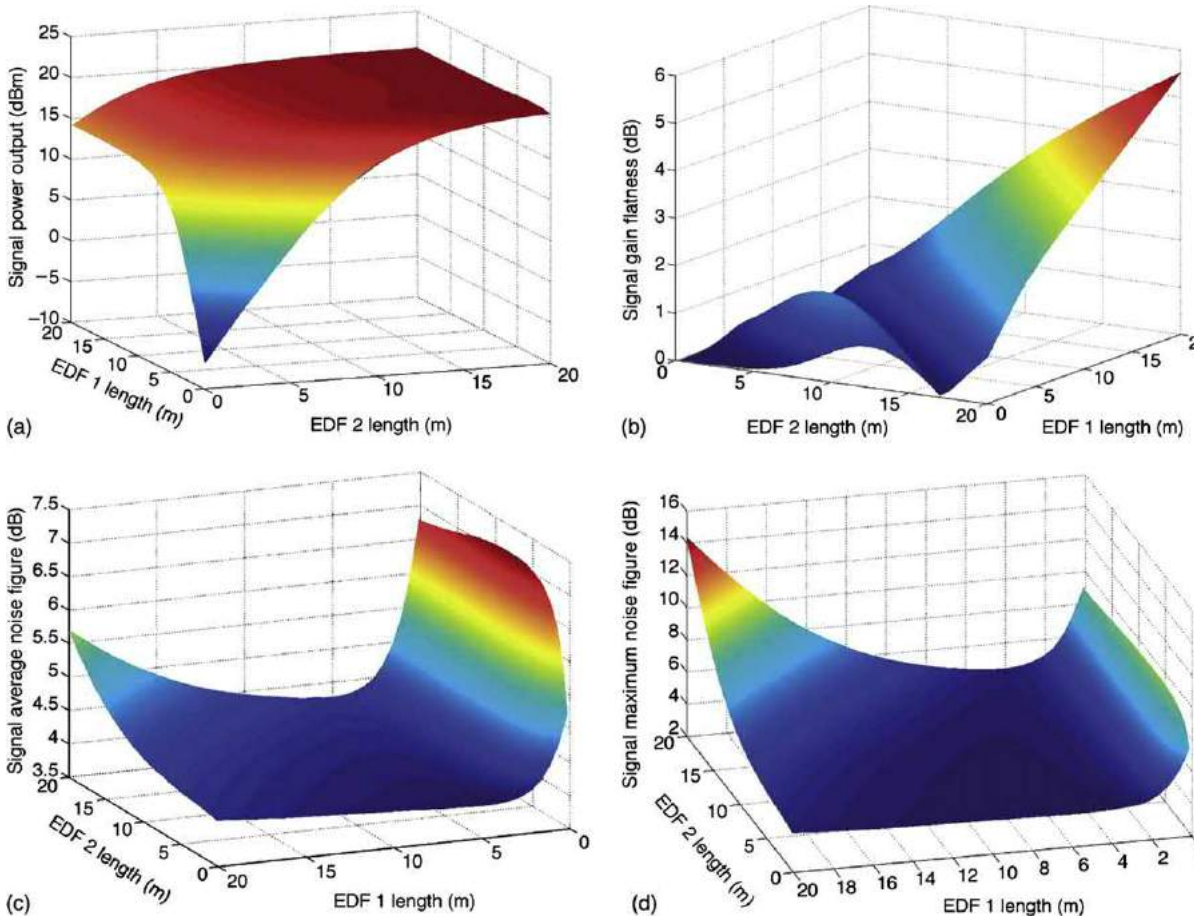


Fig. 7 (a) Total output power, (b) spectral flatness, (c) average noise figure, and (d) maximum channel noise figure of a two-stage optical amplifier with a gain-flattening filter.

Fig. 7(a) shows that the optical output power is largest when both EDF stages are approximately 20 meters. For this design, EDF lengths significantly longer than 20 meters will produce less output power, since the pump radiation will have been completely absorbed in the first few meters of EDF. The spectral gain flatness is shown in Fig. 7(b). For this design, approximately 17 meters of EDF was needed to produce a flat gain spectrum, i.e., each optical channel had the same gain. Other combinations of EDF lengths were not matched to the particular gain-flattening filter and produced large gain variations across the band. In general as the amplifier's total EDF length increases the gain spectrum exhibits a positive tilt (e.g., longer wavelength channels will experience more gain). Figs. 7(c, d) show the average channel and maximum channel NF, respectively. From a system design perspective, a link can be limited by the optical channel with the lowest optical signal-to-noise ratio (OSNR) or highest noise figure, and this is a critical parameter of interest when designing optical amplifiers. In practice the best amplifier design is a compromise between high-gain, low-gain flatness, and low NF. From extensive numerical simulations, for the two-stage gain flattened amplifier example it was for approximately 4 meters in EDF stage 1, and 13 meters in stage 2.

Successful commercial EDFA products are more than a single amplifier design that can accommodate all applications. In some cases an economical solution is a modular design approach, where smaller capacity, lower-cost amplifiers are initially installed in a network. Additional gain can be added to the basic amplifier for longer-distance spans or as network traffic increases over time. This 'pay as you grow' approach can be done using additional gain stages, or pump lasers, and can be upgraded without interruption of revenue-generating network traffic.

Low-Noise Design of EDFAs

A key design feature of commercial EDFAs is low-noise operation since the degradation of the OSNR limits the reach of the system. The NF of an EDFA can be defined as

$$\mathrm{NF} = \mathrm{SNR}_{\mathrm{in}}/\mathrm{SNR}_{\mathrm{out}} \tag{1}$$

The NF can also be calculated from

$$\mathrm{NF} = \frac{1}{G} \times \left[\frac{P_{\mathrm{ASE}}(\nu_s, L)}{h\nu_s} + 1 \right] \tag{2}$$

where P_{ASE} is the amplified spontaneous emission (ASE) noise power at the signal frequency ν_s for an EDFA of gain G (h = Plancks constant). An alternative equation for the NF for a single section of EDF is

$$\mathrm{NF} = \int_0^L \frac{\gamma_e(\nu_s, z) - \gamma_a(\nu_s, z)}{P_s(z)} P_s(0) \mathrm{d}z + 1 \tag{3}$$

where γ_e and γ_a are the EDF emission and absorption factors at the signal frequency ν_s, at distance z along the amplifying fiber. **Eq. (3)** shows that low NF can result from rapid signal gain in the initial fiber along the EDF. This result indicates that co-propagating pump and signal photons, both into the same end of the EDF, combined with the maximum inversion obtained by using a 980 nm pump, will result in low-noise amplification. The NF is typically expressed in dB units, and using a full quantum mechanical theory, the lowest possible NF for a fully inverted EDF with a large signal gain can be shown to be approximately 3 dB. Actual EDFAs will only approach the '3 dB quantum noise limit' when the signal is significantly lower than the available pump power (e.g., low channel count applications) so that high inversion can be maintained along the EDF length. Optical components in the EDFA prior to the first EDF will attenuate the signal but not the noise, since ASE is only generated in the EDF, and this input stage loss will directly add to the EDFA's NF.

The use of 1480 nm pump lasers can also result in a further NF penalty of 0.3 to 1.5 dB since complete inversion cannot be achieved when the erbium ions are operating as a two-level atomic system and more ASE is generated. The ASE, like the signal gain, has a wavelength-dependent spectral profile and this results in each channel having a different NF. Typically, the optical channel with the largest NF (lowest OSNR) is used to define the EDFA's noise performance. EDFAs with low-input signal power (e.g., -15 dBm) and high gain (e.g., 25 dB) may be described as low noise when the NF is <4 dB. For large total input signal power (e.g., $+2.5$ dBm for high channel count operation) low noise may be NF <5 dB.

For an amplifier with multiple EDF sections, the noise figure can be calculated from

$$\mathrm{NF}_{\mathrm{System}} = \frac{\mathrm{NF}_1}{L_1} + \frac{\mathrm{NF}_2}{L_1 G_1 L_2} + \ldots + \frac{\mathrm{NF}_N}{L_1 G_1 L_2 G_2 \ldots L_{N-1} G_{N-1}} \tag{4}$$

where L_i is the signal loss prior to the gain G_i of the EDF stage i. From **Eq. (4)**, multistage amplifiers with large inter-stage losses will have numerically larger NFs, but the total module NF is dominated by the first EDF stage. Undersea EDFAs, with a single short EDF section, very small L_i and 980 nm pumping, can have NF <3.5 dB.

Advanced EDFA Functions

An important trend in optical amplifiers, and systems in general, is the move towards dynamic control. Several technologies are being considered and the primary idea is to dynamically control the intensity level of each optical channel or wavelength to correct for any gain/loss inequalities in the network. This is particularly important as channel counts and bit rates rise concurrently with the functionality demanded of optical communication networks. A common requirement in this category is fast amplifier response so that in a network where individual optical channels can be added or dropped there are no power distortions at other surviving wavelengths. In a typical drop event, in less than 1 microsecond, 31 of 32 wavelengths can be completely removed from the input to the EDFA by an add/drop module located elsewhere in the network. This event will cause a total input power decrease of 15 dB and without rapid pump power adjustment the remaining channel will be excessively amplified.

Erbium ions have an excited state lifetime of approximately 10 ms, and this sets the time-scale for the transient signal response. A network transient usually manifests as a sharp optical amplifier (OA) gain fluctuation lasting up to several tens of microseconds that will pulse large signal levels through the entire optical network. Dynamic network loading necessitates controlling temporal gain transients, and this may be achieved by using a microprocessor in the amplifier module. These intelligent EDFA modules can rapidly adjust the pump laser power based on feedback from internal photo detectors to control gain transients.

The need for dynamic control also exists in the frequency domain. Wideband amplifiers have a gain variation among the optical channels of typically up to 1 dB over all operating conditions, and since these EDFAs are concatenated over many spans the electronic receiver circuit at the end of the network will have to detect optical channels with widely varying power levels. This can limit the number of spans that the network can bridge before electronic regeneration becomes necessary. This problem can be corrected using a dynamic gain equalizing filter (DGEF) to actively attenuate individual optical channels levels. A DGEF can be embedded inside an amplifier to create a loss-less component, and can be used to equalize channel powers for an entire network.

The problems of accumulation of gain ripple become more apparent in systems that employ amplifiers based on stimulated Raman scattering. A conventional EDFA amplified system may have 10 spans transmitting 100 optical channels at 10 Gbit/s each over a distance of approximately 800 km before electronic regeneration becomes necessary. To span longer distances with higher capacity, Raman amplifiers are used to amplify the transmission fiber along with EDFAs. This system architecture has the advantage of a higher effective OSNR for the individual spans, and has been demonstrated for distances over 4000 km with 40 Gbit/s channels. However, the variability of the Raman amplifier gain, due to network loading and differing transmission fiber properties, necessitates the use of loss-less DGEFs.

High-capacity transmission uses individual channels that are each modulated up to typical 2.5, 10, or 40 GHz, that requires that the optical amplifiers have low polarization mode dispersion (PMD) and low chromatic dispersion (CD). Polarization mode dispersion distorts the temporal spread of an ultra-short pulse as different polarization states travel at different speeds through the transmission fiber and amplifier components. PMD is particularly a problem in the first generation of installed transmission fiber and through polarization components (e.g., isolators) that are not PMD compensated. Without PMD compensation, 40 Gbit/s transmission, for example, can be limited to very short distances (several km). Fortunately, PMD compensators have been developed for upgrading aged installed networks, and also amplifiers are available with very low PMD.

To achieve long-distance transmission, a chromatic dispersion map is produced for all the optical spans, and by using a dispersion compensating module in each amplifier, the net dispersion across the channel wavelength window is controlled. Although optical amplifiers have approximately 1000 times less optical fiber in them compared to the transmission span that it is amplifying, chromatic dispersion in the amplifier can also be of concern for very long-haul networks. The CD of the amplifier is primarily in the EDF, and is specific to the EDF's geometric and chemical composition. Furthermore, since the erbium ion is a resonant atomic system, there may be a pump and signal power contribution to the dispersion (resonant dispersion) that could degrade network performance. Characterization of these optical effects is typically done on manufactured 'field grade' amplifiers as part of a quality control process.

Conclusion

As optical networks evolve EDFAs will continue to be an enabling technology for higher capacity and more dynamic communications networks. EDFA technology has already advanced to accommodate multiple channels, to span several wavelength windows and to provide features such as dispersion compensation, gain-transient suppression, and dynamic gain flatness. As future network architectures are introduced, to incorporate Raman amplification and additional wavelength windows, the demands placed upon EDFA design will continue to expand.

See also: Optical Amplifiers: SOAs

Further Reading

Agrawal, G., 1989. Nonlinear Fiber Optics. San Diego, CA: Academic Press.
Agrawal, G., 1997. Fiber Optic Communication Systems. New York: John Wiley and Sons.
Becker, P.C., Olsson, N.A., Simpson, J.R., 1999. Erbium-doped Fiber Amplifiers – Fundamentals and Technology. San Diego, CA: Academic Press.
Desurvire, E., 1995. Erbium Doped Fiber Amplifiers. New York: John Wiley and Sons.
Pedersen, B., *et al.*, 1991. The design of erbium-doped fiber amplifiers. Journal of Lightwave Technology 9 (9), 1105.
Sudo, S., 1997. Optical Fiber Amplifiers: Materials, Devices and Applications. Boston, MA: Artech House.

All-Optical Signal Regeneration

O Leclerc, Alcatel Research & Innovation, Marcoussis, France

Introduction

The breakthrough of optical amplification, combined with the techniques of wavelength division multiplexing (WDM) and dispersion management, have made it possible to exploit a sizeable fraction of the optical fiber bandwidth (several terahertz). Systems based on 10 Gbit/s per channel bit-rate and showing capacities of several terabit/s, with transmission capabilities of hundreds or even thousands of kilometers, have reached the commercial area.

While greater capacities and spectral efficiencies are likely to be reached with current technologies, there is potential economic interest in reducing the number of wavelength channels by increasing the channel rate (e.g., 40 Gbit/s). However, such fourfold increase in the channel bit-rate clearly results in a significant increase in propagation impairments, stemming from the combined effects of noise accumulation, fiber dispersion, fiber nonlinearities, and inter-channel interactions and contributing to two main forms of signal degradation. The first one is related to the amplitude domain; power levels of marks and spaces can suffer from random deviations arising from interaction between signal and amplified spontaneous emission (ASE) noise or with signals from other channels through cross-phase modulation (XPM) from distortions induced by chromatic dispersion. The second type of signal degradations occurs in the time domain; time position of pulses can also suffer from random deviations arising from interactions between signal and ASE noise through fiber dispersion. Preservation of high power contrast between '1' and '0', and of both amplitude fluctuations and timing jitter below some acceptable levels, are mandatory for high transmission quality, evaluated through bit-error-rate (BER) measurements or estimated by Q-factors. Moreover, in future optical networks, it appears mandatory to ensure similar but high optical signal quality at the output of whatever nodes in the networks, as to enable successful transmission of the data over arbitrary distance.

Among possible solutions to overcome such systems limitations is the implementation of Optical Signal Regeneration, either in-line for long-haul transmission applications or at the output of network nodes. Such Signal Regeneration performs, or should be able to perform, three basic signal-processing functions that are Re-amplifying, Re-shaping, and Re-timing, hence the generic acronym '3R' (Fig. 1). When Re-timing is absent, one usually refers to the regenerator as '2R' device, which has only re-amplifying and re-shaping capabilities. Thus, full 3R regeneration with retiming capability requires clock extraction.

Given system impairments after some transmission distance, two solutions remain for extending the actual reach of an optical transmission system or the scalability of an optical network. The first consists in segmenting the system into independent trunks, with full electronic repeater/transceivers at interfaces (we shall refer to this as 'opto-electronic regeneration' or O/E Regeneration forthwith). The second solution, i.e., all-optical Regeneration, is not the optical version of the first which would have higher

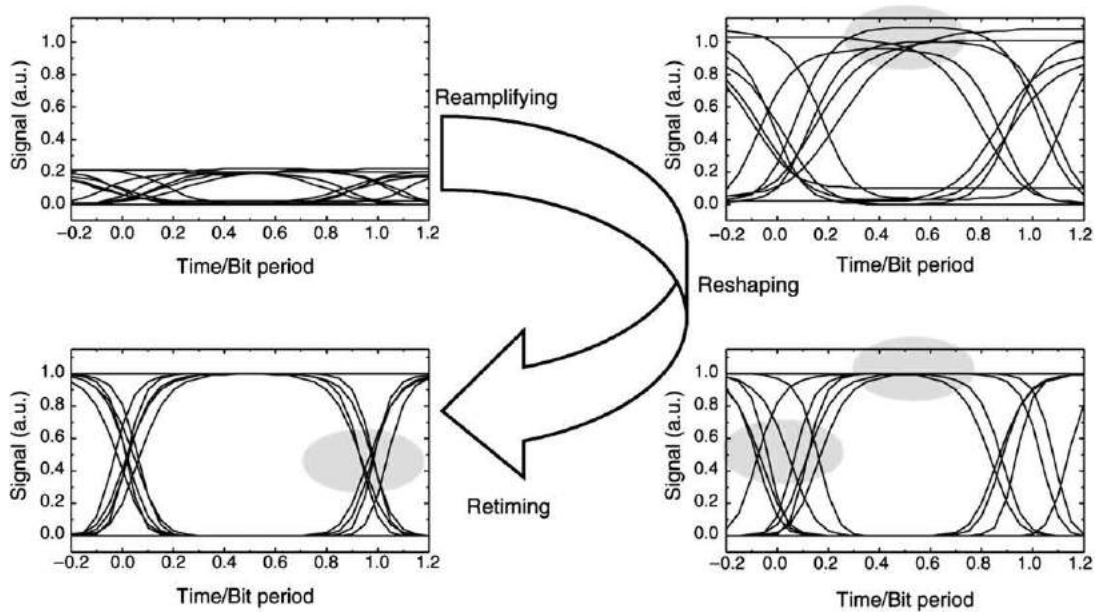


Fig. 1 Principle of 3R regeneration, as applied to NRZ signals. (a) Re-Amplifying; (b) Re-Shaping; (c) Re-Timing. NB: Such eye diagrams can be either optical or electrical eye diagrams.

 doi:10.1016/B978-0-12-809283-5.00669-8

bandwidth capability but still performs the same signal-restoring functions with far reduced complexity. At this point, it should be noted that Optical 3R techniques are not necessarily void of any electronic functions (e.g., when using electronic clock recovery and O/E modulation), but the main feature is that these electronic functions are narrowband (as opposed to broadband in the case of electronic regeneration).

Some key issues have to be considered when comparing such Signal Regeneration approaches. The first is that today's and future optical transmission systems or/and networks are WDM networks. Under this condition, the WDM compatibility – or the fact that any Regeneration solution can simultaneously process several WDM channels – represents a key advantage. The maturity of the technology – either purely optical or opto-electronic – also plays an important role in the potential (pre-)development of such solutions. But the main parameter that will decide the actual technology (and also technique) relies on the tradeoff between actual performance of the regeneration solutions and their costs (device and implementation), depending on the targeted applications (long-haul system, medium haul transport, wide area optical network, etc.).

In this article, we review the current alternatives for all-optical Signal Regeneration, considering both theoretical and experimental performance and practical implementation issues. Key advantages and possible drawbacks of each solutions are discussed, to sketch the picture in this field. However, first we must focus on some generalities about Signal Regeneration and the way to define (and qualify) such regenerator performance. In a second part, we will detail the currently-investigated optical solutions for Signal Regeneration with a specific highlight for semiconductor-based solutions using either semiconductor optical amplifiers (SOA) technology or newly-developed saturable absorbers. Optical Regeneration techniques based on synchronous modulation will also be discussed in a third section. The conclusion will summarize the key features of each solution, so as to underline the demanding challenge optical components are facing in this application.

Generalities on Signal Regeneration

Principles

In the general case, Signal Regeneration is performed using a decision element exhibiting a nonlinear transfer function. Provided with a threshold level and when associated with an amplifier, such an element then performs the actual Re-shaping of the incoming data (either in the electrical or optical domain) and completes a 2R Signal Regenerator. **Fig. 2** shows the generic structure of such a Signal Regenerator in the general case as applied to non return to zero (NRZ) data. A clock recovery block can be added (dotted lines) to provide the decision element with time reference and hence perform the third R (Re-timing) of full Signal 3R Regeneration. At this point, it should be mentioned that the decision element can operate either on electrical signals (standard electrical DFF) provided that optical → electrical and electrical → optical signal conversion stages are added or directed onto optical signals using the different techniques described below. The clock signal can be of an electrical nature, as for electrical decision element in O/E regenerator – or either an electrical or a purely optical signal in all-optical regenerators.

Prior to reviewing and describing the various current technology alternatives for such Optical Signal Regeneration, the issue of the actual characterization of regenerator performance needs to be explained and clarified. As previously mentioned, the core element of any Signal Regenerator is the decision element showing a nonlinear transfer function that can be of varying steepness. As will be seen in **Fig. 3**, the actual regenerative performance of the regenerator will indeed depend upon the degree of nonlinearity of the decision element transfer function.

Fig. 3 shows the principle of operation of a regenerator incorporating a decision element with two steepnesses of the nonlinear transfer function. In any case, the '1' and '0' symbols amplitude probability densities (PD) are squeezed after passing through the decision element. However, depending upon the addition of a clock reference for triggering the decision element, the symbol arrival time PD will be also squeezed (clocked DECISION = 3R regeneration) or enlarged (no clock REFERENCE = 2R regeneration) resulting in conversion of amplitude fluctuations to time position fluctuations.

As for system performance – expressed through BER – regenerative capabilities of any regenerator simultaneously depend upon both the output amplitude and arrival time PD of the '1' and '0' symbols. In the unusual case of 2R regeneration (no clocked decision), a tradeoff has then to be derived, considering both the reduction of amplitude PD and the enlarged arrival time PD induced by the regenerator, to ensure sufficient signal improvement. In **Fig. 3(a)**, we consider a step function for the transfer function of the decision element. In this case, amplitude PD are squeezed to Dirac PD after the decision element, and depending upon addition or not of a clock reference, arrival time PD is reduced (3R) or dramatically enlarged (2R). In **Fig. 3(b)**, the decision element exhibits a moderately nonlinear transfer function. This results in an asymmetric and less-pronounced squeezing of the

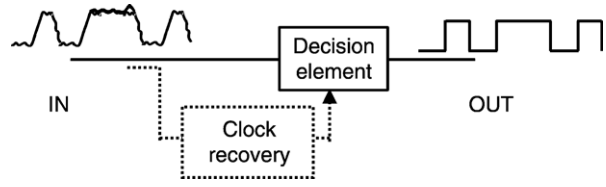


Fig. 2 Generic structure of Signal 2R/3R Regenerator based on Decision Element (2R) and Decision Element and Clock Recovery (3R).

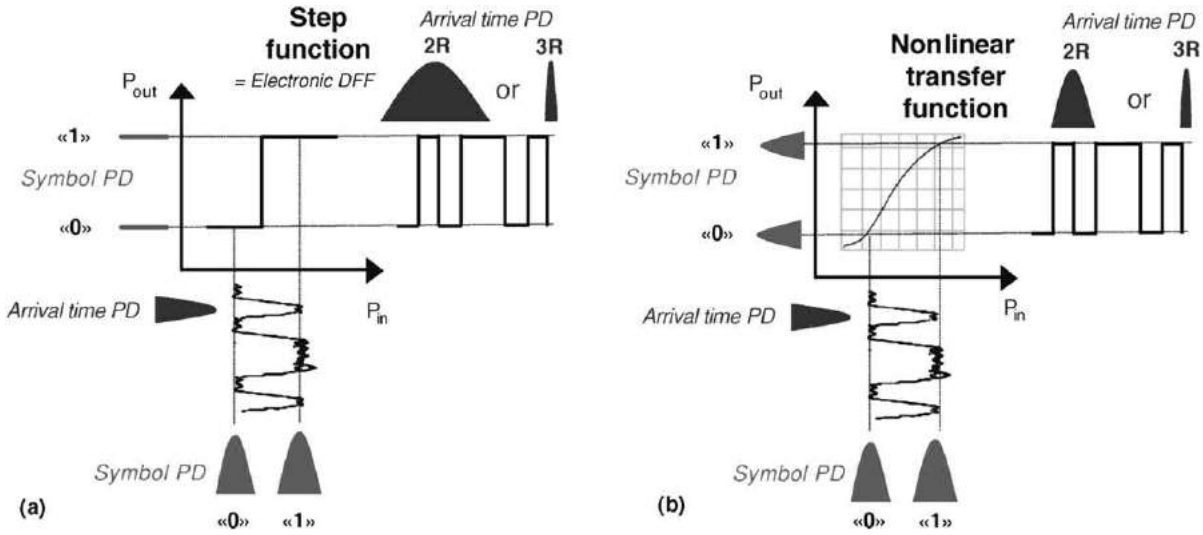


Fig. 3 Signal Regeneration process using Nonlinear Gates. (a) Step transfer function (=Electronic DFF); (b) 'moderately' nonlinear transfer function. As an illustration of the Regenerator operation '1' and '0' symbols amplitude probability density (PD) and arrival time probability density (PD) are shown in light gray and dark gray, respectively.

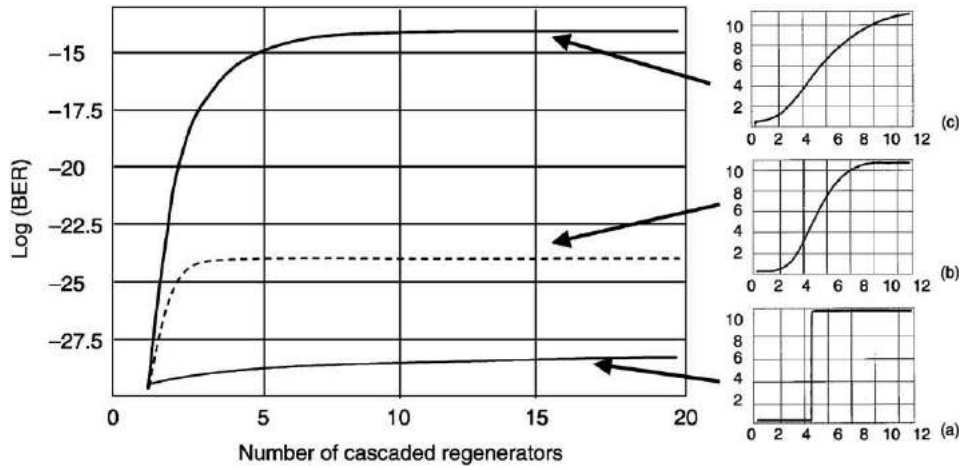


Fig. 4 Evolution of the BER with concatenated regenerators for nonlinear gates with nonlinear transfer function of decreasing depths from case (a)–(c) (step function).

amplitude PD compared to the previous case, but in turn results in a significantly less enlarged arrival time PD when no clock reference is added (2R regeneration). Comparison of these two decision element of different nonlinear transfer function indicates that for 3R regeneration applications, the more nonlinear the transfer function of the decision element the better performance, the ideal case being the step function. In the case of 2R regeneration applications, a tradeoff between the actual reduction of the amplitude PD and enlargement of timing PD is to be derived and clearly depends upon the actual shape of the nonlinear transfer function of the decision element.

Qualification of Signal Regenerator Performance

To further illustrate the impact of the actual shape of the nonlinear transfer function of the decision element in 3R application, the theoretical evolution of BER with number of concatenated regenerators have been plotted for regenerators having different nonlinear responses. Fig. 4 shows the numerically calculated evolution of the BER of a 10 Gbit/s NRZ signal with fixed optical signal-to-noise ratio (OSNR) at the input of the 3R regenerator, as a function of the cascaded regenerator incorporating nonlinear gates with nonlinear transfer function of different depths. From Fig. 4, can be seen different behaviors depending on the nonlinear function shape. As previously stated, the best regeneration performance is obtained with an ideal step function (case a), which is actually the case for O/E regenerator using electronic decision flip-flop (DFF). In that case, BER linearly increase (i.e., more errors)

in the cascade. Conversely, when nonlinearity is reduced (cases (b) and (c)), both BER and noise accumulate, until the concatenation of nonlinear functions reach some steady-state pattern, from which BER linearly increases. Concatenation of nonlinear devices thus magnifies shape differences in their nonlinear response, and hence their regenerative capabilities.

Moreover, as can be seen in Fig. 4, all curves standing for different regeneration efficiencies pass through a common point defined after the first device. This clearly indicates that it is not possible to qualify the regenerative capability of any regenerator when considering the output signal after only one regenerator. Indeed, the BER is the same for either a 3R regenerator or a mere amplifier if only measured after a single element. This originates from the initial overlap between noise distributions associated with marks and spaces, that cannot be suppressed but only minimized by a single decision element through threshold adjustment.

As a general result, the actual characterization of the regenerative performance of any regenerator should *in fine* be conducted considering a cascade of regenerators. In practice this can easily be done with the experimental implementation of the regenerator under study in a recirculating loop. Moreover, such an investigation tool will also enable access to the regenerator performance with respect to the transmission capabilities of the regenerated signal, which should not be overlooked.

Let us now consider the physical implementation of such all-optical Signal Regenerators, along with the key features offered by the different alternatives.

All-Optical 2R/3R Regeneration Using Optical Nonlinear Gates

Prior to describing the different solutions for all-optical in-line Optical Signal Regeneration, it should be mentioned that since the polarization states of the optical data signals cannot be preserved during propagation, it is required that the regenerator exhibits an extremely low polarization sensitivity. This clearly translates to a careful optimization of all the different optical components making up the 2R/3R regenerator. It should be noted that this also applies to the O/E solutions but is of limited impact, since only the photodiode has to be polarization insensitive.

Fig. 5 illustrates the generic principle of operation of an all-optical 3R Regenerator using optical nonlinear gates. Contrary to what occurs in O/E, regenerator where the extracted clock signal drives the electrical decision element, the incoming and distorted optical data signal triggers the nonlinear gate, hence generating a switching window which is applied to a newly generated optical clock signal so as to reproduce the initial data stream on the new optical carrier.

In the case of 2R Optical Signal Regeneration, a continuous wave (CW) signal is substituted for the synchronized optical clock pulses. As previously mentioned, the actual regeneration performance of the 2R/3R devices will mainly depend upon the nonlinearity of the transfer function of the decision element but in 3R applications the quality of the optical clock pulses has also to be considered. In the following, we describe current solutions to explain the two main building blocks of all-optical Signal Regenerators: the decision element (i.e., nonlinear gate) and the clock recovery (CR) elements.

Optical Decision Element

In the physical domain, optical decision elements with ideal step response – as for electrical DFF – do not exist. Different nonlinear optical transfer functions, approaching more or less the ideal case, can be realized in various media such as fiber, SOA, electro-absorption modulators (EAM), and lasers. Generally, as described below, the actual response (hence the regenerative properties of the device) of such optical gates directly depends upon the incoming signal instantaneous power. Under these conditions, it appears essential to add an adaptation stage so as to reduce intensity fluctuations (as caused by propagation or crossing routing/switching node) and provide the decision element with fixed power conditions. In practice, this results in the addition of a control circuitry (either optical or electrical) in the Re-amplification block, whose complexity directly depends on actual system environment (ultra-fast power equalization for packet-switching applications and compensation of slow power fluctuations in transmission applications).

As previously described the decision gate performs Re-shaping (and Re-timing when clock pulses are added) of the incoming distorted optical signal, and represent the regenerator's core element. Ideally, it should also act as a transmitter with characteristics ensuring the actual propagation of the regenerated data stream. In that respect, the chirp possibly induced by the optical decision gate onto the regenerated signal – and the initial quality of the optical clock pulses in 3R applications – should be carefully considered (ideally by means of loop transmission) as to adequately match line transmission requirements.

Different solutions for the actual realization of the optical decision element have been proposed and extensively investigated using, for example, cross gain modulation in semiconductor optical amplifier (SOA) devices but the most promising and flexible

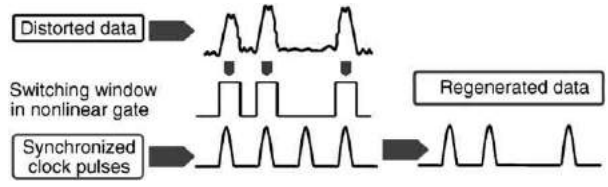


Fig. 5 Principle of operation of an all-Optical Signal 3R Regenerator using nonlinear gates.

devices probably are interferometers, for which, descriptions of the generic principle of operation follows. Consider a CW signal (probe) at λ_2 wavelength injected into an optical interferometer, in which one arm incorporates a nonlinear medium in which an input signal carried by λ_1 wavelength (command) is, in turn, injected. Such a signal, at λ_1 wavelength, induces a phase shift through cross-phase modulation (XPM) in this arm of the interferometer, the amount depending upon power level $P_{\text{in},\lambda 1}$. In turn, such phase modulation (PM) induces amplitude modulation (AM) on the signal at λ_2 wavelength when recombined at the output of the interferometer and translates the information carried by wavelength λ_1 onto λ_2. Under these conditions, such optical gates clearly act as wavelength converters (it should be mentioned that Wavelength Conversion is not necessarily equivalent to Regeneration; i.e., a linear transfer function performs suitable Wavelength Conversion but by no means Signal Regeneration).

Optical interferometers can be classified according to the nature of the nonlinearity exploited to achieve a π phase shift. In the case of fiber-based devices, such as the nonlinear optical loop mirror (NOLM), the phase shift is induced through the Kerr effect in an optical fiber. The key advantage of fiber-based devices such as NOLM lies in the near-instantaneous (fs) response of the Kerr nonlinearity, making them very attractive for ultra-high bit-rate operation ($\geq$160 Gbit/s). Polarization-insensitive NOLMs have been realized, although with the same drawbacks concerning integrability. With recent developments in highly nonlinear (HNL) fibers, however, the required NOLM fiber length could be significantly reduced, hence dramatically reducing environmental instability.

A second type of device is the integrated SOA-based Mach-Zehnder interferometers (MZI). In MZIs, the phase shift is due to the effect of photo-induced carrier depletion in the gain saturation regime of one of the SOAs. The control and probe can be launched in counter- or co-directional ways. In the first case, no optical filter is required at the output of the device for rejecting the signal at λ_1 wavelength but operation of the MZI is limited by its speed. At this point, one should mention that the photo-induced modulation effects in SOAs are intrinsically limited in speed by the gain recovery time, which is a function of the carrier lifetime and the injection current. An approach referred to as differential operation mode (DOM) and illustrated on Fig. 6, which takes advantage of the MZI's interferometric properties, makes it possible to artificially increase the operation speed of such 'slow' devices up to 40 Gbit/s.

As discussed earlier, the nonlinear response is a key parameter for regeneration efficiency. Combining two interferometers is a straightforward means to improve the nonlinearity of the decision element transfer function, and hence regeneration efficiency. This approach was validated at 40 Gbit/s using a cascade of two SOA-MZI, (see Fig. 7 (**left**)). Such a scheme offers the advantage of restoring data polarity and wavelength, hence making the regenerator inherently transparent. Finally, the second conversion stage can be used as an adaptation interface to the transmission link achieved through chirp tuning in this second device.

Such an Optical 3R Regenerator was upgraded to 40 Gbit/s, using DOM in both SOA-MZIs with validation in a 40 Gbit/s loop RZ transmission. The 40 Gbit/s eye diagram monitored at the regenerator output after 1, 10, and 100 circulations are shown in Fig. 7 (**right**) and remain unaltered by distance. With this all-optical regenerator structure, the minimum OSNR tolerated by the regenerator (1 dB sensitivity penalty at 10^{-10} BER) was found to be as low as 25 dB/0.1 nm. Such results clearly illustrate the high performance of this SOA-based regenerator structure for 40 Gbit/s optical data signals.

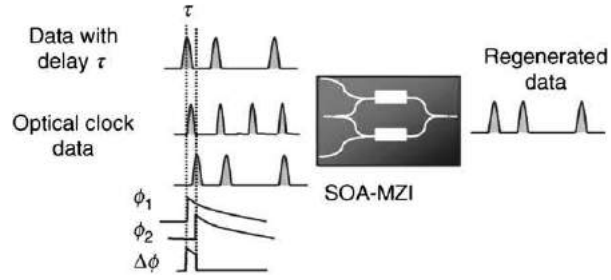


Fig. 6 Schematic and principle of operation of SOA-MZI in differential mode.

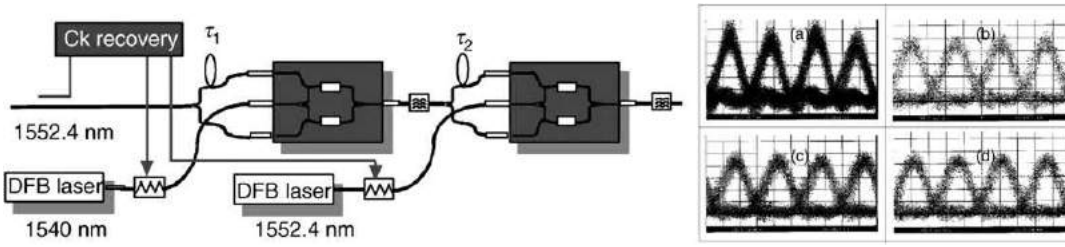


Fig. 7 Optimized structure of a 40 Gbit/s SOA-based 3R regenerator. 40 Gbit/s eye diagram evolution: (a) B-to-B; (b) 1 lap; (c) 10 laps; (d) 100 laps.

Such a complex mode of operation for addressing 40 Git/s bit-rates will probably be discarded when we consider the recent demonstration of standard-mode wavelength conversion at 40 Gbit/s, which uses a newly-designed active-passive SOA-MZI incorporating evanescent-coupling SOAs. The device architecture is flexible in the number of SOAs, thus enabling easier operation optimization and reduced power consumption, leading to simplified architectures and operation for 40 Gbit/s optical 3R regeneration.

Based on the same concept of wavelength conversion for Optical 3R Regeneration, it should be noted many devices have been proposed and experimentally validated as wavelength converters at rates up to 84 Gbit/s, but with cascadability issues still to be demonstrated to assess their actual regenerative properties.

Optical Clock Recovery (CR)

Next to the decision, the CR is a second key function in 3R regenerators. One possible approach for CR uses electronics while another only uses optics. The former goes with OE conversion by means of a photodiode and subsequent EO conversion through a modulator. This conversion becomes more complex and power-hungry as the data-rate increases. It is clear that the maturity of electronics gives a current advantage to this approach. But considering the pros and cons of electronic CR for cost-effective implementation, the all-optical approach seems more promising, since full regenerator integration is potentially possible with reduced power consumption. In this view, we shall focus here on the optical approach and more specifically on the self-pulsating effect in three-sections distributed feedback (DFB) lasers or more recently in distributed Bragg reflector (DBR) lasers. Recent experimental results illustrate the potentials of such devices for high bit rates (up to 160 Gbit/s), broad dynamic range, broad frequency tuning, polarization insensitivity, and relatively short locking times (1 ns). This last feature makes these devices good candidates for operation in asynchronous-packet regimes.

Optical Regeneration by Saturable Absorbers

We next consider saturable absorbers (SA) as nonlinear elements for optical regeneration. Fig. 8 (**left**) shows a typical SA transfer function and illustrates the principle of operation. When illuminated with an optical signal with peak power below some threshold (P_{sat}), the photonic absorption of the SA is high and the device is opaque to the signal (low transmittance). Above P_{sat}, the SA transmittance rapidly increases and asymptotically saturates to transparency (passive loss being overlooked). Such a nonlinear transfer function only applies to 2R optical regeneration.

Different technologies for implementing SAs are available, but the most promising approach uses semiconductors. In this case, SA relies upon the control of carrier dynamics through the material's recombination centers. Parameters such as on–off contrast (ratio of transmittance at high and low incident powers), recovery time (1/e) and saturation energy, are key to device optimization. In the following, we consider a newly-developed ion-irradiated MQW-based device incorporated in a micro-cavity and shown on Fig. 8 (**right**). The device operates as a reflection-mode vertical cavity, providing both a high on/off extinction ratio by canceling reflection at low intensity and a low saturation energy of 2 pJ. It is also intrinsically polarization-insensitive. Heavy ion-irradiation of the SA ensures recovery times (at 1/e) shorter than 5 ps (hence compatible with bit-rate above 40 Gbit/s), while maintaining a dynamic contrast in excess of 2.5 dB at 40 GHz repetition rate.

The regenerative properties of SA make it possible to reduce cumulated amplified spontaneous emission (ASE) in the '0' bits, resulting in a higher contrast between mark and space, hence increasing system performance. Yet SAs do not suppress intensity noise in the marks, which makes the regenerator incomplete. A solution for this noise suppression is optical filtering with nonlinear (soliton) pulses. The principle is as follows. In absence of chirp, the soliton temporal width scales in the same way as the reciprocal of its spectral width (Fourier-transform limit) times its intensity (fundamental soliton relation). Thus, an increase in pulse intensity corresponds to both time narrowing and spectral broadening. Conversely, a decrease in pulse intensity corresponds to time broadening and spectral narrowing. Thus, the filter causes higher loss when intensity increases, and lower loss when intensity decreases. The filter thus acts as an automatic power control (APC) in feed-forward mode, which causes power

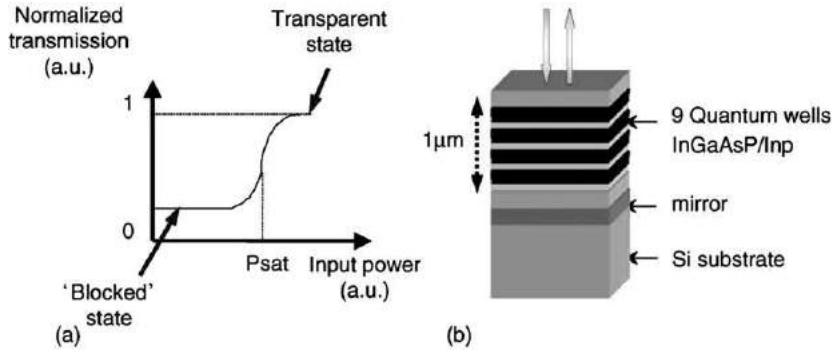


Fig. 8 (a) Saturable Absorber (SA) ideal transfer function. (b) Structure of Multi-Quantum Well SA.

stabilization. The resulting 2R regenerator (composed by the SA and the optical filter) is fully passive, which is of high interest for submarine systems where the power consumption must be minimal, but it does not include any control in the time domain (no Re-timing).

System demonstrations of such passive SA-based Optical Regeneration have been reported with a 20 Gbit/s single-channel loop experiment. Implementation of the SA-based 2R Regenerator with 160 km-loop periodicity made it possible to double the error-free distance (Q = 15.6 dB or 10^{-9} BER) of a 20 Gbit/s RZ signal. So as to extend the capability of passive 2R regeneration to 40 Gbit/s systems, an improved configuration was derived from numerical optimization and experimentally demonstrated in a 40 Gbit/s WDM-like, dispersion-managed loop transmission, showing more than a fourfold increase in the WDM transmission distance at 10^{-4} BER (1650 km without the SA-based regenerator and 7600 km when implementing the 2R regenerator with 240 km periodicity).

Such a result illustrates the potential high interest of such passive optical 2R regeneration in long-haul transmission (typically in noise-limited systems) since the implementation of SA increases the system's robustness to OSNR degradation without any extra power consumption. Reducing both saturation energy and insertion loss along with increasing dynamic contrast represent key future device improvements. Regeneration of WDM signals from the same device, such as one SA chip with multiple fibers implemented between Mux/DMux stages, should also be thoroughly investigated. In this respect, SA wavelength selectivity in quantum dots could possibly be advantageously exploited.

Synchronous Modulation Technique

All-optical 3R regeneration can be also achieved through in-line synchronous modulation (SM) associated with narrowband filtering (NF). Fig. 9 shows the basic layout of such an Optical 3R Regenerator. It is composed of an optical filter followed by an Intensity and Phase Modulator (IM/PM) driven by a recovered clock. Periodic insertion of SM-based modulators along the transmission link provides efficient jitter reduction and asymptotically controls ASE noise level, resulting in virtually unlimited transmission distances. Re-shaping and Re-timing provided by IM/PM intrinsically requires nonlinear (soliton) propagation in the trunk fiber following the SM block. Therefore, one can refer to the approach as distributed optical regeneration. This contrasts with lumped regeneration, where 3R is completed within the regenerator (see above with Optical Regeneration using nonlinear gates), and is independent of line transmission characteristics.

However, when using a new approach referred to 'black box' optical regeneration (BBOR), it is possible to make the SM regeneration function and transmission work independently in such a way that any type of RZ signals (soliton or non-soliton) can be transmitted through the system. The BBOR technique includes an adaptation stage for incoming RZ pulses in the SM-based regenerator, which ensures high regeneration efficiency regardless of RZ signal format (linear RZ, DM-soliton, C-RZ, etc.). This is achieved using a local and periodic soliton conversion of RZ pulses by means of launching an adequate power into some length of fiber with anomalous dispersion. The actual experimental demonstration of the BBOR approach and its superiority over the 'classical' SM-based scheme for DM transmission was experimentally investigated in 40 Gbit/s DM loop transmission. Under these conditions, one can then independently exploit dispersion management (DM) techniques for increasing spectral efficiency in long-haul transmission, while ensuring high transmission quality through BBOR.

One of the key properties of the SM-based all-optical Regeneration technique relies on its WDM compatibility. The first (Fig. 10, **left**) and straightforward solution to apply Signal Regeneration to WDM channels amounts to allocating a regenerator to each WDM channel. The second consists in sharing a single modulator, thus processing the WDM channels at once in serial fashion. This approach requires WDM synchronicity, meaning that all bits be synchronous with the modulation, that can be achieved either by use of appropriate time-delay lines located within a DMux/Mux apparatus (Fig. 10, **upper right**), or by making

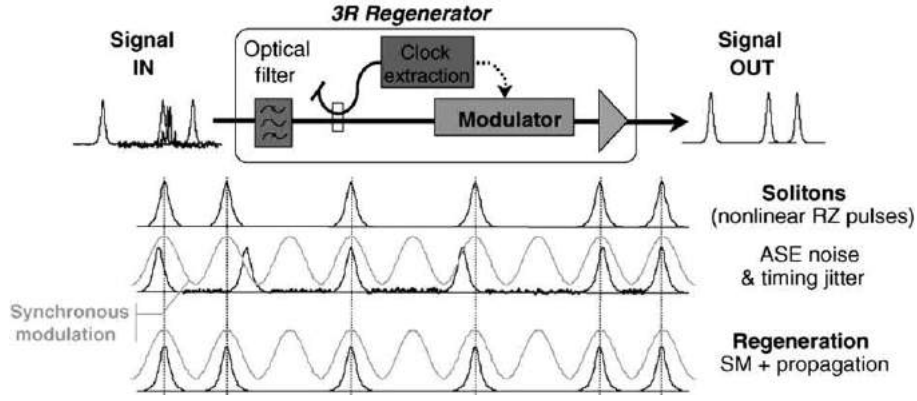


Fig. 9 Basic layout of the all-optical Regenerator by Synchronous Modulation and Narrowband Filtering and illustration of the principle of operation.

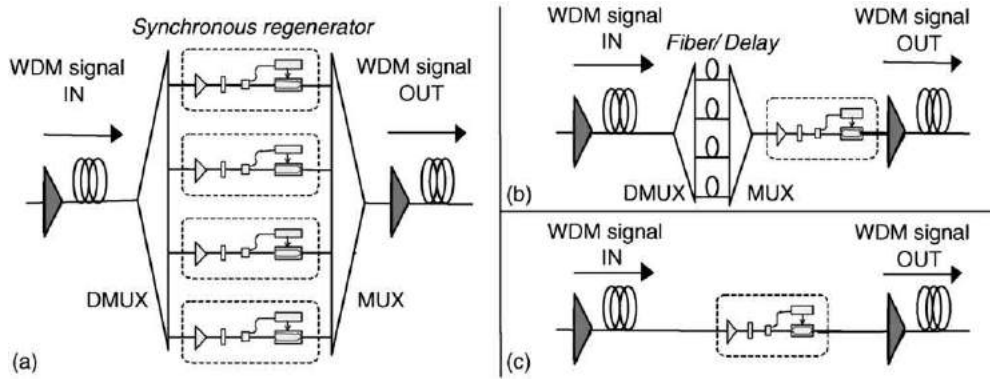


Fig. 10 Basic implementation schemes for WDM all-Optical Regeneration. (a) parallel asynchronous; (b) serial re-synchronized; (c) serial self-synchronized.

the WDM channels inherently time-coincident at specific regenerator locations (**Fig. 10**, **bottom right**). Clearly, the serial regeneration scheme is far simpler and cost-effective than the parallel version; however, optimized re-synchronization schemes still remain to be developed for realistic applications. Experimental demonstration of this concept was assessed by means of a 4×40 Gbit/s dispersion-managed transmission over 10 000 km (BER $< 5.10^{-8}$) in which a single modulator was used for the simultaneous regeneration of the 4 WDM channels.

Considering next all-optical regeneration schemes with ultra-high speed potential, a compact and loss-free 40 Gbit/s Synchronous Modulator, based on optically-controlled SOA-MZI, was proposed and loop-demonstrated at 40 Gbit/s with an error-free transmission distance in excess of 10 000 km. Moreover, potential ultra-high operation of this improved BBOR scheme was recently experimentally demonstrated by means of a 80 GHz clock conversion with appropriate characteristics through the SOA-MZI. One should finally mention all fiber-based devices such as NOLM and NALM for addressing ultra-high speed SM-based Optical Regeneration, although no successful experimental demonstrations have been reported so far in this field.

Conclusion

In summary, optical solutions for Signal Regeneration present many key advantages. These are the only advantages to date to possibly ensure WDM compatibility of the regeneration function (mostly 2R related). Such optical devices clearly exhibits the best 2R regeneration performance (wrt to O/E solutions) as a result of the moderately nonlinear transfer function (which in turn can be considered as a drawback in 3R applications), but the optimum configuration is still to be clearly derived and identified depending upon the system application. Optics also allow to foresee and possibly target ultrafast applications above 40G, for signal regeneration if needed. Among the current drawbacks, one should mention the relative lack of wavelength/formats flexibility of these solutions (compared to O/E solutions). It is complex or difficult to restore the input wavelength or address any C-band wavelength at the output of the device or to successfully regenerate modulation formats other than RZ. In that respect, investigations should be conducted to derive new optical solutions capable of processing more advanced modulation formats at 40G. Finally, the fact that the nonlinear transfer function of the optical is in general triggered by the input signal instantaneous power also turns out to be a drawback since it requires control circuitry. The issue of cost (footprint, power consumption, etc.) of these solutions, compared to O/E ones, is still open. In this respect, purely optical solutions incorporating all-optical clock recovery, the performance of which is still to be technically assessed, are of high interest for reducing costs. Complete integration of an all-optical 2R/3R regenerator or such parallel regenerators onto a single semiconductor chip should also contribute to make all-optical solutions cost-attractive even though acceptable performance of such fully integrated devices is still to be demonstrated.

From today's status concerning the two alternative approaches for in-line regeneration (O/E or all-optical), it is safe to say that the choice between either solution will be primarily dictated by both engineering and economical considerations. It will result from a tradeoff between overall system performance, system complexity and reliability, availability, time-to-market, and rapid returns from the technology investment.

Further Reading

Bigo, S., Leclerc, O., Desurvire, E., 1997. All-optical fiber signal processing for solitons communications. IEEE Journal of Selected Topics on Quantum Electronics 3 (5), 1208–1223.

Dagens B, Labrousse A, Fabre S, *et al.* (2002) New modular SOA-based active–passive integrated Mach–Zehnder interferometer and first standard mode 40 Gb/s all-optical wavelength conversion on the C-band. paper PD 3.1, Proceedings of ECOCOL 02, Copenhague.

Durhuus, T., Mikkelsen, B., Joergensen, C., Danielsen, S.L., Stubkjaer, K.E., 1996. All-optical wavelength conversion by semiconductor optical amplifiers. Journal of Lightwave Technology 14 (6), 942–954.

Lavigne B, Guerber P, Brindel P, Balmefrezol E and Dagens B (2001) Cascade of 100 optical 3R regenerators at 40 Gbit/s based on all-active Mach–Zehnder interferometers. paper We.F.2.6, Proceedings of ECOC 2001, Amsterdam.

Leclerc, O., Lavigne, B., Chiaroni, D., Desurvire, E., 2002. All-optical regeneration: Principles and WDM implementation. Kaminow, I.P., Li, T. (Eds.), 40 Gbit/s transmission and cascaded all-optical wavelength conversion over 1,000,000 km. 732–784. chap. 15.

Leuthold, J., Raybon, G., Su, Y., 2002. 40 Gbit/s transmission and cascaded all-optical wavelength conversion over 1,000,000 km. Electronic Letters 38 (16), 890–892.

Öhlén, P., Berglind, E., 1997. Noise accumulation and BER estimates in concatenated nonlinear optoelectronic repeaters. IEEE Photon Technol. Letters 9 (7), 1011–1013.

Otani T, Suzuki M and Yamamoto S (2001) 40 Gbit/s optical 3R regenerator for all-optical networks. We.F.2.1, Proceedings of ECOC 2001, Amsterdam.

Oudar, J.-L., Aubin, G., Mangeney, J., 2004. Ultra-fast quantum-well saturable absorber devices and their application to all-optical regeneration of telecommunication optical signals. Annales des Télécommunications 58 (11–12),

Sartorius, B., Mohrle, M., Reichenbacher, S., 1997. Dispersive self-Q-switching in self-pulsating DFB lasers. IEEE Journal of Quantum Electronics 33 (2), 211–218.

Ueno Y, Nakamura S, Sasaki J, *et al.* (2001) Ultrahigh-speed all-optical data regeneration and wavelength conversion for OTDM systems. Th.F.2.1, Proceedings of ECOC 2001, Amsterdam.

Heterodyning

T-C Poon, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Heterodyning, also known as frequency mixing, is a frequency translation process. Heterodyning has its root in radio engineering. The principle of heterodyning was discovered in the late 1910s by radio engineers experimenting with radio vacuum tubes. Russian cellist and electronic engineer Leon Theremin invented the so-called Thereminvox, which was one of the earliest electronic musical instruments that generates an audio signal by combining two different radio signals. American electrical engineer Edwin Armstrong invented the so-called super-heterodyne receiver. The receiver shifts the spectrum of the modulated signal, that is, the frequency contents of the modulated signal, so as to demodulate the audio information from it. Commercial amplitude modulation (AM) broadcast receivers nowadays are super-heterodyne type. After illustrating the basic principle of heterodyning by some simple mathematics, we will discuss some examples on how the principle of heterodyning is used in radio and in optics.

For a simple case, when signals of two different frequencies are heterodyned or mixed, the resulting signal produces two new frequencies, the sum and difference of the two original frequencies. **Fig. 1** illustrates how signals of two frequencies are mixed or heterodyned to produce two new frequencies, $\omega_1 + \omega_2$ and $\omega_1 - \omega_2$, by simply multiplying the two signals $\cos(\omega_1 t + \theta_1)$ and $\cos(\omega_2 t)$, where ω_1 and ω_2 are radian frequencies of the two signals and θ_1 is the phase angle between the two signals. Note that when the frequencies of the two signals to be heterodyned are the same, i.e., $\omega_1 = \omega_2$, the phase information of $\cos(\omega_1 t + \theta_1)$ can be extracted to get $\cos\theta_1$ if we use an electronic lowpass filter (LPF) to filter out the term $\cos(2\omega_1 t + \theta_1)$. This is shown in **Fig. 2**. Heterodyning is often referred to as homodyning for the mixing of two signals of the same frequency.

For a general case of heterodyning, we can have a signal represented by $s(t)$ with its spectrum $s(\omega)$, which is given by the Fourier transform of $s(t)$, and when it is multiplied by $\cos(\omega_2 t)$, the resulting spectrum is frequency-shifted to new locations in the frequency domain as:

$$\mathcal{F}\{s(t)\cos(\omega_2 t)\} = \frac{1}{2}S(\omega - \omega_2) + \frac{1}{2}S(\omega + \omega_2) \tag{1}$$

where $\mathcal{F}\{s(t)\} = S(\omega)$ and $\mathcal{F}\{\cdot\}$ denotes the Fourier transform of the quantity being bracketed. The spectrum of $s(t)\cos(\omega_2 t)$, along with the spectrum of $s(t)$, is illustrated in **Fig. 3**. It is clear that multiplying $s(t)$ with $\cos(\omega_2 t)$ is a process of heterodyning, as we have translated the spectrum of the signal $s(t)$. This process is known as modulation in communication systems.

In order to appreciate the process of heterodyning let us now consider, for example, heterodyning in radio. In particular, we consider AM. While the frequency content of an audio signal (from 0 to around 3.5 kHz) is suitable to be transmitted over a pair of wires or coaxial cables, it is, however, difficult to be transmitted in air. By impressing the audio information onto a higher

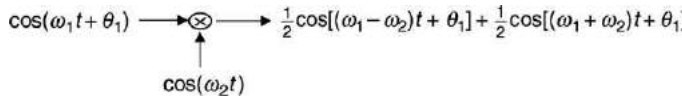


Fig. 1 Heterodyning of two sinusoidal signals.

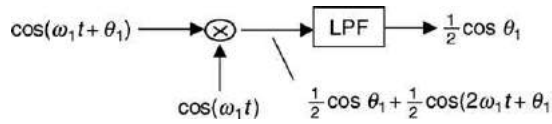


Fig. 2 Heterodyning becomes homodyning when the two frequencies to be mixed are the same: homodyning allows the extraction of the phase information of the signal.

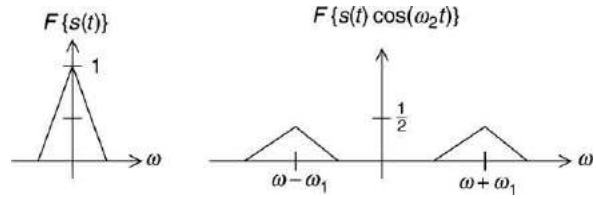


Fig. 3 Spectrum of $s(t)$ and $s(t)\cos(\omega_2 t)$. It is clear that spectrum of $s(t)$ has been translated to new locations upon multiplying a sinusoidal signal.

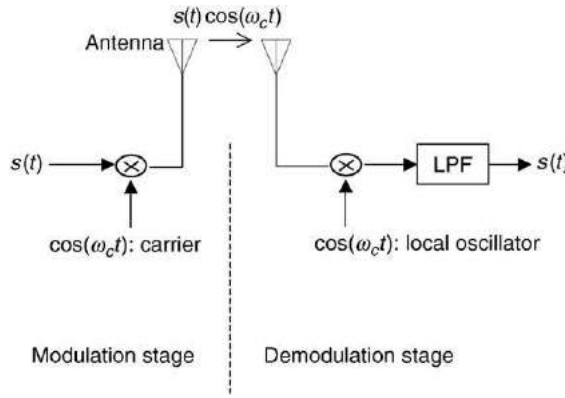


Fig. 4 Radio link: heterodyning in the demodulation stage is often known as heterodyne detection.

frequency, say 550 kHz, as one of the broadcast bands in AM radio, i.e., by modulating the audio signal, the resulting modulated signal can now be transmitted using antennas of reasonable dimensions. In order to recover (demodulate) the audio signal from the modulated signal, the modulated signal is first received by an antenna in a radio receiver, multiplied by the so-called local oscillator with the same frequency as the signal used to modulate the audio signal, then followed by a lowpass filter to eventually obtain the audio information back. The situation is illustrated in **Fig. 4**. We shall now describe the process mathematically with reference to **Fig. 4**.

We denote $s(t)$ as the audio signal. After multiplying by $\cos(\omega_c t)$, which is usually called a carrier in radio in the modulation stage, where ω_c is the radian frequency of the carrier. The modulated signal $s(t)\cos(\omega_c t)$. is transmitted via an antenna. When the modulated signal is received in the demodulation stage, the modulated signal is then multiplied by the local oscillator of signal waveform $\cos(\omega_c t)$. The output of the multiplier is given by

$$s(t)\cos(\omega_c t)\cos(\omega_c t) = s(t)\frac{1}{2}[1+\cos(2\omega_c t)] \tag{2}$$

Note that the two signals in the demodulation stage, $s(t)\cos(\omega_c t)$ and $\cos(\omega_c t)$, are heterodyned to produce two new frequencies, the sum $\omega_c + \omega_c = 2\omega_c$ to give the term $\cos(2\omega_c t)$ and the difference $\omega_c - \omega_c = 0$ to give the term $\cos(0) = 1$. Since the two frequencies to be multiplied in the demodulation stage are the same, this is homodyning as explained above. Now by performing lowpass filtering (LPF), we can recover our original audio information $s(t)$. Note that if the frequency of the local oscillator in the demodulation stage within the receiver is higher than the frequency of the carrier used in the modulation stage, heterodyning in the receiver is referred to as super-heterodyning, which most receivers nowadays use for amplitude modulation. In general, we have two heterodyning processes in the radio system just described, one in the modulation stage and the other in the demodulation stage. However, it is unusual to speak of heterodyning in the modulation stage, and so we just refer to the process in the demodulation stage as 'heterodyne detection.'

In summary, heterodyne detection can extract the information from a modulated signal and can also extract the phase information of a signal if homodyning is used. We shall see, in the next section, how optical heterodyne detection is employed.

Optical information is usually carried by coherent light such as a laser. Let $\psi_p(x,y)$ be a complex amplitude of the light field, which may physically represent a component of the electric field. We further assume that the light field is oscillating at temporal frequency ω_0. Therefore, we can write the light field as $\psi_p \exp(i\omega_0 t)$. By taking the real part of $\psi_p \exp(i\omega_0 t)$, i.e., $\mathrm{Re}\left[\psi_p \exp(i\omega_0 t)\right]$, we recover the usual physical real quantity. For a simple example, if we let $\psi_p = A\exp(-ik_0 z)$, where A is some constant and k_0 is the wavenumber of the light, $\mathrm{Re}\left[\psi_p \exp(i\omega_0 t)\right] = \mathrm{Re}[A\exp(-ik_0 z)\exp(i\omega_0 t)] = A\cos(\omega_0 t - k_0 z)$, which is a plane wave propagating along the positive z direction in free space. To detect light energy, we use a photodetector (PD), as shown in **Fig. 5**. Assuming a plane wave for simplicity, as $\psi_p = A$, we have also taken $z=0$ at the surface of the photodetector. As the photodetector responds to intensity, i.e., $|\psi_p|^2$, which gives the current, i, as output by spatially integrating the intensity:

$$i \propto \int_S |\psi_p \exp(i\omega_0 t)|^2 \mathrm{d}x\mathrm{d}y = A^2 S \tag{3}$$

where S is the surface area of the photodetector. We can see that the photodetector current is proportional to the intensity, A^2, of the incident light. Hence the output current varies according to the intensity of the optical signal intensity. This mode of photodetection is called direct detection or incoherent detection in optics.

Let us now consider the heterodyning of two plane waves on the surface of the photodetector. We assume an information-carrying plane wave, also called the signal plane wave, $A_s \exp\{i[(\omega_0 + \omega_s)t + s(t)] \times \exp(-ik_0 x\sin\phi)\}$, and a reference plane wave, $A_r \exp(i\omega_0 t)$, or called a local oscillator in radio. The situation is shown in **Fig. 6**.

Note that the frequency of the signal plane wave is ω_s higher than that of the reference signal, and it is inclined at an angle ϕ with respect to the reference plane wave, which is normal incident to the photodetector. Also, the information content $s(t)$ is in the phase of the signal wave. In the situation shown in **Fig. 6**, we see that the two plane waves are interfered on the surface of the

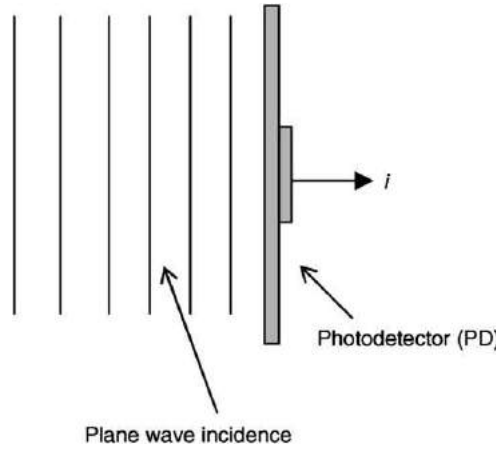


Fig. 5 Optical direction detection or optical incoherent detection.

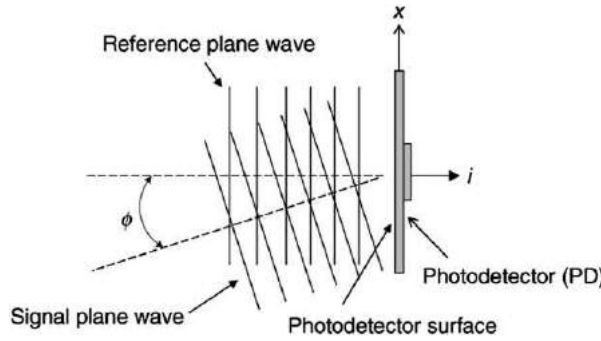


Fig. 6 Optical heterodyne detection.

photodetector, giving the total light field ψ_v given by

$$\psi_t = A_r \exp(i\omega_0 t) + A_s \exp\{i[(\omega_0 + \omega_s)t + s(t)]\} \times \exp(-\mathrm{i}k_0 x \sin\phi) \tag{4}$$

Again, the photodetector responds to intensity, giving the current

$$i \propto \int_S |\psi_t|^2 \mathrm{d}x\mathrm{d}y = \int_{-a}^{a}\int_{-a}^{a} [A_r^2 + A_s^2 + 2A_r A_s \cos(\omega_s t + s(t) - k_0 x \sin\phi)]\mathrm{d}x\mathrm{d}y \tag{5}$$

where we have assumed that the photodetector has a $2a \times 2a$ square area. The current can be evaluated to be:

$$i(t) \propto 2a\left(A_r^2 + A_s^2\right) + 4A_r A_s \frac{\sin(k_0 a \sin\phi)}{k_0 \sin\phi}\cos(\omega_s t + s(t)) \tag{6}$$

The current output has two parts: the DC current and the AC current. The AC current at frequency ω_s is commonly known as the heterodyne current. Note that the information content $s(t)$ originally embedded in the phase of the signal plane wave has now been preserved and transferred to the phase of the heterodyne current. The above process is called optical heterodyning. In optical communications, it is often referred to as optical coherent detection. In contrast, if the reference plane wave has not been used for the detection, we have the incoherent detection. The information content carried by the signal plane wave would be lost, as it is evident that for $A_r = 0$, the above equation gives only a DC current at a value proportional to the intensity of the plane wave, A_s^2.

Let us now consider some of the practical issues encountered in coherent detection. Again, the AC part of the current given by the above equation is the heterodyne current $i_{\mathrm{het}}(t)$, given by

$$i_{\mathrm{het}}(t) \propto A_r A_s \frac{\sin(k_0 a \sin\phi)}{k_0 \sin\phi}\cos(\omega_s t + s(t)) \tag{7}$$

We see that since the two plane waves propagate in slightly different directions, the heterodyne current output is degraded by a factor of

$$\frac{\sin(k_0 a \sin\phi)}{k_0 \sin\phi} = a \sin c(k_0 a \sin\phi)$$

where $\mathrm{sin}c(x) = \sin(x)/x$. For small angles, i.e., $\sin\phi \approx \phi$, the current amplitude falls off as $\mathrm{sin}c(k_0 a\phi)$. Hence, the heterodyne current is at a maximum when the angular separation between the signal plane wave and the reference plane wave is zero, i.e., the

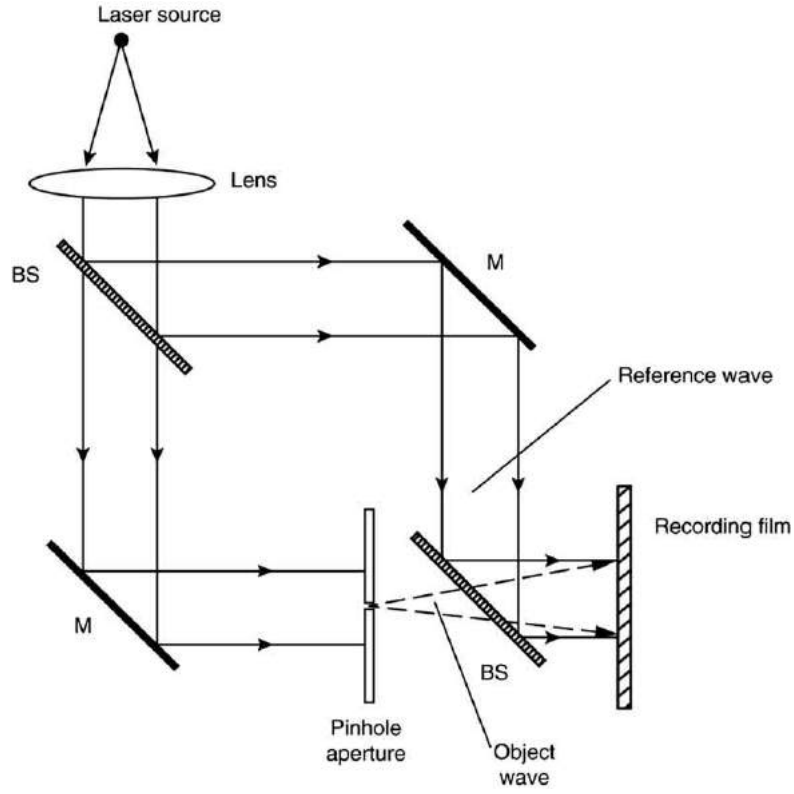


Fig. 7 Holographic recording of a point source object.

two plane waves are propagating exactly in the same direction. The current will go to zero when $k_0 a\phi = \pi$, or $\phi = \lambda_0/2a$, where λ_0 is the wavelength of the laser light. To see how critical it is for the angle ϕ to be aligned in order to have any heterodyne output, we assume the size of the photodetector $2a = 1$cm and the laser used is red, i.e., $\lambda_0 \approx 0.6\mu m$; ϕ is calculated to be about 2.3×10^{-3} degrees. Hence to be able to work with coherent detection, we need to have precise optomechanical mounts for angular rotation.

Finally we describe a famous application of heterodyning in optics – holography. As we have seen, the mixing of two light fields with different temporal frequencies will produce heterodyne current at the output of the photodetector. But we can record the two spatial light fields of the same temporal frequency with photographic films instead of using electrical devices. Photographic films respond to light intensity as well. We discuss holographic recording of a point source object as a simple example. **Fig. 7** shows a collimated laser split into two plane waves and recombined by the use of two mirrors (M) and two beamsplitters (BS). One plane wave is used to illuminate the pinhole aperture (our point source object), and the other illuminates the recording film directly. The plane wave that is scattered by the point source object, which is located z_0 away from the film, generates a diverging spherical wave. This diverging wave is known as an object wave in holography. The object wave arising from the point source object on the film is given by, according to Fresnel diffraction:

$$\psi_0 = A_0 \exp(-ik_0 z_0)\frac{ik_0}{2\pi z_0} \times \exp\left[-ik_0\left(x^2+y\right)^2/2z_0\right]\exp(i\omega_0 t) \tag{8}$$

This object wave is a spherical wave, where A_0 is the amplitude of the spherical wave.

The plane wave that directly illuminates the photographic plate is known as a reference wave, ψ_r. For the reference plane wave, we assume that the plane wave has the same phase with the point source object at a distance z_0 away from the film. Its field distribution on the film is, therefore, $\psi_r = A_r \exp(-ik_0 z_0)\exp(i\omega_0 t)$, where A_r is the amplitude of the plane wave. The film now records the interference of the reference wave and the object wave, i.e., what is recorded on the film as a 2D pattern is given by $t(x,y) \propto |\psi_r + \psi_0|^2$. The resulting recorded 2D pattern, $t(x,y)$, is called the hologram of the object. This kind of recording is known as holographic recording, distinct from a photographic recording in which the reference wave does not exist and hence only the object wave is recorded. Now:

$$\begin{aligned} t(x,y) \propto |\psi_r + \psi_0|^2 &= \left|A_r \exp(-ik_0 z_0)\exp(i\omega_0 t) + A_0 \exp(-ik_0 z_0)\frac{ik_0}{2\pi z_0}\exp[-ik_0(x^2+y)^2/2z_0]\exp(i\omega_0 t)\right|^2 \\ &= A + B\sin\left\{\frac{k_0}{2z_0}\left[(x^2+y)^2/2z_0\right]\right\} \end{aligned} \tag{9}$$

where A and B are some inessential constants. Note that the result of recording two spatial light fields of the same temporal

frequency preserves the phase information (noticeably the depth parameter z_0) of the object wave. This is considered optical homodyning as it is clear from the result shown in Fig. 2.

The intensity distribution being recorded on the film, upon being developed, will have transmittance given by the above equation. This expression is called the Fresnel zone plate, which is the hologram of a point source object and we shall call it the point-object hologram. Fig. 8 shows the hologram for a particular value of z_0 and k_0. The importance of this phase-preserving recording is that when we illuminate the hologram with a plane wave, called the reconstruction wave, a point object is reconstructed at the same location of the original point source object if we look towards the hologram as shown in Fig. 9, that is the point source is reconstructed at a distance z_0 from the hologram as if it were at the same distance away from the recording film. For an arbitrary 3D object, we can imagine the object as a collection of points, and therefore, we can see that we have a collection of Fresnel zone plates on the hologram. Upon reconstruction, such as the illumination of the hologram by a plane wave, the observer would see a 3D virtual object behind a hologram.

As a final example, we discuss a state-of-the-art holographic recording technique called optical scanning holography, which employs the use of optical heterodyning and electronic homodyning to achieve real-time holographic recording without the use of films. We will take the holographic recording of a point object as an example. Suppose we superimpose a plane wave and a spherical wave of different temporal frequencies and use the resulting intensity pattern as an optical scanning beam to raster scan a point object located z_0 away from the point source that generates the spherical wave. The situation is shown in Fig. 10. Point C is the point source that generates the spherical wave (shown with dashed lines) on the pinhole object, our point object. This point source, for example, can be generated by a focusing lens. The solid parallel rays represent the plane wave. The interference of the two waves generates a Fresnel zone plate type pattern on the pinhole object:

$$I_s(x,y;z_0,t) = \left|A_r\ \exp(-\mathrm{i}k_0 z_0)\exp(i\omega_0 t) + A_0\exp(-\mathrm{i}k_0 z_0)\,\frac{\mathrm{i}k_0}{2\pi z_0}\exp\left[-\mathrm{i}k_0\left(x^2+y\right)^2/2z_0\right]\exp[i(\omega_0+\Omega)t]\right|^2$$

$$= 1 + \left(\frac{1}{\lambda_0 z_0}\right)^2 + \frac{1}{\lambda_0 z_0}\sin\left[\frac{\pi}{\lambda_0 z_0}\left(x^2+y^2\right) - \Omega t\right] \tag{10}$$

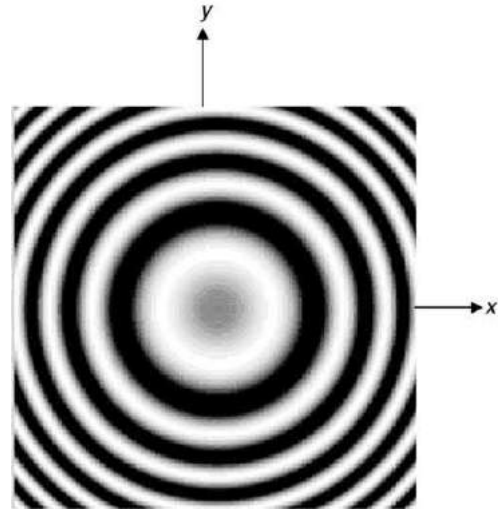


Fig. 8 Point-object hologram.

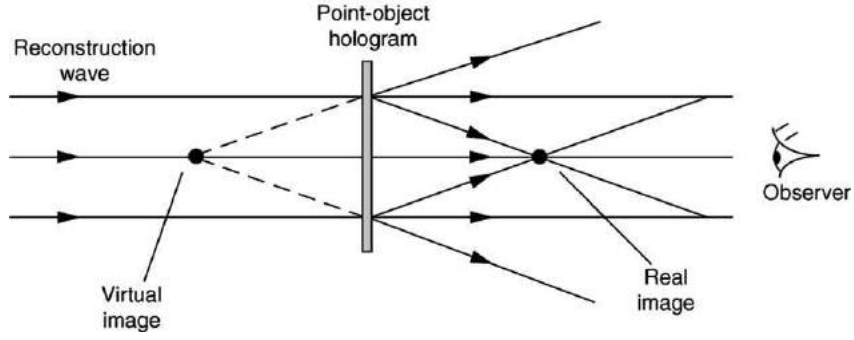


Fig. 9 Reconstruction of a point-object hologram.

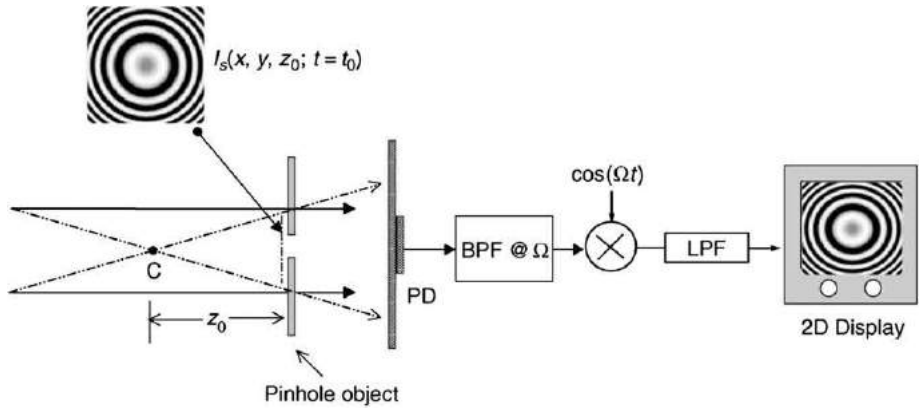


Fig. 10 Optical scanning holography (use of optical heterodyning and electronic homodyning to record holographic information upon scanning the object in two dimensions).

assuming $A_r = A_0 = 1$ for simplicity. Note that the plane wave is at temporal frequency ω_0, and the spherical wave is at temporal frequency $\omega_0 + \Omega$. The expression $I_s(x, y; z_0, t)$ is a temporally modulated Fresnel zone plate and is known as the time-dependent Fresnel zone plate (TDFZP). If we freeze time, say at $t = t_0$, we have the Fresnel zone plate pattern on the pinhole object as shown in **Fig. 10**. However, if we let the time run in the above expression, physically we will have running zones that would be moving away from the center of the pattern. These running zones are the result of optical heterodyning of the plane wave and the spherical wave of different temporal frequencies. It is this TDFZP that is used to raster scan a 3D object to obtain holographic information of the scanned object and such technique is called optical scanning holography.

Upon scanning by the TDFZP, the photodetector, captures all the transmitted light and delivers a current, which consists of a heterodyne current at frequency Ω. After electronic bandpass filtering (BPF) at Ω, the heterodyne current is homodyned electronically by $\cos(\Omega t)$ and lowpass filtered (LPF) to extract the phase information of the current. When this final scanned and processed current is display in a 2D display, as shown in **Fig. 10**, we have the Fresnel zone plate, which is the hologram of the pinhole object, on the display. We can photograph this 2D display to obtain a transparency and have it illuminated by a plane wave to have the reconstruction as shown in **Fig. 9**, or, since the hologram is now in electronic form, we can store it in a PC and reconstruct it digitally. When holographic reconstruction is performed digitally, we have so-called digital holography.

See also: All-Optical Multiplexing/Demultiplexing

Further Reading

Banerjee, P.P., Poon, T.C., 1991. Principles of Applied Optics. MA: Irwin.
Palais, J.C., 1988. Fiber Optic Communications. New Jersey: Prentice-Hall.
Poon, T.C., Banerjee, P.P., 2001. Contemporary Optical Image Processing with MATLAB. Oxford: Elsevier.
Poon, T.C., Qi, Y., 2003. Novel real-time joint-transform correlation using acousto-optic heterodyning. Applied Optics 42, 4663–4669.
Poon, T.C., Wu, M.H., Shinoda, K., Suzuki, Y., 1996. Optical scanning holography. Proc. IEEE 84, 753–764.
Pratt, W.K., 1969. Laser Communication Systems. New York: Wiley.
Stremler, F.G., 1990. Introduction to Communication Systems. MA: Addison-Wesley.
VanderLugt, A., 1992. Optical Signal Processing. New York: Wiley.

Terahertz Lasers

Benjamin S Williams, University of California, Los Angeles, CA, United States
Qing Hu, MIT, Cambridge, MA, United States

Optically Pumped Molecular Gas Lasers

The first commercially available terahertz lasers (more traditionally known as far-infrared or sub-millimeter wave lasers) were gas lasers, where stimulated emission takes place between rotational levels of excited vibrational states of low-pressure molecular gasses. The first such demonstration was made in 1963, where lasing in water vapor excited with a pulsed electric discharge was observed by Crocker *et al.* (1964). Gas lasers continue to be an important THz laser source, although now the most common arrangement is to use optical pumping by a CO_2 gas laser (Chang and Bridges, 1970), which significantly increases the pumping selectivity and efficiency (**Fig. 1**). A large variety of discrete lines can be obtained between 0.1 and 8 THz (Inguscio *et al.*, 1986), although most strongly pumped lines are below 3 THz. Continuous-wave power levels of milliwatts are not uncommon, with powers of hundreds of milliwatts possible for the strongest lines (e.g. the methanol lines at 118.8 and 184.3 μm). There have been many review articles published on far-IR gas lasers, including (Jacobsson (1989), Inguscio *et al.* (1986), Dodel (1999)), and they are commercially available from several vendors.

The selection of laser frequencies is limited by available gasses, some of which are "inconvenient" to handle from a safety and environmental point of view. Only discrete tuning of the lasers is available because of the limited transition lines, however harmonic generators can be used to generate tunable sidebands (Farhoomand *et al.*, 1985). Although these laser systems present limitations in terms of size and weight, this has not prevented them from being flown into space. For example, a far-IR gas laser was used as a 2.5 THz local oscillator source in the Earth Observing System (EOS) Aura satellite, in the Microwave Limb Sounder heterodyne instrument. This laser delivered ~30 mW of cw power while consuming 120 W of electrical power (Mueller *et al.*, 2007).

It is notable that THz molecular gas lasers can operate at room temperature, even though the THz photon energy ($h(\nu) \sim 1$–30 meV) is smaller than the thermal energy k_BT. This is in contrast to the THz semiconductor lasers to be discussed below, all of which require varying degrees of cryogenic cooling. A naïve assumption is sometimes made that a population

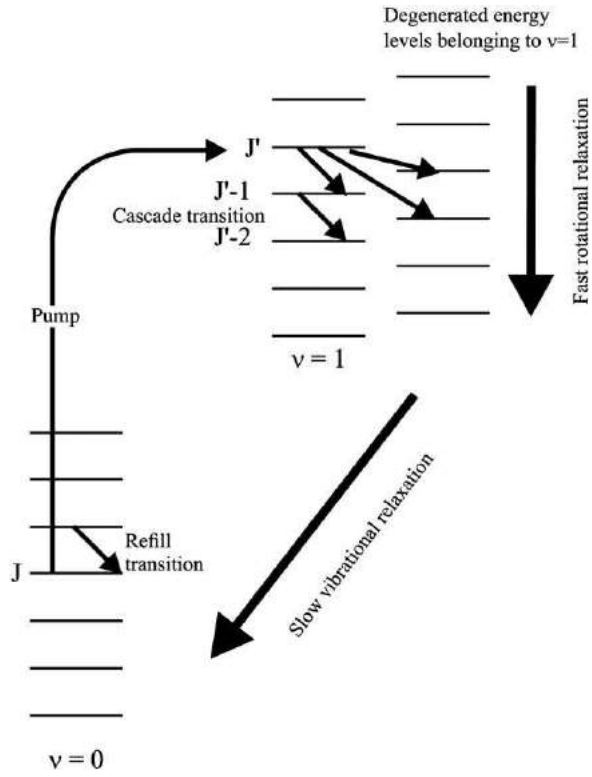


Fig. 1 Schematic of energy levels and transitions in a generic far-IR molecular gas laser. The pump photon excites the selected molecule from its vibration ground state to an excited vibrational state. FIR laser action takes place between rotational levels in the excited state. Figure from Jacobsson, S., 1989. Optically pumped far infrared lasers. Infrared Phys. 29, 853–874.

inversion is not possible if $(h(\nu) < k_B T)$. However, one must remember that a pumped laser system is by its nature not in thermal equilibrium. A population inversion is possible as long as the lower state lifetime is shorter than the upper to lower state relaxation time ($\tau_1 < \tau_{21}$). Rotational transitions in a low-pressure gas have very little broadening ($\sim$1–20 MHz), due to the weak scattering processes (Jacobsson, 1989). As a result, the energy states are very distinct, and can maintain different relaxation rates. Second, the narrow linewidths lead to large cross sections, and high peak levels of gain even for modest population inversion levels.

The inherent limitations of the molecular gas lasers (i.e. size, weight, limited tunability, and low power efficiencies etc.) have created interest in developing terahertz semiconductor lasers. One need look no further than the success of the interband diode lasers in the visible and near-infrared wavelengths, to see the inspiration for developing (potentially) low cost, compact, coherent sources. However, the terahertz photon energies are 100–1000 times smaller than visible photons - appropriate materials that have both a small bandgap and can support a population inversion simply do not exist in nature. As a result, semiconductor terahertz laser research has focused on unconventional sources of stimulated emission. We will highlight three of these here: the germanium intra-valence band lasers, silicon impurity-state lasers, and quantum-cascade lasers.

Germanium Intra-Valence Band Lasers

The first type of THz semiconductor laser demonstrated was the hot-hole p-Ge laser. In this device, lasing action results from a hole population inversion in bulk p-Ge that is established between the light and heavy hole bands due to a "streaming motion" that takes place in crossed electric and magnetic fields (Golka *et al.*, 1991). For low lattice temperatures ($<$20 K), holes have kinetic energies far below the optical phonon energy (37 meV), and their energy relaxation is mediated by (relatively) slow acoustic phonons and impurity scatterings. However, for critical ratios of the electric to magnetic field strength, heavy holes are less bent by the magnetic field due to their heavier mass and therefore can acquire more energy from the electric field and be accelerated to have sufficient kinetic energy to emit an optical phonon, and quickly relax to lower energies. Light holes acquire insufficient energy and do not scatter by the optical phonon. This acts as a pumping mechanism which establishes a population inversion between the light hole and heavy hole bands (**Fig. 2**).

Large fields are required: typical electric fields strengths are from 1 to 2 kV/cm and magnetic fields greater than 1 T. Hence, a liquid-helium cooled superconducting magnet is typically used. The gain is relatively low, as the pumping mechanism is highly nonselective and inefficient, and the upper state lifetime is still relatively short. Free-carrier absorption is also an issue, and the doping must be kept low ($<10^{15}$ cm^{-3}). Nonetheless, due to the large total semiconductor volume, large peak powers of several Watts have been obtained in pulsed mode. Absent a cavity mode selection mechanism, broadband lasing is obtained (spanning over 10–20 cm^{-1}) that can be tuned from 1–4 THz by the applied electric/magnetic fields. Its broadband gain has proven useful for mode-locked operation, where pulses with a 100-ps width have been obtained (Hovenier *et al.*, 2000). However the need for a strong magnetic field, high voltage, and cryogenic operation (T$<$20 K) has limited the utility of this source. Traditionally, due to high power consumption and low efficiency the maximum duty cycle is usually limited to less than 10^{-4}, although a report of up to 5% duty cycle has been demonstrated (Bründermann *et al.*, 2000).

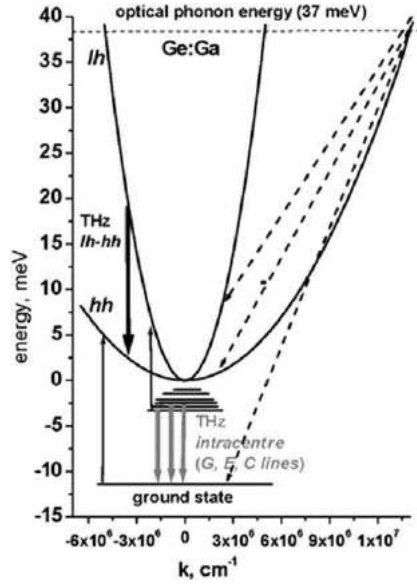


Fig. 2 Schematic of intra-valence band transitions in p-Ge lasers with the hole energies plotted in positive direction. Dashed arrows correspond to relaxation of holes due to interaction with optical phonons; the bold black downward arrow indicates stimulated emission from light- to heavy-hole subband. Figure from Hübers, H.-W., Pavlov, S.G., Shastin, V.N., 2005. Terahertz lasers based on germanium and silicon. Semicond. Sci. Technol. 20, S211–S221.

A related device is the strained p-Ge resonant state laser. *cw* lasing that was tunable with pressure from 2.5 to 10 THz was demonstrated (Gousev *et al.*, 1999). Power levels of tens of microwatts were observed and operation takes place at liquid helium temperatures. In this device, application of strain lifts the degeneracy of the light-hole and heavy-hole band such that the 1 s impurity state of the heavy-hole band is brought into resonance with the light-hole band. This leads to a population inversion between the heavy hole 1 s state and the light-hole impurity states, which are depopulated by electric field ionization. No magnetic field is necessary. A similar laser was demonstrated in 2000 in SiGe/Si quantum wells, where the mechanism of lasing is the same but the strain is provided instead by the epitaxial mismatch (Kagan *et al.*, 2000; Blom *et al.*, 2001). This is a promising method which eliminates the need for externally applied strain, but the power level is still quite low and high voltage (300–1500 V) is required.

Silicon Impurity State Lasers

Silicon is not ordinarily considered a promising material for semiconductor lasers, due to its indirect bandgap. However, *n*-type silicon crystals can be used to make optically pumped 4-level terahertz lasers, based upon donor intracenter radiative transitions. Reviews are given in Hübers *et al,*. (2005), Pavlov *et al.* (2013). Silicon is an attractive material for THz applications in general; unlike III-V semiconductors it is non-polar and has low lattice absorption in the THz range. Various donor impurities have been demonstrated for lasing, including As, P, Sb, and Bi. Excepting a few outliers, lasing has mostly been observed between 1–2 THz and 5–7 THz as shown in **Fig. 3**. Population inversion is obtained between states based upon varying rates of electron relaxation mediated by acoustic and optical phonons. Cavities are formed by polishing bulk crystals with sizes of several millimeters on a side. If desired, stress can be applied to modify the lasing frequency or improve the performance.

A schematic of the energy states and relaxation pathways is shown in **Fig. 3** for several different donor species. Typical donor densities are in the range of 10^{15}–10^{16} cm^{-3}. Pumping occurs via photoionization of the impurity ground state into the conduction band, typically performed using a mid-infrared CO_2 laser. After the electron is excited high into the conduction band, it relaxes via emission of optical and acoustic phonons until it is captured in the high-excited donor states, whereupon it relaxes via emission of acoustic phonons along specific paths that depend upon the energy structure of the particular donor. For example, in Si:Sb and Si:P this relaxation ends in a long lived state ($2p_0$) with an estimated lifetime of hundreds of picoseconds. Lasing occurs between the $2p_0 \rightarrow 1s$ (T_2), where the lower radiative state is estimated to have a lifetime of 10–100ps.

Typical pump threshold intensities are in the range of 10–300 kW/cm^2, depending upon the donor type, concentration, and stress. Peak powers on the order of milliwatts have been measured. In order to avoid thermal ionization of the impurities, and thermal backfilling of the lower radiative state, operation requires cryogenic temperatures typically <10 K, although operation up to 30 K has been observed for Si:Bi lasers. Furthermore, decay dynamics of the nonequilibrium phonon distribution limits the laser operation to pulsed mode – for example in Si:P and Si:Sb lasers emission is quenched after $\sim$150 ns. The operating temperature may be potentially improved by the use of impurities with deeper energy states, so as to increase the energy barrier for thermal backfilling. Research is ongoing to better understand the complexities of the relaxation physics, as well as to improve the performance.

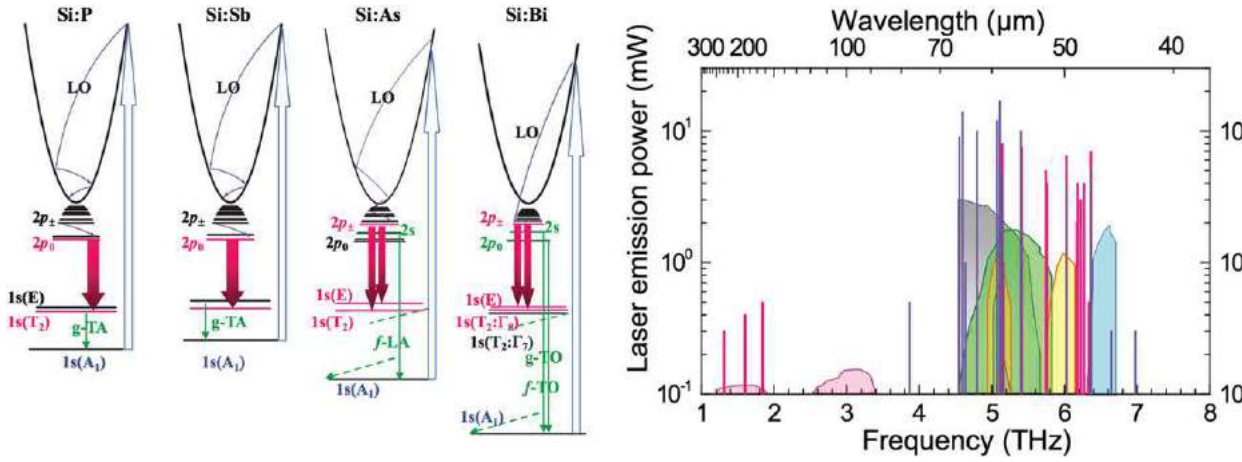


Fig. 3 (left) Laser schemes realized in Si: P, Si: Sb, Si:As and Si:Bi when pumped with a CO_2 laser into the conduction band. The straight blue arrows up indicate the optical pumping and the bold red arrows down are for stimulated emission. Resonant coupling with intervalley phonons is shown by green vertical arrows down. (right) Various reported laser lines from various optically pumped Si impurity state lasers. Figures from Pavlov, S.G., Zhukavin, R.K., Shastin, V.N., H.-W., 2013. The physical principles of terahertz silicon lasers based on intracenter transitions. Phys. Status Solidi B250, 9–36.

Quantum-Cascade Lasers

At present, the most active area in THz laser research is in the field of terahertz quantum-cascade (QC) lasers. The QC-laser is a unipolar intraband laser in which photons are generated as electrons within the conduction band make radiative transitions between quantized subbands engineered into heterostructure quantum wells. By designing sophisticated sequences of coupled quantum wells, "one dimensional artificial molecules" are created with energy levels of the subbands designed to emit photons in the mid-infrared or terahertz spectral range. Furthermore, it is possible to achieve a fine degree of control over electron transport, such that electrons can be injected (usually) via resonant tunneling into an upper radiative state, and can be selectively depopulated from the lower state with some combination of tunneling and scattering processes. The radiative cross-section of the radiative transition can likewise be engineered by controlling the overlap and parity of the various wavefunctions.

The first THz quantum-cascade lasers were demonstrated in 2001 and shortly thereafter (Köhler *et al.*, 2002; Rochat *et al.*, 2002; Williams *et al.*, 2003); their technological progress has been summarized in various review articles (Williams, 2007; Kumar, 2011; Vitiello *et al.*, 2015). The most common material system is GaAs/Al_xGa_{1-x} As heterostructures grown by molecular beam epitaxy, although other material systems have been demonstrated as well (e.g. InGaAs/AlInAs, InGaAs/GaAsSb, etc.). A common feature of these polar III-V semiconductors, is that they all exhibit strong electron interaction with longitudinal optical (LO) phonons (which typically have energies from 30–50 meV). At this time various THz QC-lasers currently have been demonstrated between 1.2 and 5.6 THz (and at frequencies as low as 0.6 THz in a very strong magnetic field).

So far cryogenic operation is still required for THz QC-lasers; the current temperature record is $T_{max}=200$ K in pulsed mode ($<1\%$ duty cycles), and $T_{max}=129$ K in cw mode without the assistance of a magnetic field. By applying a strong magnetic field perpendicular to the planes of quantum wells, additional quantization can be achieved in the transverse dimensions, which reduces the available scattering volume in the momentum space and consequently increases the upper-state lifetimes. As a result, a significant increase of the maximum operating temperature is achieved, from 165 K to 225 K, and the latter is still the record of all the solid-state THz lasers at this writing (Wade *et al.*, 2009). While efforts to improve temperature performance of the active material continue, with ongoing improvements in cryogen-free coolers, and considering the modest power consumption requirements, even operation in the 40–90 K range is very feasible for many applications. For devices operating above 77 K in cw, milliwatt level powers are typical, although in some cases very high powers have been observed. Current record results stand at over 2 W peak power in pulsed mode, and 230 mW in continuous-wave (cw) mode for devices cooled by liquid helium. Furthermore, wall-plug power efficiencies (WPE) greater than 1% have been achieved by several groups.

Active Regions

The heart of the quantum-cascade laser is the multiple-quantum-well active region. Most active region designs can be grouped into three categories: bound-to-continuum (BTC), resonant-phonon (RP) designs, as well as hybrid designs which combine features of each. Schematics for the bandstructure and energy levels are shown in **Fig. 4**.

The key to the BTC design is the coupling of several quantum wells together to create a superlattice which supports a "miniband" of subband states when the appropriate electric field is applied. The radiative transition takes place between an isolated upper "bound" state that resides above the miniband, and a lower state which is the top state of the miniband. Generally speaking, intra-miniband scattering is favored over scattering out of the bound state, which creates a population inversion at lower temperatures. Owing to the relatively small widths of the minibands (about 15–20 meV) for THz QCLs, LO-phonons are not directly involved in the depopulation process. Since the energy drop per module is relatively small, and the oscillator strength is large, this design is often characterized by low threshold voltages and currents, although performance tends to drop off rapidly with rising temperature above 100 K.

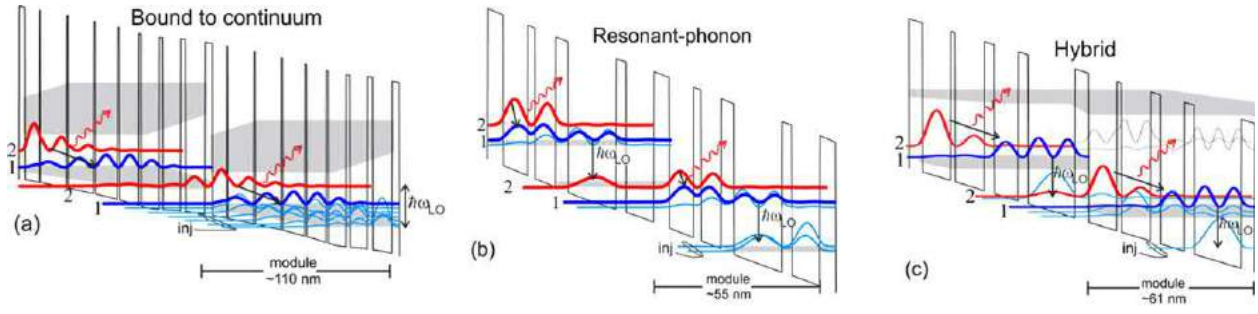


Fig. 4 Schematic figure of the conduction band edge (Gaas/Al_xGa_{1-x}As heterostructures) and energy levels of three types of active regions for quantum-cascade lasers. Two repeated modules are shown in each case. Energy levels are plotted with the probability density of the bound-state. Radiative transition occurs between levels 2 and 1. Also shown for (b) and (c) are electron relaxation via emission of longitudinal optical phonons. Figure adapted from Williams, B.S., 2007. Terahertz quantum cascade lasers. Nat. Photon. 1, 517–524.

The other major active region type is the resonant phonon (RP) scheme. This design, a reservoir/injector state is designed to reside 36 meV (one LO phonon energy) below the lower radiative state. This encourages fast depopulation via electron-LO-phonon scattering into the reservoir state. In order to prevent simultaneous depopulation of the upper state, the energy levels are designed so that electrons are selectively removed via resonant tunneling from the lower state only. In general, RP and hybrid designs have better temperature performance than the BTC designs. Good RP/hybrid designs regularly operate above 130 K and even as high as 200 K in pulsed mode. The explicit inclusion of an LO-phonon scattering event for depopulation results in a more robust depopulation mechanism, as well as a larger energetic barrier to thermal backfilling of electrons from the reservoir.

Several issues are limiting further improvements of the temperature performance of THz QC-lasers so far. Unlike gas lasers, electrons within QC-lasers exist within a high density solid-state matrix: they interact strongly with the lattice, other carriers, and have a large density of states available for scattering. As a result, lifetimes are picoseconds or shorter, and the energy level broadenings are on the order of hundreds of GHz or larger. This not only reduces the peak gain cross-section, but also tends to blur energy levels, making it difficult to maintain selective injection and depopulation essential to achieve a population inversion. This problem is exacerbated for THz QC-lasers at lower frequencies (<2 THz), as the energy spacing becomes comparable with the level broadening.

The second major issue is thermally activated scattering and leakage that degrade the upper state lifetime for high electron temperatures. For example, as electrons in the upper subband acquire sufficient in-plane kinetic energy, they may emit an LO-phonon and relax to the lower subband, or tunnel into the continuum. There is active research underway to suppress these mechanisms, which primarily involves the development of novel active region designs. In addition, new heterostructure material systems are also under investigation. For example, GaN-based materials are attractive since GaN has a much larger LO-phonon energy (90 meV) compared to most III-Vs, which should in principle suppress thermally activated LO-phonon scattering. An even more radical approach is to develop quantum-dot based cascade lasers: the discrete density of states is predicted to suppress nonradiative and dephasing scattering to improve high temperature operation. Nonetheless, no lasers have been reported to date using these new material/structure approaches.

Waveguides and Cavities

The biggest difference between THz QC-lasers and semiconductor lasers at shorter wavelengths lies in the techniques for waveguiding. Conventional dielectric waveguides (as used for semiconductor diode lasers and mid-IR QCLs) are impractical for THz QC-lasers. Due to the $\lambda 2$ scaling of free-carrier loss, the use of doped cladding layers introduces excessive loss at THz wavelengths. Hence, two types of unique waveguides for THz QC-lasers have been developed that minimize the overlap of the mode with doped semiconductor, and instead use metal layers partially or fully for optical confinement; unlike at shorter wavelengths losses are quite modest for noble metals in the terahertz range. These two schemes are shown in Fig. 5.

The first type is the surface-plasmon (SP) waveguide, which involves the growth of a thin (0.2–0.8 μm thick) heavily doped layer underneath the 10 μm-thick GaAs/AlGaAs quantum-well active region, but on top of a semi-insulating GaAs substrate. The resulting mode is a compound surface plasmon tightly confined by the top metal contact, and loosely bound to the heavily doped lower plasma layer. The mode extends far into the substrate (by tens to hundreds of microns); since the substrate is semi-insulating, the free carrier loss is minimal. The downside is that ridges narrower than ~ 100 μm tend to squeeze the mode out of the active region and into the substrate, which effectively puts a floor on the minimum device area (and power dissipation), which in turn limits the maximum achievable cw operating temperature.

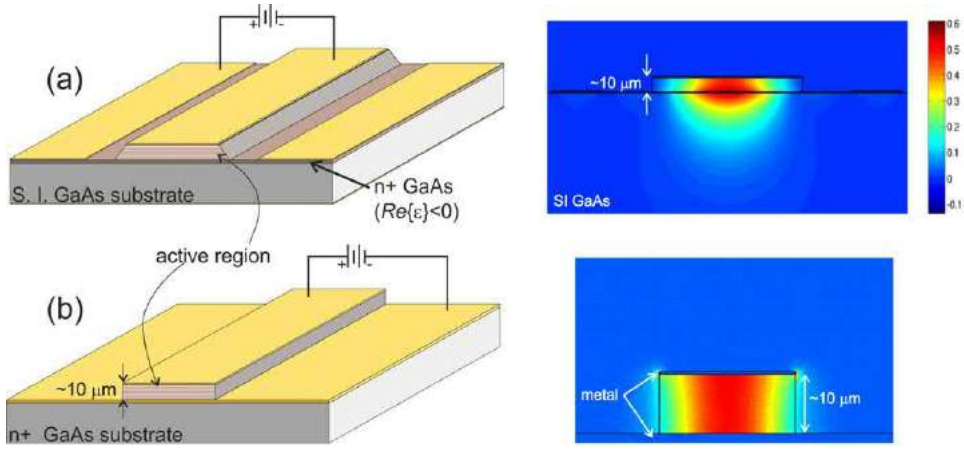


Fig. 5 Schematic of terahertz QC-laser waveguides. Schematic diagram (left) and typical two-dimensional mode intensity pattern (right) for (a) surface plasmon waveguides and (b) metal-metal waveguides. Re(ε) <0 indicates that the real part of the permittivity is less than zero in the heavily doped GaAs laser. Figure from Williams, B.S. 2007. Terahertz quantum cascade lasers. Nat. Photon. 1, 517-524.

An alternative to the SP waveguide is the so-called "metal-metal" (MM) waveguide, in which the waveguide mode is tightly confined between metal cladding placed immediately above and below the ~10 μm thick epitaxial active region. The overall result resembles a microstrip transmission line. MM waveguides tend to have the best high-temperature performance, mostly as a result of lower overall losses (both absorption and radiative) compared to the SP waveguide. Furthermore, the strong modal confinement of MM waveguides allows both the vertical and lateral dimensions to be made much smaller than the wavelength. This feature is unique for a new genre of lasers termed "photonic wire lasers", in which a large fraction of mode propagates outside of the solid core. This in turn reduces the total thermal dissipation and required cooling power, which was key to observation of the highest temperature cw operation (129 K) in a metal-metal waveguide (Wienold *et al.*, 2014). Furthermore, it is possible to mechanically perturb the portion of the mode outside the core using MEMS actuators, which has allowed broad single-mode tuning of up to 330 GHz (Qin *et al.*, 2011).

While metal-metal waveguides are preferred in terms of temperature performance, if a MM ridge waveguide is simply cleaved to form a Fabry-Pérot cavity, it performs poorly as an edge emitting laser. The cleaved facet radiates as a sub-wavelength sized aperture, which exhibits an extremely divergent beam with a poor radiation efficiency. Hence, an active area of research has emerged in the design of novel cavity types to allow high quality beams simultaneously with high output power.

A straightforward approach is to mount a silicon lens flush to the metal-metal waveguide facet, which helps to improve the beam quality and improve the impedance matching. Another class of approaches uses Bragg scattering from periodic structures to reshape the beam. Second-order distributed feedback (DFB) and 2D photonic crystal cavities have demonstrated surface emission; the beam is much improved since the waveguide surface, not the facet, is now the radiating aperture. One challenge that emerges is that the DFB laser prefers to lase in the mode with the lowest total losses, which usually is a weakly radiating (high-Q) band-edge mode with low output power. A variety of schemes are under investigation, including graded photonic heterostructures that force the laser to lase in the strongly radiating symmetric band-edge mode, and other structures (e.g. quasi-periodic crystals, dual slit gratings, and others) that break the lattice symmetry to increase radiative efficiency.

End-fire QC-lasers based upon 3rd-order DFBs or antenna-coupled DFB cavities are another attractive option; they can achieve very narrow far field beams even when the transverse waveguide cross section is sub-wavelength (Amanti *et al.*, 2009). The beam is formed by a linear phased array of scatters along the length of the cavity; the beam divergence scales inversely with square root of the cavity length for a perfectly phase-matched 3rd-order DFB laser. Hence, this allows narrow photonic wire cavities to be used, which further keeps the power dissipation low for good cw operation. Slope efficiencies can be made large by including antenna-like structures into the cavity. These approaches are particularly effective for milliwatt-scale power output.

Still another class of approaches involves the phase locking of multiple smaller emitting elements to achieve high quality beams and to scale up the power by power combining. While arrays of 2nd order DFBs have been demonstrated, it is challenging to phase lock more than a few elements. Two new approaches address this issue. One is to use radiative coupling to couple large arrays of sub-wavelength sized QC-laser cavities – for appropriate lattice spacings the individual cavities will prefer to lase in an in-phase collective supermode that provides a directive high quality beam with high slope efficiency (Kao *et al.*, 2016). If this array is biased slightly below threshold, it can be used as a reflective THz amplifier. The other approach is similar, but deliberately uses arrays of sub-wavelength cavities with low-quality factor (so they do not self-oscillate) to form a reflective amplifying "metasurface". When this metasurface is placed into an external cavity, these emitters lock to the cavity mode (Xu *et al.*, 2015). Such a device, known as a metasurface vertical-emitting-cavity surface-emitting-laser (VECSEL), not only allows scalable power with a high-quality beam, but detailed control over the phase, spectral, and polarization response of the metasurface.

THz QC-Laser Frequency Combs

A rapidly evolving area of current research is the development of broadband, multi-mode THz QC-lasers that operate as frequency combs (Burghoff *et al.*, 2014). QC-lasers can be designed to exhibit gain over very large fractional gain bandwidths. This feature emerges naturally because certain designs of active region exhibit emission from multiple broadened transitions; furthermore, the active region can be made deliberately inhomogeneous by sandwiching multiple stacks, each intended to emit at different wavelengths (Rösch *et al.*, 2015). If the intracavity dispersion is properly managed (so that each longitudinal mode sees the same round trip time), then the multiple modes will interact via four-wave-mixing to establish a phase coherent comb. This is a different mechanism than the traditional method of comb generation based upon ultrafast mode-locked lasers - instead of a circulating intracavity pulse, the intensity inside the laser cavity is much closer to a constant value.

Perhaps the largest application is rapid, high-precision dual-comb spectroscopy, in which two combs (signal and local oscillator) with slightly detuned comb tooth spacings are mixed. This technique effectively mimics the functionality of Fourier-Transform Spectroscopy, but without the need for a mechanically moving mirror. The best reported comb bandwidths are approximately 1 THz, which still falls short of the octave spanning bandwidth needed to allow f-2f self-stabilization of the comb. For this reason, a major focus of current research is the effort to increase the comb bandwidth, by engineering both the active region and the cavity dispersion, in order to increase the gain bandwidth while simultaneously managing dispersion.

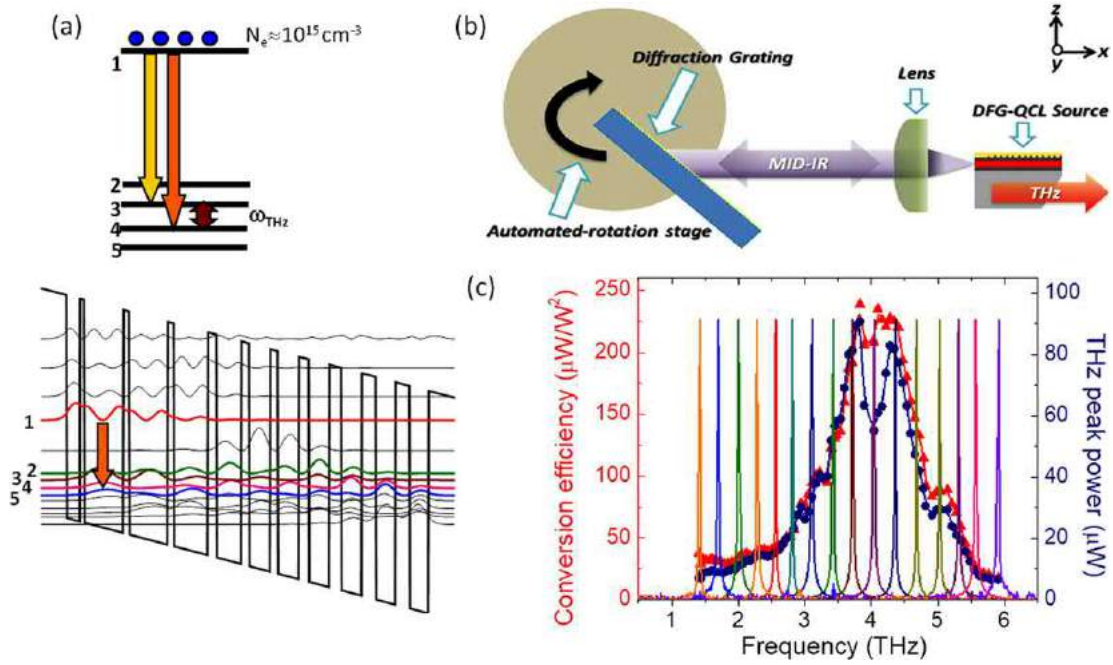


Fig. 6 (a) Schematic of mid-IR QC-laser active region band diagram, indicating the levels primarily involved with the DFG process. (b) Schematic of external cavity setup for tunable THz DFG. (c) Spectra of THz DFG over tuning range. Figure adapted from Vijayraghavan, K., Jiang, Y., Jang, M. *et al.*, 2013. Broadly tunable terahertz generation in mid-infrared quantum cascade lasers. Nat. Commun. 4, 2021.

Nonlinear Generation of THz Radiation in Mid-IR QC-Lasers

There exists a separate approach for generating THz radiation in QC-lasers that uses nonlinear downconversion from shorter wavelengths rather than direct lasing. Specifically, room-temperature THz radiation has been generated using intracavity difference frequency generation (DFG) within mid-infrared QC-lasers lasing at two wavelengths (Belkin *et al.*, 2007). Generation of THz radiation in nonlinear crystals via downconversion has a long history - however usually the limited strength of the available $\chi^{(2)}$ nonlinearity has required large laser intensities; as a result these techniques have required high power pulsed lasers (both Q-switched and mode-locked) in the visible and near-IR.

Mid-IR QC-lasers with wavelengths near 10 μm are more attractive, since the smaller pump photon energy results in an improved Manley-Rowe ratio, and the lasers themselves are physically smaller. Furthermore, the quantum-cascade laser material itself can be engineered to possess an extremely large resonant $\chi^{(2)}$ nonlinearity based upon the mid-IR intersubband transitions (**Fig. 6**). Ordinarily, resonant nonlinearities are accompanied by large linear absorption losses, but because this nonlinearity is associated with the same intersubband transitions that produce the mid-IR laser gain, the resonant absorption is avoided. This allows efficient mixing to take place within the laser cavity, where the circulating intensities are high. This approach leverages the greater technological maturity of mid-IR lasers in the 4–10 μm range, where room-temperature continuous-wave operation is possible with hundreds of mW or more.

While the initial demonstrations of THz DFG had very low conversion efficiency and sub-microwatt pulsed powers, significant advances have taken place recently. Currently, the highest reported THz power levels at room-temperature are 1.9 mW of peak power in pulsed mode, and 14 μW of power in cw mode (Lu and Razeghi, 2016). These sources are particularly promising for applications which require broad tunability, since only modest tuning of one of the mid-IR modes translates to large tuning of the THz beat note. For example, tuning of a single mode from 1.7–5.3 THz was shown in (Vijayraghavan *et al.*, 2013). Work is underway to further increase the nonlinear conversion efficiency, improve the efficiency of out-coupling the THz radiation from the cavity, as well as the increase the range and ease of tunability.

See also: Broadband Terahertz Sources. THz Molecular Spectroscopy

References

Amanti, M.I., Fischer, M., Scalari, G., Beck, M., Faist, J., 2009. Low-divergence single-mode terahertz quantum cascade laser. Nature Photon. 3, 586–590.
Belkin, M.A., Capasso, F., Belyanin, A., *et al.*, 2007. Terahertz quantum-cascade-laser source based on intracavity difference-frequency generation. Nature Photon. 1, 288–292.

Blom, A., Odnoblyudov, M.A., Cheng, H.H., Yassievich, I.N., Chao, K.A., 2001. Mechanism of terahertz lasing in SiGe/Si quantum wells. Appl. Phys. Lett. 79, 713.
Bründermann, E., Chamberlin, D.R., Haller, E.E., 2000. High duty cycle and continuous terahertz emission from germanium. Appl. Phys. Lett. 76, 2991.
Burghoff, D., Kao, T.-Y., Han, N., *et al.*, 2014. Terahertz laser frequency combs. Nat. Photon. 8, 462–467.
Chang, T.Y., Bridges, T.J., 1970. Laser action at 452, 496, and 541 μm in optically pumped CH_3F. Opt. Commun. 9, 423–426.
Crocker, A., Gebbie, H.A., Kimmitt, M.F., Mathias, L.E.S., 1964. Stimulated emission in the far infra-red. Nature 201, 250–251.
Dodel, G., 1999. On the history of far-infrared (FIR) gas lasers: thirty-five years of research and application. Infrared Phys. Technol. 40, 127–139.
Farhoomand, J., Blake, G.A., Frerking, M.A., Pickett, H.M., 1985. Generation of tunable sidebands in the far-infrared region. Proc. SPIE 598, 84–87.
Golka, S., Pflügl, C., Schrenk, W., Strasser, G., 1991. Special issue – Far-infrared semiconductor lasers. Opt. Quantum Electron. 23, S111.
Gousev, Y.P., Altukhov, I.V., Korolev, K.A., *et al.*, 1999. Widely tunable continuous-wave THz laser. Appl. Phys. Lett. 75, 757.
Hovenier, J.N., Diez, M.C., Klassen, T.O., *et al.*, 2000. The p-Ge terahertz laser propeties under pulsed- and mode-locked operation. IEEE Trans. Microwave Theory Tech. 48, 670.
Hübers, H.-W., Pavlov, S.G., Shastin, V.N., 2005. Terahertz lasers based on germanium and silicon. Semicond. Sci. Technol. 20, S211–S221.
Inguscio, M., Moruzzi, G., Evenson, K.M., Jennings, D.A., 1986. A review of frequency measurements of optically pumped lasers from 0.1 to 8 THz. J. Appl. Phys. 60, R161.
Jacobsson, S., 1989. Optically pumped far infrared lasers. Infrared Phys 29, 853–874.
Kagan, M.S., Altukhov, I.V., Sinis, V.P., *et al.*, 2000. Terahertz emission of SiGe/Si quantum wells. Thin Solid Films 380, 237.
Kao, T.-Y., Reno, J.L., Hu, Q., 2016. Phase-locked laser arrays through global antenna mutual coupling. Nat. Photon.
Köhler, R., Tredicucci, A., Beltram, F., *et al.*, 2002. Terahertz semiconductor-heterostructure laser. Nature 417, 156.
Kumar, S., 2011. Recent progress in terahertz quantum cascade lasers. IEEE J. Sel. Top. Quantum Electron. 17, 38–47.
Lu, Q., Razeghi, M., 2016. Recent advances in room temperature, high-power terahertz quantum cascade laser sources based on difference-frequency generation. Photonics 3, 42.
Mueller, E.R., Henschke, R., William, E., *et al.*, 2007. Terahertz local oscillator for the Microwave Limb Sounder on the Aura satellite. Appl. Opt. 46, 4907–4915.
Pavlov, S.G., Zhukavin, R.K., Shastin, V.N., Hubers, H.-W., 2013. The physical principles of terahertz silicon lasers based on intracenter transitions. Phys. Status Solidi B 250, 9–36.
Qin, Q., Reno, J.L., Hu, Q., 2011. MEMS-based tunable terahertz wire-laser over 330GHz. Opt. Lett. 36, 692–694.
Rochat, M., Ajili, L., Willenberg, H., *et al.*, 2002. Low-threshold terahertz quantum-cascade lasers. Appl. Phys. Lett. 81, 1381.
Rösch, M., Scalari, G., Beck, M., Faist, J., 2015. Octave-spanning semiconductor laser. Nat. Photon. 9, 42–47.
Vijayraghavan, K., Jiang, Y., Jang, M., *et al.*, 2013. Broadly tunable terahertz generation in mid-infrared quantum cascade lasers. Nat. Commun. 4, 2021.
Vitiello, M.S., Scalari, G., Williams, B.S., Denatale, P., 2015. Quantum cascade lasers: 20 years of challenges. Opt. Express 23, 5167–5182.
Wade, A., Federov, G., Smirnov, D., *et al.*, 2009. Magnetic-field-assisted terahertz quantum cascade laser operating up to 225 K. Nat. Photon. 3, 41–45.
Wienold, M., Röben, B., Schrottke, L., *et al.*, 2014. High-temperature, continuous-wave operation of terahertz quantum-cascade lasers with metal-metal waveguides and third-order distributed feedback. Opt. Express 22, 3334–3348.
Williams, B.S., 2007. Terahertz quantum cascade lasers. Nat. Photon. 1, 517–524.
Williams, B.S., Callebaut, H., Kumar, S., Hu, Q., Reno, J.L., 2003. 3.4-THz quantum cascade laser based on longitudinal-optical-phonon scattering for depopulation. Appl. Phys. Lett. 82, 1015.
Xu, L., Curwen, C.A., Hon, P.W.C., *et al.*, 2015. Metasurface external cavity laser. Appl. Phys. Lett. 107, 221105.

THz Molecular Spectroscopy

Zbigniew Kisiel, Institute of Physics of the Polish Academy of Sciences, Warsaw, Poland

Introduction

The THz region is nominally defined by the International Telecommunication Union (ITU) as covering 0.3–3 THz. It has been exploited by several branches of molecular spectroscopy, each of which describes it in terms of its customary units of either frequency, wavelength or wavenumber. **Table 1** provides a convenient summary.

The most useful and, at the same time, the most challenging molecular spectroscopy application of the THz region is high resolution spectroscopy in the gas phase. This takes advantage of the narrow linewidths of molecular absorption lines and of the ability of contemporary instrumental techniques to deliver molecular spectra, which are only limited by molecular properties and not by instrumental factors. The apparent widths of room temperature lines free from broadening effects are at the MHz level for THz frequency transitions, so that frequency control of significantly better than one part in 10^7 is mandatory. Absorption coefficients of such transitions are usually smaller or much smaller than 10^{-3} per cm of absorption path so that considerable source stability and multiple techniques for enhancing the recovered signal are also required.

Rotational Spectroscopy

The principal physical phenomenon observed in THz molecular spectra is quantized molecular rotation. In the microscopic world rotational energy can only take on very well defined values, or energy levels. Those are labeled by using quantum numbers corresponding to the total molecular angular momentum and to its projection onto a selected axis in the molecule. The rotational energies also depend on the molecular moment of inertia, which is expressed in terms of its components I_a, I_b, I_c along a unique set of center of mass Cartesian axes, called principal axes a, b, c. Transitions between rotational energy levels are allowed if the molecule possesses a permanent electric dipole moment, μ, and then its components μ_a, μ_b, μ_c, along the three principal axes are of relevance. Non-zero value of each such component leads to specific selection rules, or allowed changes in quantum number values for absorption of electromagnetic radiation inducing a change in rotational energy. Of relevance to all types of rotational spectra is the fact that frequencies of transitions depend directly on rotational constants, A, B, C, which are inverse quantities to the three principal moments of inertia. Since the moments of inertia only depend on atomic masses and on their positions in the molecule, their spectroscopic determination provides a route to precise experimental molecular structures. Furthermore, rotational transitions of a given molecule form a very characteristic and precisely measurable fingerprint pattern. This enables various analytical applications, as the high specificity allows unambiguous determination of molecular compositions in complex gas phase mixtures. This is actually the only method for identifying and quantifying the presence of molecules in the interstellar medium, currently at 200 or so species. Precise measurement of line intensities also allows quantitative determination of molecular abundances.

There are two limiting spectroscopic molecular rotation cases, depending on the general molecular shape: whether it is close to a cylinder (prolate rotor) or to a disk (oblate rotor). Rotational energies (in frequency units) are then approximately given by:

$$\text{prolate rotor } (A>B=C): \quad E_{JK} = BJ(J+1) + (A-B)K^2 + \text{smaller terms} \tag{1}$$

$$\text{oblate rotor } (A=B>C): \quad E_{JK} = BJ(J+1) + (B-C)K^2 + \text{smaller terms} \tag{2}$$

where J and K are the quantum numbers corresponding to the total molecular angular momentum and to its projection onto the molecular symmetry axis, respectively. The selection rules in both limiting cases are identical, being $\Delta J=1$ and $\Delta K=0$, and leading to a deceptively simple expression for transition frequencies:

$$\nu = 2B(J+1) + \text{smaller terms} \tag{3}$$

Most molecules are actually asymmetric rotors, with all three rotational constants different. Their quantum mechanical description uses the values of K for the two limiting symmetric top cases, renamed K_a (prolate) and K_c (oblate) to label asymmetric

Table 1 Limits of the THz region in units used in various branches of molecular spectroscopy

	Lower bound	*Upper bound*	*Unit*
Frequency	0.3	3	THz
	300	3000	GHz
Wavelength	1	0.1	mm
	1000	100	μm
Wavenumber	10	100	cm^{-1}

rotor energy levels, by means of the notation J_{KaKc}. The frequencies of the most relevant transition types can also be usefully predicted by replacing $2B$ in Eq. (3) with $(B+C)$ for a prolate asymmetric rotor or by $(A+B)$ for an oblate one. In practice, precise reproduction of measured transition frequencies requires also inclusion of terms from several other effects, such as distortion of molecular structure on rotation (centrifugal distortion), or splitting due to the presence in the molecule of atoms with non-zero nuclear quadrupole (nuclear quadrupole hyperfine splitting).

From the point of view of THz spectroscopy disk type asymmetric rotors are less relevant due to a more compressed energy level structure as implied by Eq. (2). On the other hand, prolate rotors can have sizable values of the A rotational constant, and Eq. (1) allows a much faster increase in rotational energies with increasing values of quantum number K_a, leading to significant numbers of THz region transitions. This is also partly due to basic spectroscopic properties associated with the dipole moment components responsible for the observed transitions. The selection rules leading to the K independent transitions described by Eq. (3) arise from the non-zero dipole moment component directed close to the symmetry axis of the molecule (μ_a for a prolate, μ_c for an oblate rotor). In an asymmetric rotor the third, μ_b, dipole moment component is possible and is associated with rather more complex selection rules for transitions, which introduce frequency dependence on the K_a quantum number, and thus directly on the value of the potentially sizable A rotational constant. In this way the $\mu_b \neq 0$ case can give rise to plentiful b-type rotational transitions in the THz.

As a boundary example: in the very light water molecule μ_b is the only non-zero dipole moment component and the rotational constants are $A=835.8$, $B=435.3$, $C=278.1$ GHz. In consequence, only two significant rotational transitions of water, $6_{16} \leftarrow 5_{23}$ at 22.235 GHz and $3_{13} \leftarrow 2_{20}$ at 183.410 GHz, fall below the THz region, whereas the actual water rotational spectrum extends through the THz region and beyond. While this is an important case, it is not typical, as rotational constants for molecules with significant THz region spectra are most often in the 1–10 GHz region. A more representative example is the relatively rigid acrylonitrile molecule, $CH_2=CHC\equiv N$, which gives rise to the THz spectrum displayed in **Fig. 1**. The rotational constants in this case are $A=49.95$, $B=4.97$, $C=4.51$ GHz and lead to plentiful transitions allowed by non-zero μ_a and μ_b dipole moment components, due to the orientation of the molecular dipole moment shown in **Fig. 2**. At 1 THz the b-type transitions dominate and comprise of ^bR-branch ($\Delta J=1$) and ^bQ-branch ($\Delta J=0$) transitions. While it is not possible to derive simple expressions for frequencies of ^bR-branch transitions those involve an $(A+B)(J+1)$ leading term. A more successful approximate relation for the dense bands made up of ^bQ-branch transitions is:

$$\nu = [A-(B+C)](2K_a+1) + \text{smaller terms} \tag{4}$$

Of course, an appropriate treatment of the problem with a suitable quantum mechanical Hamiltonian allows reproduction of experimental frequencies to measurement accuracy, but this is the domain of dedicated computer programs. Nevertheless, it is easy to see from the above how a sizable value of the rotational constant A can result in a rich THz region spectrum. The relative intensities of a-type and b-type transitions and their temperature dependence are illustrated in **Fig. 2**. The a-type transitions lose relevance in relation to b-type transitions simply because at a given THz frequency the former involve a much greater value of J.

THz region rotational transitions in excited vibrational states of molecules, can also be of appreciable intensity, as has already been seen in **Fig. 1**. The intensities of such transitions will be reduced relative to those in the ground state by the standard Boltzmann population factor of $e^{-\Delta E/kT}$. At room temperature, rotational transitions in an excited vibrational state positioned 100 cm^{-1} above the ground state will have over 60% of the intensity of the ground state transitions, while the $\nu_{11}=1$ state lines in

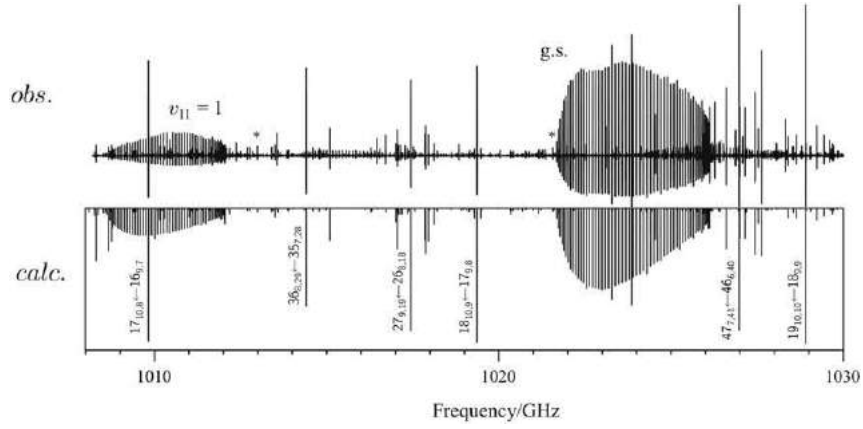


Fig. 1 The room temperature spectrum of acrylonitrile recorded at 1 THz. It consists almost entirely of b-type rotational transitions in the ground state (g.s.) and in the lowest excited vibrational states (such as the marked $\nu_{11}=1$). Quantum numbers for the strongest ^bR-branch ($\Delta J=1$) transitions are indicated on the figure, while the two bands of dense lines are ^bQ-branch transitions ($\Delta J=0$) for $K_a=11 \leftarrow 10$. The strongest a-type transitions in this region are marked with asterisks. Reproduced from Kisiel, Z., Pszczółkowski, L., Drouin, B.J., *et al.*, 2009. The rotational spectrum of acrylonitrile up to 1.67 THz. Journal of Molecular Spectroscopy 258, 26–34.

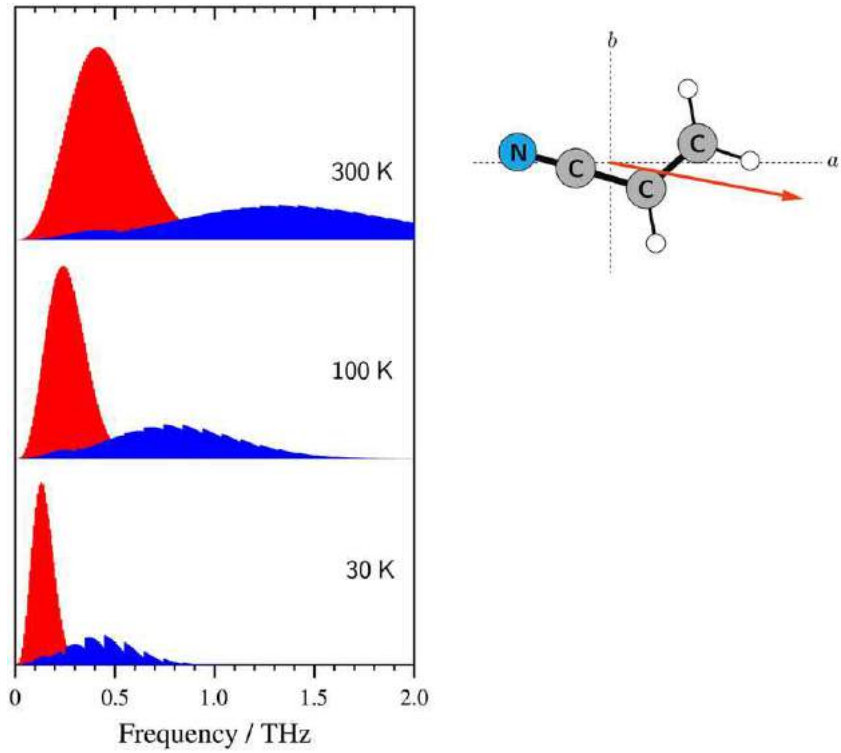


Fig. 2 The nature of the THz rotational spectrum of acrylonitrile and its temperature dependence. The dipole moment of the molecule (with orientation depicted by the arrow) is made up of a large μ_a component giving rise to stronger but lower frequency transitions (red bands), and a smaller μ_b component resulting in weaker but higher frequency transitions (blue bands). At the lower, astrophysical temperatures, the nominally weaker *b*-type transitions dominate above 0.5 THz. Reproduced from Kisiel, Z., Pszczółkowski, L., Drouin, B.J., *et al.*, 2009. The rotational spectrum of acrylonitrile up to 1.67 THz. Journal of Molecular Spectroscopy 258, 26–34.

Fig. 1 originate from a state with vibrational energy of 228 cm^{-1} and are consequently at 33% of ground state intensity. The situation becomes more complex when the molecule has a clearly defined internal rotor, which can have several equivalent orientations relative to the remaining molecular frame. The most common case is that of a C_{3v}-symmetry internal rotor (typically a CH_3 group) in a C_s-symmetry molecule (containing a symmetry plane). The existence of three equivalent minima possible on internal rotation (torsional motion) of the methyl group gives rise to a splitting of the ground state, and of each excited vibrational state, into two substates, called A and E after the C_3 group classification of such motion. Separate rotational level stacks are possible for each substate, giving rise to two sets of rotational transitions of comparable intensity. For a sizable internal rotation barrier the frequencies of A and E-substate transitions are not too different resulting in characteristic A–E doubling of lines in the spectrum. When the internal rotation barrier is lower the vibrational E–A splitting increases and for the important methanol molecule is close to 9 cm^{-1}. Since methanol is a rather light molecule with comparable masses of the internal rotor and of the OH frame the frequencies of A and E ground state transitions become much more difficult to analyse.

Rotation Vibration Spectroscopy

Pure rotation is not the only phenomenon of relevance to high resolution THz molecular spectroscopy. A related one is vibrational spectroscopy with resolved rotational structure, in the form of transitions between rotational level stacks belonging to different vibrational states. Typical vibrational energy level differences are in excess of 100 cm^{-1} so that transitions between vibrational states with different vibrational quantum numbers typically fall in the infrared region and are relatively rare in the THz. But there are special cases, one of which is an important class of molecules endowed with a relatively low-barrier inversion motion. When two symmetry equivalent orientations of the molecule are possible then quantum tunneling through the separating barrier leads to splitting of each vibrational state of the inversion motion into two components. In this way vibrational state $v=0$ is split into 0^+ and 0^- substates, $v=1$ into 1^+ and 1^- substates, etc. Rotation-vibration transitions between a pair of such substates are allowed by a vibrational dipole moment along the inertial axis connecting the two minima. The most celebrated case of this type is the inversion splitting spectrum between the 0^+ and 0^- substates of the ammonia molecule, which consists of a multitude of transitions centered at 23.8 GHz. This was the first microwave molecular spectrum observed as far back as the 1930s, and caused some confusion since it was not a pure rotation spectrum. The tunneling splitting between the next higher, 1^+ and 1^-, inversion doublet of ammonia is 1056 GHz so that the respective rotation-vibration transitions fall in that region and have been studied by THz spectroscopy.

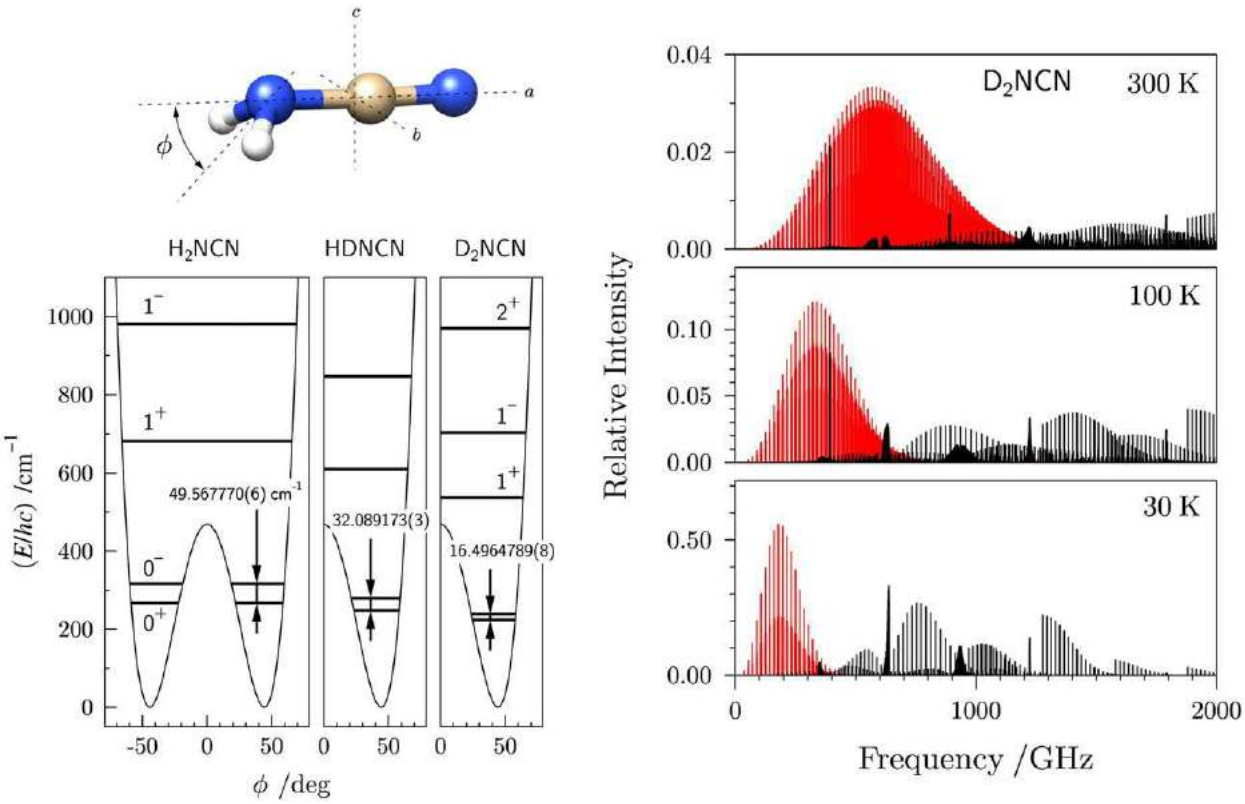


Fig. 3 The origin and temperature dependence of THz transitions for the cyanamide molecule. The pure rotation spectrum (red in the right panel for D_2NCN) arises from the large permanent μ_a dipole moment component, while higher frequency lines (black in the right panel) are due to *c*-type rotation-vibration transitions between the two inversion substates, 0^+ and 0^-, of the ground state. Global analysis of the available transitions allowed precise determination of inversion splitting (marked in the left panel) illustrating its considerable isotope dependence. Reproduced from Kisiel, Z., Kraśnicki, A., Jabs, W., *et al.*, 2013. Rotation and rotation–vibration spectroscopy of the 0^+–0^- inversion doublet in deuterated cyanamide. Journal of Physical Chemistry A 117, 9889–9898 with permission.

A related molecule to ammonia is cyanamide, $H_2NC{\equiv}N$, in which the inversion motion at the pyramidal nitrogen nucleus is characterized by a significantly lower barrier of below 500 cm^{-1}. In consequence, the $0^+ - 0^-$ inversion splitting ranges from 0.495 to 1.486 THz (Fig. 3) and creates ample opportunities for THz molecular spectroscopy. Furthermore, cyanamide is a light molecule so that its pure rotation spectrum is intertwined with the vibration-rotation spectrum, even for the relatively heavy $D_2NC{\equiv}N$ isotopologue, as shown in Fig. 3.

Laboratory Molecular Spectroscopy

The ability to record THz molecular spectra in the laboratory is critically determined by the techniques used to generate highly controlled electromagnetic radiation in this frequency region. Generation needs to be complemented by suitable methods of detection and, in an associated step, by efficient extraction of molecular signals. Various modulation techniques are employed for the latter, generally making use of the narrow natural width of molecular absorption lines. Linewidths have a complex dependence on quantum numbers of the involved transitions but can, in general, be regarded to range from 1 MHz full-width at half height (FWHH) at the lower end of the THz region, to 10 MHz FWHH at the upper end. Several influential techniques of laboratory spectroscopy are described below, chosen for their current relevance, or promise for further molecular spectroscopy applications.

Sources

There are many sources that give access to the THz frequency region, but molecular spectroscopy requires high quality performance on routine operation. There is considerable premium on resolution and on broadband access, namely recording of spectra over considerable frequency regions. The spectra contain many features that are not amenable to direct interpretation, and can often be satisfactorily understood only upon a protracted off-line analysis. The highest resolution is possible with tunable monochromatic radiation, although promising time domain microwave techniques are appearing. Somewhat lower resolution, but very broad

Table 2 THz region coverage available with the cascaded harmonic multiplication sources used at the Jet Propulsion Laboratory[a]

Synthesizer frequency (GHz)	N[b]	*Output frequency (GHz)*	*Synthesizer frequency (GHz)*	N	*Output frequency (GHz)*
16.1–17.8	18	290–320	17.6–20.0	54	950–1080
13.0–17.7	30	390–530	14.7–17.2	72	1055–1240
14.8–17.0	36	534–612	12.1–13.2	108	1305–1425
12.6–14.8	54	680–800	13.1–14.8	108	1420–1600
12.8–14.2	60	770–850	14.8–17.1	108	1600–1850
11.8–12.9	72	850–930	16.5–18.5	108	1780–2000
			15.2–17.2	162	2470–2780

[a]Yu, S. Personal communication.
[b]The total multiplication factor of the frequency of the driving microwave synthesizer.

frequency coverage is also possible with Fourier transform far-infrared interferometers, especially those using synchrotron radiation sources.

High-resolution spectroscopic access to the THz region was pioneered in the 1950s by the research group of Gordy at Duke University, Durham, United States (see Chapter 2.1 of Rao (1976) for a review), who carried out many seminal studies including those of the many isotopologues of the water molecule. They used point contact diode multiplication of radiation from microwave klystrons and point contact diode detectors, a technology that was adopted in many other laboratories. The use of klystrons and of oscilloscope based detection limited the spectroscopic access to relatively narrow windows. Nevertheless, there are similarities between this approach and the technique that has over the recent years turned out to be the optimum source for routine THz molecular spectroscopy. This is cascaded harmonic multiplication of microwave radiation, in which a 20 GHz region microwave synthesizer drives a stack of active and passive microwave multipliers. The principal advantage is that it is left to the synthesizer to provide frequency stabilization and tunability, which is then simply transferred to THz frequencies. In practice, years of development have gone into optimizing the multiplier combinations, and into ensuring harmonic purity. The broadest coverage has been developed at the Jet Propulsion Laboratory (JPL), Pasadena, United States (Drouin *et al.*, 2005) as summarized in **Table 2**. Usable output powers range from the mW level at the low frequency end to sub-μW level at the highest frequency bands. The total multiplication factor of the driving synthesizer frequency is the product of the multiplication factors of the individual multipliers in the stack and, for example, $18 = 6 \times 3$ for the lowest frequency source in **Table 2**, and $108 = 6 \times 2 \times 3 \times 3$ for the 1.3–2.0 THz sources. This is a reasonably mature technology and cascaded multiplication sources in many configurations are available commercially.

A very attractive class of sources are high frequency backward wave oscillators (BWO). These are vacuum tubes, developed in the Soviet Union, and revealed to the spectroscopic community in the 1970s in publications of the research group at the Institute of Applied Physics, Nizhny Novgorod, Russia (then Gorky, Soviet Union) (reviewed in Chapter 2.2 of Rao (1976)). These sources generate THz radiation directly, have relatively high output power, excellent electronic tunability, and linewidth estimated at less than 100 Hz. The drawbacks are that high voltage and a sizable magnetic field are required for operation, and it is necessary to provide external frequency stabilization by means of phase locking techniques. Devices with operation of up to 1.4 THz have been produced and have been used in many spectroscopic laboratories throughout the world. The most prolific adopters have been the research group at the University of Cologne, Germany (as reviewed in Schlemmer *et al.* (2009)) who used such sources in both fundamental generation mode and with harmonic multiplication. The unique properties of these BWOs have also allowed development of several novel spectrometer designs: the fast scan spectrometer at The Ohio State University, Columbus, United States, and the resonator spectrometer for broadband measurements of atmospheric absorption lines at the Nizhny Novgorod laboratory. Both designs are discussed further below.

More specialized, high-resolution access to the THz region is also possible by using difference frequency laser radiation, or laser sideband generation. One such approach has been to use a frequency stabilized far-infrared molecular laser and mix it with a 300 GHz BWO source, by means of a rooftop optical mixer (Schlemmer *et al.*, 2009). This approach allowed μW level electronically swept frequency operation near the end of the THz region.

Broadband access to most of the THz region has also been possible for a considerable time with far-infrared configurations of Fourier-transform infrared (FTIR) interferometers. The available resolution has been steadily increasing with the increase in optical path difference available from the movable mirror of the interferometer. Currently, resolutions in the region of 0.001 cm^{-1} (30 MHz) are possible. The considerable length of the optical path associated with the very long travel of the movable mirror requires a strong radiation source for sample illumination, so that the majority of such instruments are installed on beamlines of synchrotron radiation sources. A comparison of the 1.8 THz molecular spectrum of the same molecule recorded with microwave techniques and with optical interferometric techniques is shown in **Fig. 4**. While the advantages of microwave based techniques are clearly apparent it should be borne in mind that FTIR gives access to practically the whole THz region in one continuous spectrum. The difference in resolution and frequency measurement precision between the two traces in **Fig. 4** is only an order of magnitude and may in some cases be acceptable.

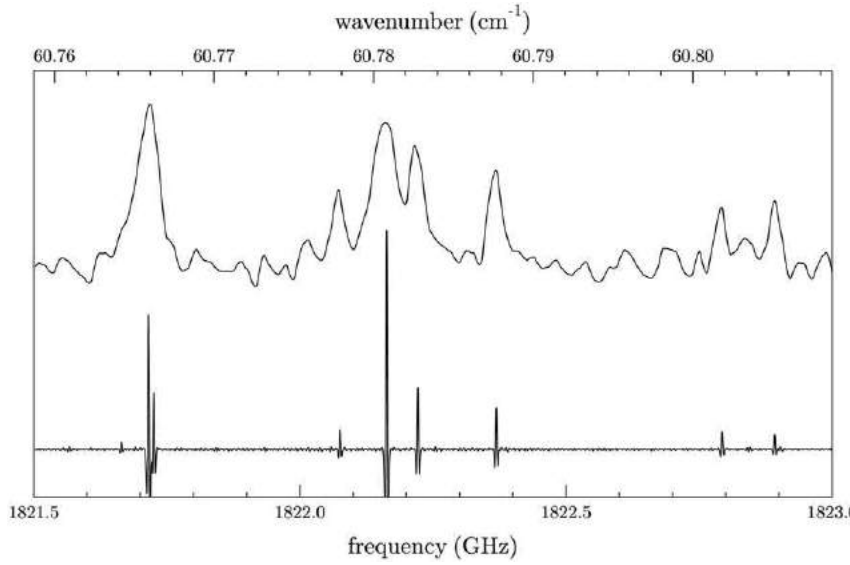


Fig. 4 Comparison of the room temperature THz spectrum of acrylonitrile recorded with microwave multiplication techniques (lower) (Kisiel *et al.*, 2012) and by using a Fourier transform interferometer with an 8.8 m optical path difference installed on the AILES beamline at the SOLEIL synchrotron source (upper) (Kisiel *et al.*, 2015).

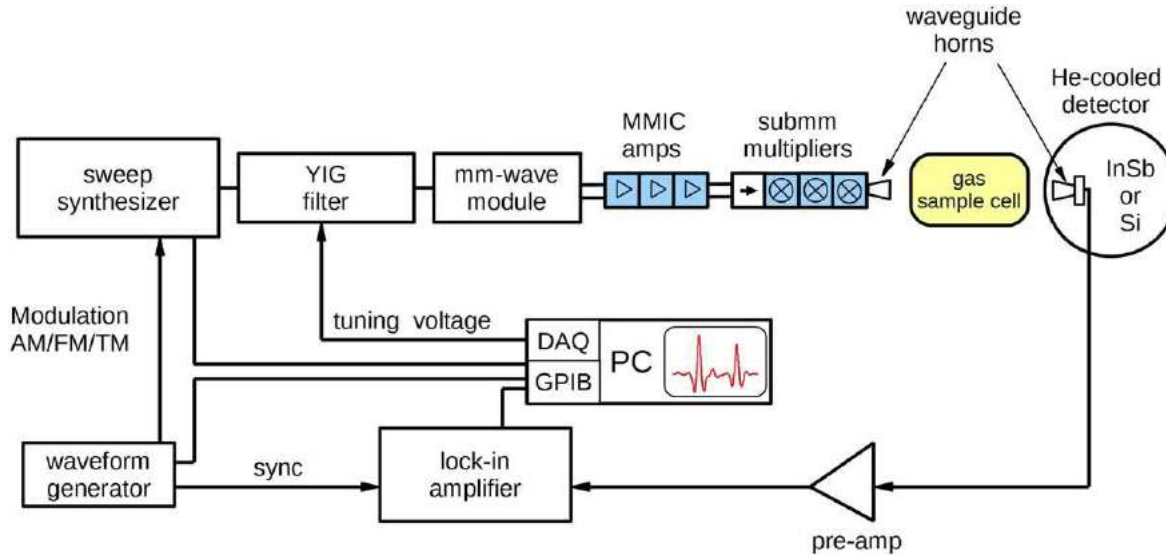


Fig. 5 Schematic diagram of the JPL cascaded harmonic multiplication THz spectrometer, utilizing various types of source modulation and several different detectors. Reproduced from Drouin, B.J., Maiwald, F.W., Pearson, J.C., 2005. Application of cascaded frequency multiplication to molecular spectroscopy. Review of Scientific Instruments 76, 093113 with permission.

Detection and the Cascaded Multiplication Spectrometer

The preferred detection method is to use liquid helium cooled bolometers, typically of the Indium Antimonide (InSb) type. A single detector of this type provides coverage of the complete THz region at the sensitivity and bandwidth ample for most studies. Room temperature, zero-bias GaAs Schottky diode detectors are also an attractive alternative for applications below 1 THz, where freedom from the need to use liquid helium offsets the somewhat lower sensitivity.

Detection of molecular signals is intimately associated with various stratagems aiming to increase signal to noise ratio and attempting to eliminate unwanted baseline variation. The most typical experimental setup is an absorption experiment, as illustrated in **Fig. 5**. THz radiation is passed through an absorption cell containing the gaseous molecular sample, and the transmitted signal is detected. Since molecular absorption coefficients are relatively small the first step is to increase absorption length, by using reasonably long sample cells, typically 1–3 m in length. The most common further technique is electronic source modulation.

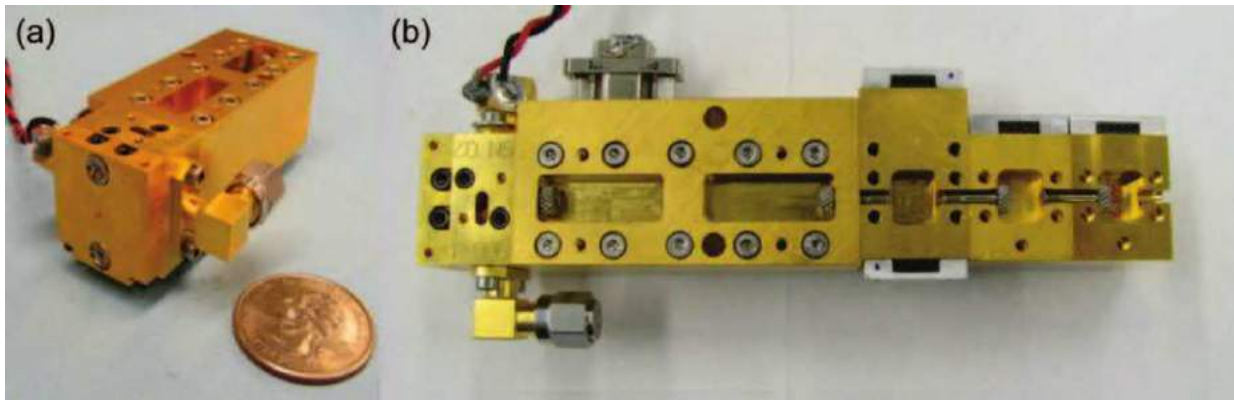


Fig. 6 Illustration of the small size of the 2.7 THz cascaded multiplication source: (a) the three final stage triplers with a 0.69×0.69 mm output port in the center of the output plate (b) the final 27x output multipliers with preceding Gaas and GaN simplifiers. Reproduced from Pearson, J.C., Drouin, B.J., Maestrini, A., *et al.*, 2011. Demonstration of a room temperature 2.48–27.5 GHz coherent spectroscopy source. Review of Scientific Instruments 82, 093105 with permission.

The frequency of an otherwise monochromatic source is modulated by an amount comparable to widths of absorption lines. In this way there will be a small alternating signal at the detector in the presence of a molecular absorption line and little or no signal in the absence of a line. A suitable type of phase sensitive detection will then provide separation from the large direct current detector signal proportional to the unabsorbed radiation, at the same time considerably improving the spectroscopic baseline. Two types of source modulation and of associated phase sensitive detection are possible. The first is standard frequency modulation (FM) and detection at twice the frequency f of the rate at which this modulation is applied. In this $2f$ detection the observed line profile is a second derivative of the natural line profile, as in the bottom trace in **Fig. 4**. This has the added bonus that the width of the central peak of the second derivative of a typical near-Lorentzian absorption profile is close to a third of width of the undifferentiated profile, introducing some useful resolution enhancement. The second source modulation method is to use toneburst modulation. In this case modulation is applied in an on-off mode determined by a square wave of frequency f and phase sensitive detection is also at this frequency. The resulting line profile is also second derivative in appearance, although it is made up of positive and negative lobes comparable in width to the natural waveform. Traditionally, toneburst modulation (TM) was somewhat easier to implement than the FM method for detection of broader THz region, although the distinction has largely disappeared. Finally, in high accuracy lineshape work the signal distortion introduced by source modulation is not acceptable so that amplitude modulation (AM), with either a mechanical chopper or by electronic means, is employed. A complete spectrometer design using a cascaded harmonic multiplication source is shown in **Fig. 5**. The miniature size of a THz cascaded multiplication source can be assessed from **Fig. 6**.

The performance of a source modulation cascaded multiplication spectrometer is critically dependent on the available THz power. The power profile of the source is particularly important at higher frequencies, where it results from the operation of many multiplication/amplification stages. The critical dependence of the actual spectral profile on the available power is shown in **Fig. 7**. The calculated methanol spectrum for the frequency region 2.45–2.75 THz has relatively uniform intensity of the strongest lines, confirming that the intensity profile of the spectrum in **Fig. 7** is an instrumental effect due to the source power characteristics shown in the top part of this figure. Nonetheless, **Fig. 7** covers a very considerable fragment of the molecular spectrum of methanol recorded at high resolution, while the much narrower 4 GHz width inset displays relative intensities of lines at close to prediction. At this smaller scale other instrumental effects on intensities can also be important. These are standing waves from reflections between the source and the detector, and from unavoidable reflections at all optical surfaces crossed by the radiation, especially the cell windows. If necessary, suitable techniques can be used to rescale the intensity of broadband spectra in order to remove the effect of the source power variation.

The Chirped Pulse THz Spectrometer

Broadband spectroscopy based on chirped pulse excitation has recently revolutionized microwave spectroscopy of supersonic expansion. In this method the sample is first excited by means of a pulse containing an internal frequency sweep and the resulting free induction decay (FID) signal emitted by the sample is then collected and Fourier transformed to give the frequency domain spectrum. High resolution time domain methods based on chirped pulse excitation have now been advanced to the millimeter wave and the THz regions. Generation of THz frequencies is by means of a multiplier chain and detection is by recording the time domain response signal with a fast digital oscilloscope, after downconversion with a similar multiplier chain. Chirped pulse molecular spectroscopy in the THz region has been demonstrated by Neill *et al.* (2013) as shown for methanol in **Fig. 8**. This spectrum was obtained in a record acquisition time of 58 μs, and results from 232 sequential measurements. Each measurement took 250 ns and consisted of a 25 ns excitation pulse, followed by 225 ns of acquisition. Each excitation pulse was a chirp covering

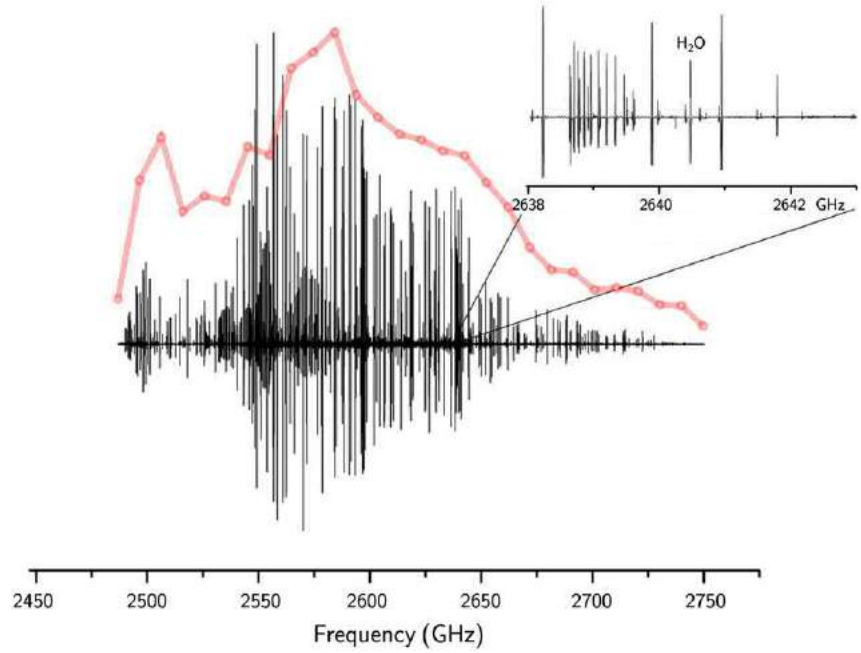


Fig. 7 Illustration of the broad, high resolution coverage available with a single cascaded frequency multiplication source. The spectrum in the lower panel is that of methanol at room temperature and spans 270 GHz at Doppler limited resolution, as shown in more detail in the 4 GHz width inset (reproduced from Pearson, J.C., Drouin, B.J., Yu, S., Gupta, H., 2011. Microwave spectroscopy of methanol between 2.48 and 2.77 THz. Journal of the Optical Society of America B 28, 2549–2577). The overall intensity profile is largely determined by the power performance of the source, with typical measured output in the 1–14 μW range plotted in red. The overall intensity profile is largely determined by the power performance of the source, with typical measured output in the 1–14 μW range plotted in red. Reproduced from Pearson, J.C., Drouin, B.J., Maestrini, A., *et al.*, 2011. Demonstration of a room temperature 2.48–27.5 GHz coherent spectroscopy source. Review of Scientific Instruments 82, 093105 with permission.

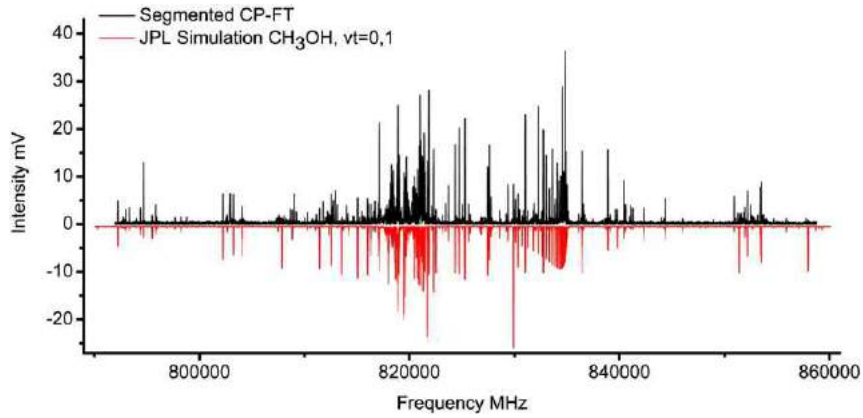


Fig. 8 Example of chirped pulse, Fourier transform access to the THz region. The spectrum is that of methanol and covers 67 GHz at a signal to noise ratio of 120:1. It results from Fourier transform of a single measurement cycle consisting of 232 free-induction decays recorded sequentially over a total 58 μs acquisition time. Reproduced from Neill, J.L., Harris, B.J., Steber, A.L., *et al.*, 2013. Segmented chirped-pulse Fourier transform submillimeter spectroscopy for broadband gas analysis. Optics Express 21, 19743–19749 with permission.

288 MHz in the frequency domain. This was obtained by generating a chirped pulse of 4 MHz bandwidth with programmed arbitrary waveform generator, and then subjecting it to 72 times frequency multiplication. The center frequency was ramped for successive pulses in order to provide continuous frequency coverage by means of the 232 FID segments. The technique shows considerable promise, even though some teething problems of spurious signals and additional instrumental line broadening are still to be ironed out. Comparison in **Fig. 8** between the experimental spectrum and the simulation shows that there is still considerable local variation in sensitivity, as apparent in the profile of the *Q*-branch near 835 GHz, or in the near disappearance of some predicted strong lines. Nevertheless, sufficient system stability for time-domain signal averaging has already been demonstrated and very rapid THz analytical applications appear feasible. Further refinement of such designs and broader adoption may

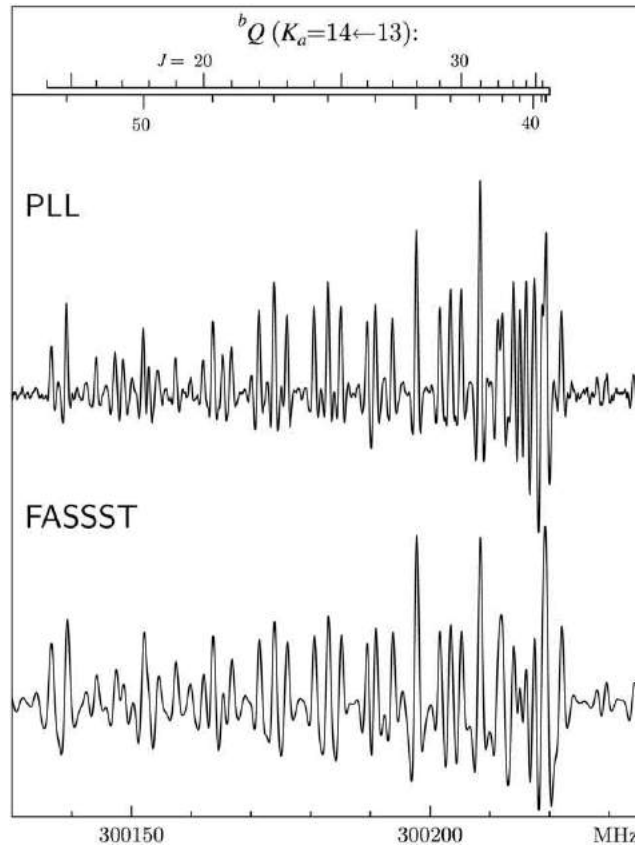


Fig. 9 Comparison of trans-gauche diethyl ether spectrum obtained in two different modes of using the BWO sources: under standard phase lock loop frequency control (PLL) or in the fast scanning mode calibrated with Fabry–Perot resonator fringes and SO_2 reference lines (FASSST). The FASSST method offered continuous access to many tens of GHz of the rotational spectrum at scan speeds exceeding 10 GHz s^{-1}, with only a small decrease in resolution. Reproduced from Medvedev, I., Winnewisser, M., De Lucia, F.C., *et al.*, 2004. The millimeter- and submillimeter-wave spectrum of the trans–gauche conformer of diethyl ether. Journal of Molecular Spectroscopy 228, 314–328.

be expected, while the ultimate limitation appears to be imposed by the decay rate of the free induction signal, found in the discussed work to be described by a time constant of 100 ns.

The FASSST Spectrometer

This is one of the spectrometer designs making use of the specific properties of high frequency BWO sources. It was recognized that these sources have very low noise on fast sweep operation so that sweep speeds exceeding 10 GHz s^{-1} were achievable without significantly compromising resolution. At such sweep speeds the source has to be operated in free running mode so that external, post-measurement frequency calibration was employed. This was achieved by means of fringes of a 15 m long folded Fabry-Perot resonator (9.5 MHz fringe spacing) and absolute calibration was transferred from a reference cell containing SO_2 gas. The fringe channel and the SO_2 spectrum were recorded simultaneously with the main spectrum. Considerable signal filtering is also required, as well as averaging of up and down scans in order to eliminate low level frequency nonlinearities. This spectrometer was developed at The Ohio State University, Columbus, United States (see Medvedev *et al.* (2004) and the references cited therein) and the authors coined the name "Fast Scan Submillimeter Spectroscopic Technique" (FASSST) which is the acronym by which it is known. The use of three different BWO sources allowed recording of practically continuous 110–370 GHz spectra for many molecules. Calibration against a more standard spectrometer employing a phase locked source revealed that the fast scanning mode introduced only a small penalty in resolution and frequency accuracy, see **Fig. 9**. While frequencies in the original FASSST spectrometer only overlap with the lower limit of the THz region, the technique was powerfully extended by multiplying the output of the 300 GHz source with a frequency tripler. Frequency calibration was carried out at the fundamental frequency, while frequency tripling allowed over 200 GHz of spectroscopic coverage around 1 THz.

The Broadband Resonator Spectrometer

Resonator spectrometers are a class of spectrometer in which the quality factor (Q) of an optical resonator is affected by the medium filling the resonator. The most robust measurement is that of the shape of a given resonator fringe, which yields precise

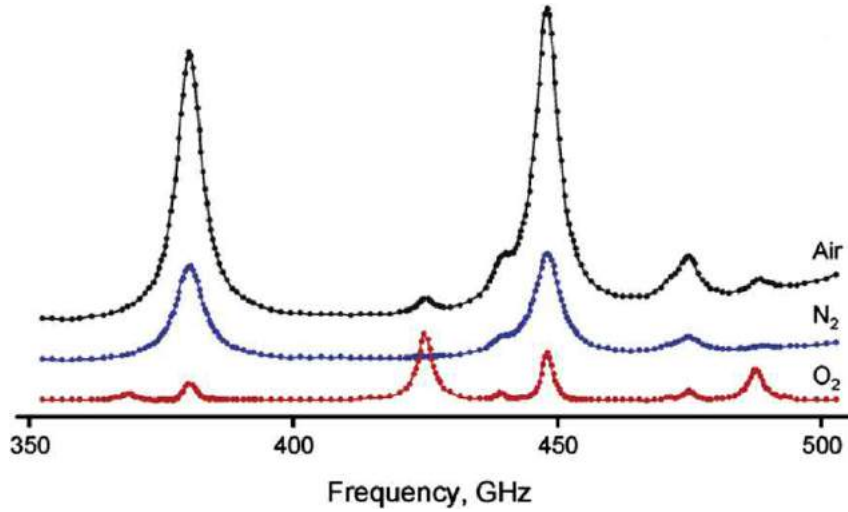


Fig. 10 Absorption spectra of the atmospheric gas composition and of individual atmospheric constituents at atmospheric pressure and with varying water vapor content (measured at 6.6%, 1.9% and near 0% relative humidity for air, N_2, and O_2, respectively). The spectra are composed of precise point by point absolute absorption measurements made with a single source of a BWO based resonator spectrometer. Reproduced from Krupnov, A.F., Tretyakov, M.Y., Belov, S.P., *et al.*, 2012. Accurate broadband rotational BWO-based spectroscopy. Journal of Molecular Spectroscopy 280, 110–118.

values of molecular absorption coefficients and broadening parameters. A powerful recent version of this type of spectrometer design is also based on the broad tunability and clean output power of high frequency BWOs (Krupnov *et al.*, 2012). In this case a phase locked BWO is used to make measurements on successive resonator fringes for a resonator of fixed physical size, without any need for moving any of the resonator mirrors. This technique allows measurement of gas phase rotational spectra in the important atmospheric pressure regime, **Fig. 10**, where the strongest lines in the three spectra are those of water. Nitrogen does not have any lines, but contributes to pressure broadening. The lines at 368.5, 424.8, and 487.2 GHz, which are visible only in the O_2 and in the air spectrum, are a spin triplet for the $N=3\leftarrow 1$ rotational transition of the oxygen molecule. Oxygen is a rare molecule in which its electronic ground state is a triplet state ($^3\Sigma$) due to the presence of two unpaired electrons. In such case, even though O_2 does not have a permanent electric dipole moment, it will have pure rotational transitions allowed by the magnetic dipole moment. The N quantum number accounts also for the electronic angular momentum, which is combined with the usual rotational angular momentum. The widths of molecular lines in **Fig. 10** are in the region of 5 GHz, and are four orders of magnitude greater than in the low pressure regime. This makes it clear why the standard signal modulation methods discussed above are not applicable and a specialized technique is necessary. The incentive for its development was that the atmospheric THz molecular spectrum is of particular applied importance as will be seen below in the discussion of astrophysical applications.

THz Lamb-Dip Spectroscopy

This is a method at the other extreme to that described immediately above, as it is a variant of molecular spectroscopy in which precision and resolution are enhanced over those available with the more routine techniques. Lamb-dip spectroscopy, or saturation spectroscopy, is usually associated with laser spectroscopy since it relies on the use of high radiation power density. In a longitudinal THz molecular absorption cell it is most convenient to achieve sufficient power density by a double pass arrangement, where a rooftop reflector is placed at the end of the cell, allowing detection to be at the cell entrance. In such case molecules with a zero Doppler velocity component relative to the probing radiation (moving perpendicular to the cell axis) will show preferentially decreased absorption at the transition frequency. Their lower energy levels participating in the transition will be rapidly depopulated and further photons arriving at this frequency will not be absorbed. At frequencies on the slopes of the Doppler absorption profile the absorption on the forward and backward pass will correspond to molecules moving with different velocity vectors so that depopulation by radiation will be much less efficient. In this way a frequency sweep will produce a narrow peak at the absorption maximum directed toward zero absorption (the Lamb-dip). In many cases high effective radiation power can also be set up through unavoidable Fabry–Perot type resonances in the absorption cell. For this reason observation of Lamb dips on absorption lines of water is a relatively common occurrence when the cell is pumped down, but trace water is still desorbing from the cell walls. Nevertheless, a well designed Lamb-dip experiment as described by Cazzoli and Puzzarini (2013) allows frequency measurement precision at the kHz level, as is implicit from the examples in **Fig. 11**. The conditions for successful saturation measurements require sample pressures of less than 1 mTorr and reduced frequency modulation widths comparable with the achievable sub 100 kHz linewidths of the Lamb-dip components. Such results have clear relevance to derivation of very precise spectroscopic constants, including those implicated in various types of small scale spectroscopic splitting, that is not resolvable by other means.

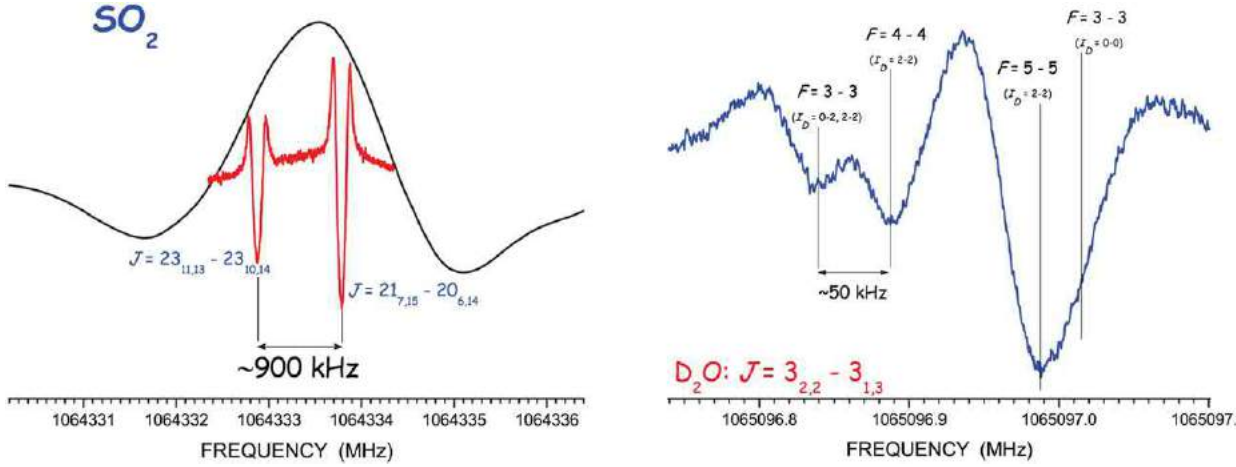


Fig. 11 Illustration of the enhanced resolution possible on application of the Lamb-dip technique in the THz region. The left panel shows how two transitions of SO_2 blended in the standard Doppler profile limited spectrum (black trace) can be completely resolved (red trace). The right panel demonstrates the resolution of components in a transition of D_2O that result from nuclear quadrupole hyperfine splitting due to the two deuterium nuclei. Reproduced from Cazzoli, G., Puzzarini, C., 2013. Sub-doppler resolution in the THz frequency domain: 1 kHz accuracy at 1 THz by exploiting the Lamb-dip technique. Journal of Physical Chemistry A 117, 13759–13766 with permission.

Table 3 THz region bands available at the ALMA observatory

Band number	*Frequency (GHz)*	*Wavelength (mm)*
7	275–373	0.804–1.091
8	385–500	0.600–0.779
9	602–720	0.417–0.498
10	787–950	0.316–0.381

Astrophysical Molecular Spectroscopy

In this case the emphasis is on detection. Extra-terrestrial THz spectra are mostly emission spectra resulting from excitation by distant sources, so that recording such spectra involves efficient collection of incoming radiation by means of suitable antennas and then extracting the spectra by heterodyne techniques. Precision in antenna construction is paramount since adhering to even the $\lambda/10$ rule of surface precision for a radio telescope dish of several meter diameter is challenging for sub-millimeter wavelengths. The premier radio telescope installation is currently the Atacama Large Millimeter Array (ALMA), which is operated on the Chajnantor plain in the Atacama desert in Chile, at an altitude of 5000 m. This consists of fifty movable 12-meter diameter antennas made to 25 μm surface precision, which can be configured into various baseline configurations for increased spatial resolution. Heterodyne detection is implemented with the use of liquid helium cooled downconversion "back ends" allowing simultaneous coverage of a signal bandwidth of up to 7.5 GHz. Detection is in the form of position/intensity/frequency "data cube", so that mapping of the spatial distribution of signals from several molecules is possible on the basis of a single observation.

ALMA makes available four frequency bands for THz spectroscopic studies, as listed in **Table 3**. Since the observatory is subject to atmospheric absorption its location was chosen in a particularly dry, high-altitude area in order to minimize the absorption from the atmospheric spectrum already shown in the laboratory version in **Fig. 10**. The actual atmospheric absorption at the ALMA site is reproduced in **Fig. 12** and it is immediately apparent how THz atmospheric molecular absorption lines affect the coverage chosen for the various bands. The main culprits in this case are water lines, as listed in **Fig. 12**, although the 60 GHz oxygen absorption also enforces the gap between bands B1 and B2. Many other molecular lines are also visible in **Fig. 12**, although those do not pose such a hindrance to observations. Those lines are mainly due to ozone, which is an important atmospheric constituent, and such lines are, for example, the 'grass' between 230 GHz and the 325 GHz water line, the multiplets around 655 GHz, and the multiplets around 835 GHz.

Although ALMA is currently the most capable ground based astrophysical THz observatory, it has had several ground based predecessors, and THz molecular spectroscopy has also been possible with satellite observatories. The advantage of satellite observatories is their immunity to atmospheric absorption, but the disadvantage has been relatively short operational lifetime due to limits on liquid helium coolant supply used to achieve the highest detection sensitivity. The best known THz capable satellite has been the Herschel Space Observatory, carrying a 3.5 m diameter collecting mirror and a heterodyne spectroscopy instrument

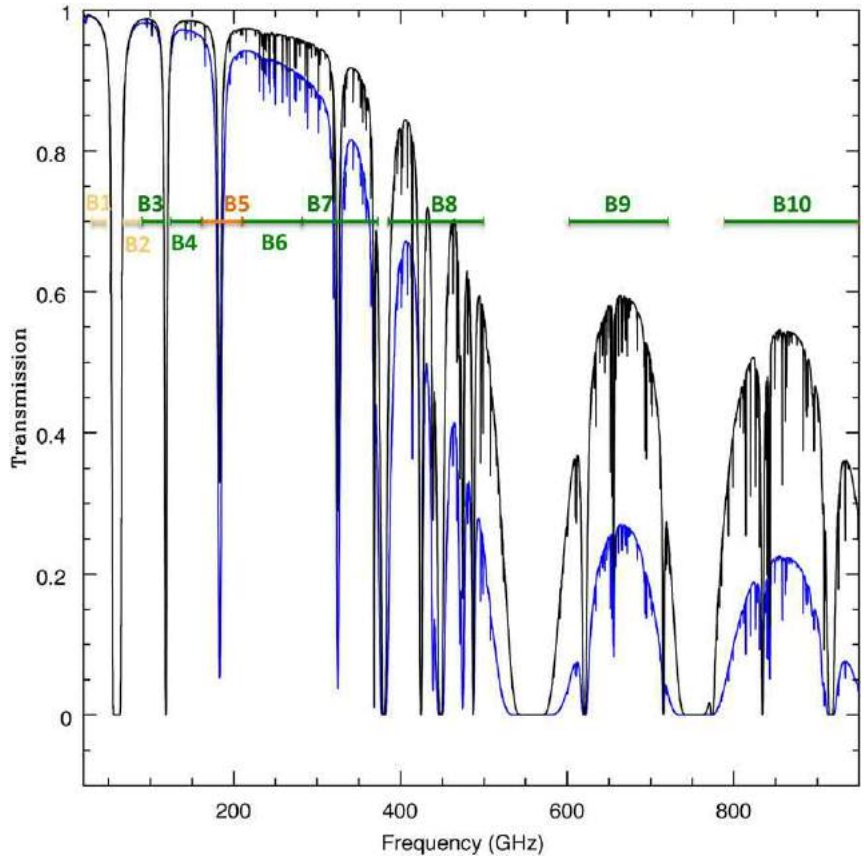


Fig. 12 Comparison of the band choice of the ALMA observatory with the atmospheric water spectrum recorded in its location and corresponding to 50% (blue) and 12.5% (black) of the available observing conditions. The water THz spectrum commences with the relatively benign $5_{14} \leftarrow 4_{22}$ line at 325.2 GHz, but it then contains the fully attenuating lines, $4_{13} \leftarrow 3_{21}$ at 380.2 GHz, $1_{10} \leftarrow 1_{01}$ at 556.9 GHz, and $2_{11} \leftarrow 2_{02}$ at 752 GHz, which enforce the three gaps in coverage between bands B7 to B10. Note that all of the atmospheric lines from the laboratory resonator spectrum shown in **Fig. 10** are visible. Reproduced from Schieven, G. (Ed.), 2016. Observing With ALMA – A Primer, ALMA Doc. 4.1, Ver. 3. Available at: https://almascience.eso.org/documents-and-tools/cycle4/alma-early-science-primer.

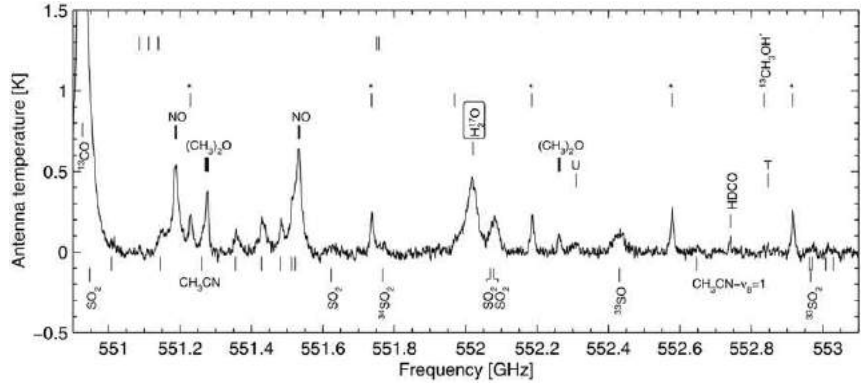


Fig. 13 A small excerpt from the Odin satellite spectral survey of the Orion KL interstellar molecular cloud core. Unlabeled markers placed at the intensity level of 0.9 K identify a *Q*-branch line sequence of CH_3OH in the first excited torsional state. These lines and the strong ^{13}CO and $H_2{}^{17}O$ lines indicate how increasing sensitivity of astrophysical spectra requires complete understanding also of transitions in rare isotopic species, and in excited vibrational states of many molecules. Reproduced from Olofsson, A.O.H., Persson, C.M., Koning, N., *et al.*, 2007. A spectral line survey of Orion KL in the bands 486–492 and 541–577 GHz with the Odin satellite I. The observational data. Astronomy and Astrophysics 476, 791–806 with permission.

(HIFI) covering 480–1280 GHz and 1419–1910 GHz. Herschel was operated from 2009 to 2013, when its 2300 liters of onboard helium coolant ran out. There have been many scientific results from Herschel, even after its lifetime, such as identification of water vapor on the dwarf planet Ceres.

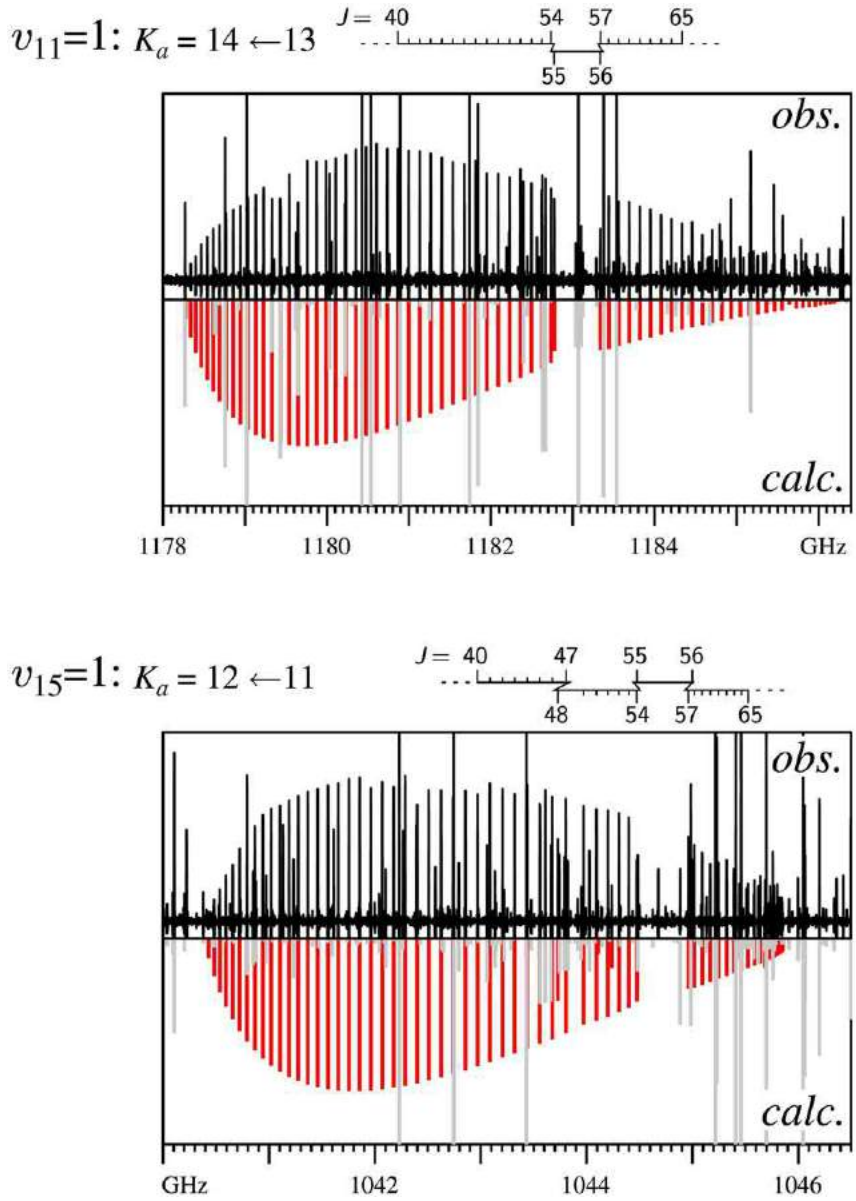


Fig. 14 Example of clearly discernible perturbations in the THz spectrum of acrylonitrile. The striking gaps in ^{b}Q-branch bands in two different vibrational states, $v_{11}=1$ (upper) and $v_{15}=1$ (lower) are the result of matching resonance interactions, which take place for specific values of J between 54 and 57. Such interactions are successfully accounted for in the complete quantum mechanical treatment of the problem, as shown in the calculated spectrum. Reproduced from Kisiel, Z., Pszczółkowski, L., Drouin, B.J., *et al.*, 2012. Broadband rotational spectroscopy of acrylonitrile: Vibrational energies from perturbations. Journal of Molecular Spectroscopy 280, 134–144.

A preview of the rich astrophysical molecular spectroscopy in the THz is visible in Fig. 13, which displays only a small fragment of a spectroscopic survey of the diffuse molecular cloud in the Orion nebula. This result comes from the Odin satellite, equipped with a moderate 1.1 m diameter telescope operating at 486–580 GHz. It is noteworthy that understanding this spectrum requires the knowledge not only of the main molecular species, but also of their, sometimes rare, isotopes (^{17}O has a terrestrial abundance of 0.04%) and of transitions in excited vibrational states. In the spectrum in Fig. 13 this concerns relatively few molecules, but more sensitive observational instruments such as ALMA require such information for a large number of molecules.

Influence of the THz Region on Molecular Spectroscopy

The advances in experimental techniques, which currently allow broadband access to high resolution molecular spectra in the THz region, have created new challenges for spectroscopists. The first is simply the need for efficient handling of spectra as illustrated by

the fact that high-resolution coverage of the whole THz region at 0.1 MHz point spacing requires 27 million data points and such spectrum will contain well in excess of one hundred thousand lines. There is, therefore, the need to efficiently navigate very long spectra, to compare spectra with predictions, and to set up data sets for programs fitting spectroscopic constants in a Hamiltonian suitable for the problem at hand. Various computer programs have been developed to meet this challenge and those are generally available for use by the spectroscopic community on specialized websites.

Measurement of THz molecular spectra usually took place as an extension of previous lower frequency studies of a given molecule. The primary advance brought in by the THz data was increased numerical precision in accounting for transition frequencies of considerably greater numbers of measured lines. But there have also been cases where THz measurements allowed access to qualitatively new molecular information, which is was not revealed at lower frequencies. One such example is for the already discussed acrylonitrile molecule and is illustrated in **Fig. 14**. The rotational spectrum of acrylonitrile has been subjected to many preceding spectroscopic studies, but it was only when transitions around 1 THz were measured that several intriguing features became apparent. It does not take special insight to notice that there are unusual gaps in the otherwise uniform progressions of *Q*-branch transitions in the two lowest excited vibrational states of the molecule. The pertinent transitions are not missing but are shifted due to interactions resulting from near coincidence in energy of rotational levels in the two vibrational states (Kisiel *et al.*, 2012). Similar, but less spectacular, interactions were observed between the ground state and the $v_{11}=1$ vibrational state, also at frequencies close to 1 THz, but for values of the *J* quantum number exceeding 100 (Kisiel *et al.*, 2009). A suitable quantum mechanical analysis of such interactions allowed precise determination of the vibrational frequencies of the two lowest normal modes of acrylonitrile: $v_{11}=228.29986(2)$ and $v_{15}=332.67811(2)$ cm^{-1}, even though the actual vibrational transitions from the ground state fall outside the THz region. This analysis was later confirmed by synchrotron based high resolution rotation vibration spectroscopy (Kisiel *et al.*, 2015).

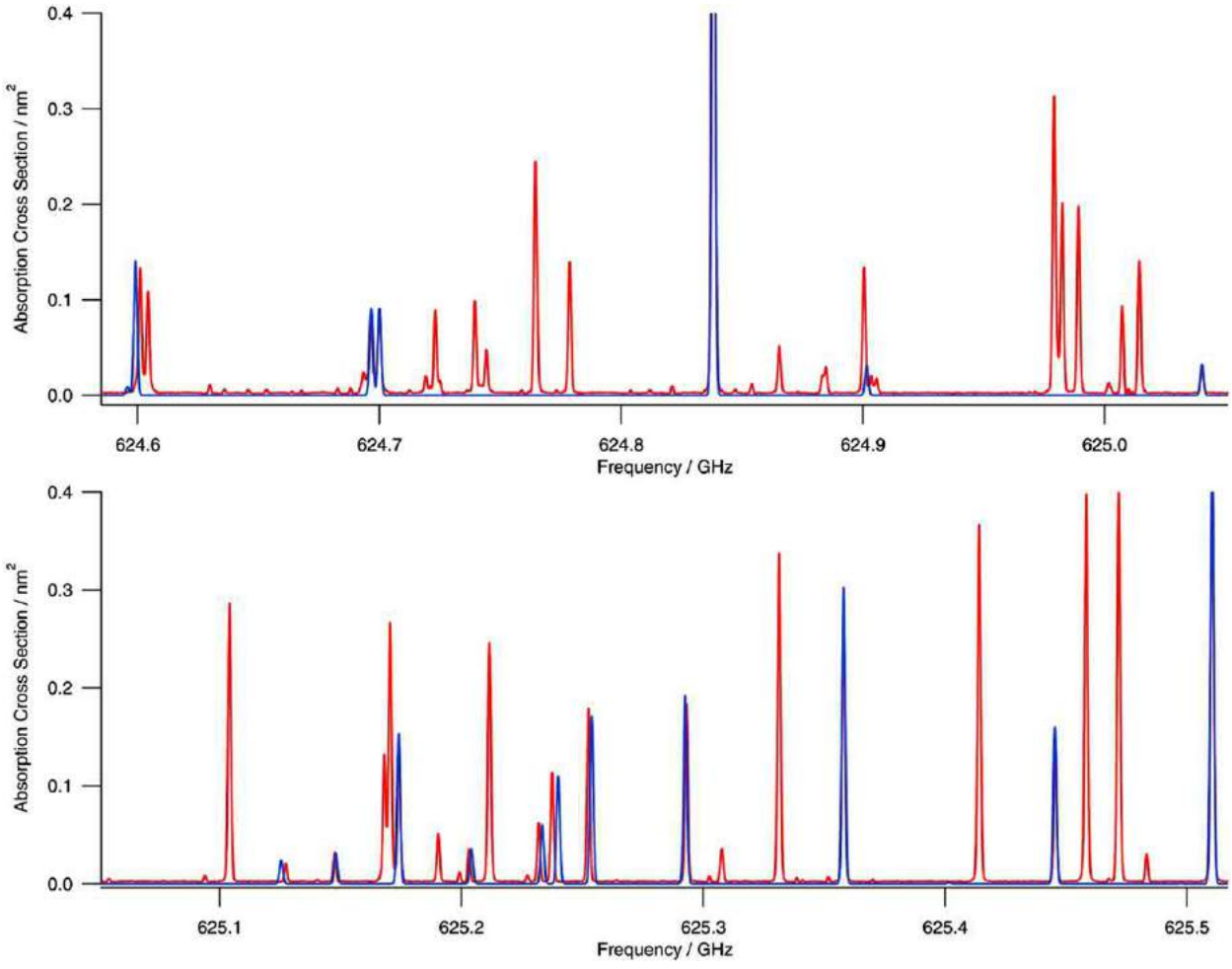


Fig. 15 Comparison of the experimentally derived, absolute intensity calibrated, 300 K spectrum of methanol (red) with prediction from the available catalog data (blue). The experimental trace is a point by point simulation based on recording 166 experimental spectra over the 248–398 K temperature range. The number of inconsistencies between the two traces indicates that considerable progress still has to be made in understanding this complex spectrum. Reproduced from Fortman, S.M., Neese, C.F., De Lucia, F.C., 2014. The complete, temperature resolved experimental spectrum of methanol (CH_3OH) between 560 and 654 GHz. Astrophysical Journal 782, 1–8 with permission.

Molecular Spectroscopy Databases and Their Verification

The molecules with extensive THz spectra, which are of greatest applied interest, such as water and methanol, are some of the most challenging molecules for quantum mechanics. The extensive coverage of their spectra that is now possible and the need to reproduce the experimental frequencies to within their experimental uncertainty require very specialized techniques of constructing and fitting quantum mechanical Hamiltonians. The number of parameters of fit becomes considerable and a ratio of the number of fitted lines to the number of fitted parameters exceeding 100 is considered satisfactory. This means that for data sets reaching tens of thousands of lines the number of fitted parameters runs into hundreds.

At the same time most users of such data, such as the astrophysical community, are not familiar with the spectroscopic details but simply require line lists of frequencies and temperature scalable intensities. In order to cater for this demand several very useful databases have been established. These are the JPL Catalog, the Cologne Database for Molecular Spectroscopy (CDMS), and the High-resolution transmission molecular absorption database (HITRAN). The first two databases are oriented more towards the astrophysical user, while the development of HITRAN is geared to the needs of atmospheric spectroscopy.

The databases are continuously updated as new experimental data is acquired, but their coverage of a given molecule is still far from complete. The complexity of some spectroscopic problems, especially of the spectrum of methanol, is such that transition frequencies that are not among the fitted data, but are only predicted, may not yet be sufficiently reliable. For this reason dedicated measurements have been undertaken with the specific aim of database verification, both in transition frequency and in absolute intensity. This was carried out for several molecules of astrophysical importance by recording their laboratory spectra at multiple temperatures. Suitable processing allowed derivation of an absolute intensity spectrum for any desired temperature, and the result for the methanol spectrum at 0.6 THz is shown in **Fig. 15**. Comparison with the data available in the JPL and CDMS catalogs (see Relevant Websites section) reveals that there are still significant gaps in the coverage of transitions and that there is still room for improvement in the (frequency, intensity) parameters of some of the catalog lines.

Conclusions

Relatively routine experimental access to the THz region and the resulting opportunities for high-resolution molecular spectroscopy allowed considerable advances in spectroscopy of several important molecules. The astrophysical community is an important consumer of such data and it has stimulated much laboratory work, largely in an attempt to characterize the spectroscopic 'weeds' that hinder identification of new molecules. At the same time significant improvement in characterization of rotational and vibrational properties of several key molecules has been reached, although the abundance of spectroscopic data shows that there are still significant gaps in identification of observed lines and there is still much further work ahead.

References

Cazzoli, G., Puzzarini, C., 2013. Sub-doppler resolution in the THz frequency domain: 1 kHz accuracy at 1 THz by exploiting the Lamb-dip technique. Journal of Physical Chemistry A 117, 13759–13766.

Drouin, B.J., Maiwald, F.W., Pearson, J.C., 2005. Application of cascaded frequency multiplication to molecular spectroscopy. Review of Scientific Instruments 76, 093113.1–10.

Kisiel, Z., Martin-Drumel, M.-A., Pirali, O., 2015. Lowest vibrational states of acrylonitrile from microwave and synchrotron radiation spectra. Journal of Molecular Spectroscopy 315, 83–91.

Kisiel, Z., Pszczółkowski, L., Drouin, B.J., *et al*., 2009. The rotational spectrum of acrylonitrile up to 1.67 THz. Journal of Molecular Spectroscopy 258, 26–34.

Kisiel, Z., Pszczółkowski, L., Drouin, B.J., *et al*., 2012. Broadband rotational spectroscopy of acrylonitrile: Vibrational energies from perturbations. Journal of Molecular Spectroscopy 280, 134–144.

Krupnov, A.F., Tretyakov, M.Y., Belov, S.P., *et al*., 2012. Accurate broadband rotational BWO-based spectroscopy. Journal of Molecular Spectroscopy 280, 110–118.

Medvedev, I., Winnewisser, M., De Lucia, F.C., *et al*., 2004. The millimeter- and submillimeter-wave spectrum of the trans–gauche conformer of diethyl ether. Journal of Molecular Spectroscopy 228, 314–328.

Neill, J.L., Harris, B.J., Steber, A.L., *et al*., 2013. Segmented chirped-pulse Fourier transform submillimeter spectroscopy for broadband gas analysis. Optics Express 21, 19743–19749.

Rao, K.N. (Ed.), 1976. Molecular Spectroscopy: Modern Research, vol. II. New York: Academic Press. (Chapter 2).

Schlemmer, S., Giesen, T., Lewen, F., Winnewisser, G., 2009. High-resolution laboratory terahertz spectroscopy and applications to astrophysics. In: Laane, J. (Ed.), Frontiers of Molecular Spectroscopy. Amsterdam: Elsevier, pp. 241–265.

Relevant Websites

http://www.almaobservatory.org/
ALMA.

http://www.almascience.org/
ALMA.

http://www.astro.uni-koeln.de/cdms/
CDMS – Cologne Database for Molecular Spectroscopy.

http://www.istokmw.ru/vakuumnie-generatori-maloy-moshnosti/
Examples of THz region BWO tubes produced by Istok Company, Russia.
http://vadiodes.com/en/products/custom-transmitters
Examples of THz transmitter modules produce by Virginia Diode Inc., USA.
http://www.esa.int/SPECIALS/Herschel
Herschel.
https://www.cfa.harvard.edu/hitran/
High-resolution transmission molecular absorption database, HITRAN.
http://hitran.iao.ru/
HITRAN on the web.
https://spec.jpl.nasa.gov/
JPL Molecular Spectroscopy, home of the JPL Catalog and of the SPFIT/SPCAT program suite.
http://www.mwl.sci-nnov.ru/2016.html
Microwave Spectroscopy Laboratory at Nizhny Novgorod.
http://www.nasa.gov/herschel
NASA.
http://www.snsb.se/en/Home/Space-Activities-in-Sweden/Satellites/Odin/
Odin satellite.
http://info.ifpan.edu.pl/~kisiel/prospe.htm-PROSPE
Programs for Rotational SPEctroscopy.

Broadband Terahertz Sources

Kang Liu, University of Rochester, Rochester, NY, United States
Xi-Cheng Zhang, University of Rochester, Rochester, NY, United States; ITMO University, Saint-Petersburg, Russia; and Capital Normal University, Beijing, China

Photo-Induced Carriers in Solid Materials

Photoconductive Antenna

Broadband Terahertz (THz) pulses can be generated by a biased photoconductive (PC) antenna excited with femtosecond optical pulses. The PC antenna is one of the most widely used components for THz generation and detection. **Fig. 1(a)** gives a schematic demonstration of a typical THz generation PC antenna structure. It is composed of two metal electrodes deposited on a semiconductor substrate. When generating THz radiation, a DC voltage is applied across the electrodes. The femtosecond optical pulses which have photon energy higher than the semiconductor energy band gap will radiate the area between the electrodes and generate photo-induced free carriers in the substrate. Sometimes, free carriers can also be formed using excitation pulses with photon energy lower than the semiconductor band gap, due to multi-photon absorption in the substrate. Once the free carriers are formed, they will be accelerated by the electric field between the electrodes, and a photocurrent J is produced. As the photocurrent varies with time, it radiates electromagnetic wave pulses with an electric field E_{THz} at the far field that can be express as (Zhang and Xu, 2010)

$$E_{THz} \propto \frac{\partial J(t)}{\partial t} \tag{1}$$

Fig. 1(b) plots the calculated results of a photocurrent induced by an optical pulse and its corresponding THz far-field waveform with typical material and physical parameters, which should give a straightforward intuition regarding the relation between the optical pulse, the photocurrent, and the THz radiation electric field. The emitted THz electric field has a polarization parallel to the direction of the bias field, which can be flipped by switching the direction of the bias field.

The power and bandwidth of the THz emission from PC antenna largely depend on the structure of the electrodes (Tani *et al.*, 1997). **Fig. 2** shows the temporal and spectral amplitude of THz radiation from three different electrode structures: dipole, bow tie and strip line, deposited on low-temperature grown GaAs. The THz pulse from the strip line structure has a shorter pulse duration, corresponding to a spectrum extending up to 4 THz with a FWHM of about 1.3 THz.

The output power of a PC emitter also depends on the bias voltage and the optical pump power. When the optical pump power is low and the bias field is weak, the THz radiation field amplitude increases linearly with these two parameters. However, increasing the bias voltage has a limitation because a high electric field may result in the breakdown of the substrate material. For example, the breakdown field of LT-GaAs is reported as ~300 kV/cm (Tani *et al.*, 1997). In addition, the THz radiation output power will saturate as the optical pump power reaches a certain level, for the screening effect of the bias field by the photo-induced carriers.

Ever since the first introduction of THz generation by PC antenna in the 1980s (Auston *et al.*, 1984a), this technique has undergone dramatic developments, due to the soaring amount of research on the THz-time-domain-spectroscopy techniques. Up to now, the strongest broadband THz pulse energy emitted by PC antenna was reported to be 8.3 μJ with a peak electric

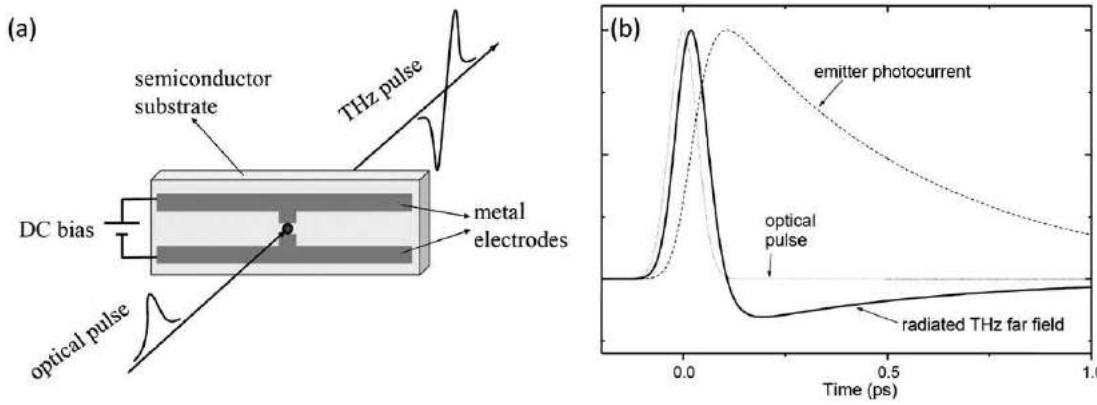


Fig. 1 (a) Schematic diagram of THz pulse emission from a PC antenna excited by a femtosecond laser pulse and (b) calculated photocurrent (dashed line) vs. time. Reprinted from Lee, Y.S. 2010. Principles of Terahertz Science and Technology. New York: Springer, p. 61.

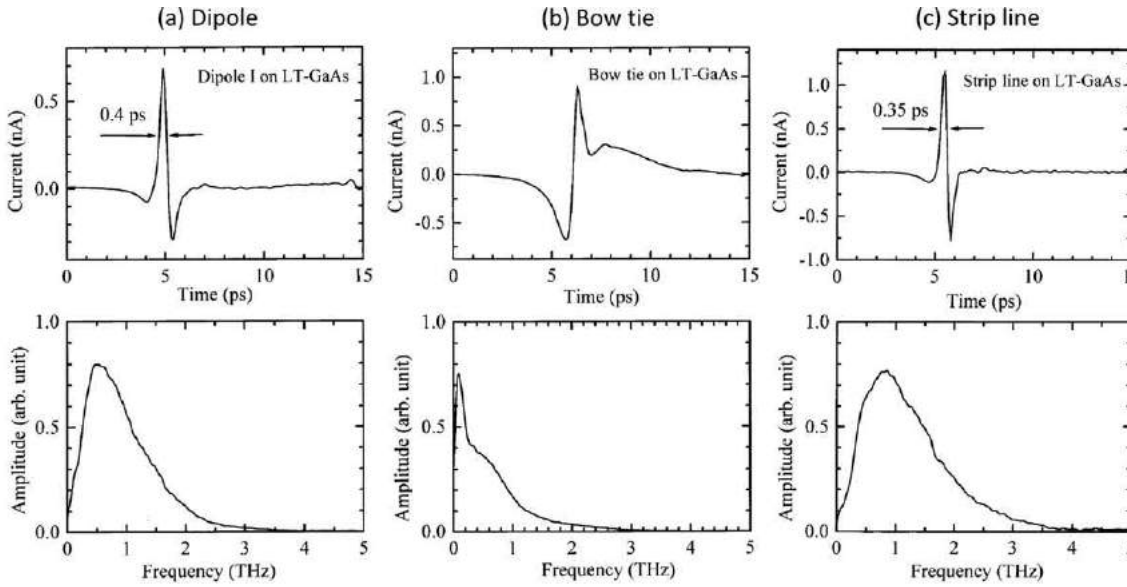


Fig. 2 THz radiation pulse shapes and amplitude spectra from PC emitters with (a) dipole, (b) bow tie and (c) strip line. Reprinted from Tani, M., Matsuura, S., Sakai, K., Nakashima, S.-i., 1997. Emission characteristics of photoconductive antennas based on low-temperature-grown Gaas and semi-insulating GaAs. Applied Optics 36, 7853–7859.

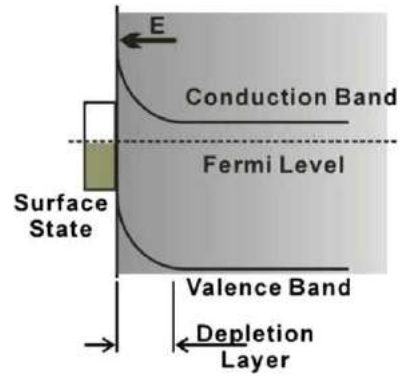


Fig. 3 Schematic demonstration of band bending and surface field of a n-type GaAs wafer. Reprinted from Zhang, X.-C., Xu, J., 2010. Introduction to THz Wave Photonics. New York: Springer, p. 32.

field of 331 kV/cm, achieved with an encapsulated interdigitated ZnSe Large Aperture Photo-Conductive Antenna (LAPCA) (Ropagnol *et al.*, 2016). A 20 THz broadband generation using semi-insulating GaAs interdigitated photoconductive antennae was also reported in 2014 by Hale *et al.*, (2014).

Built-in Field in Semiconductor

THz generation from the surface of semiconductor illuminated by femtosecond laser was reported byZhang *et al.* (1990). This generation takes advantage of the surface built-in field of semiconductors, such as GaAs (Zhang and Xu, 2010). The Fermi level of the surface state at the interface between the material and the air is different from the Fermi level of the bulk material, which leads to the semiconductor band bending close to the interface. In this area, which is also called the depletion layer, forms a built-in surface field that causes the carriers to drift toward the inside of the bulk material and eventually a balance is reached between the drift and the diffusion of the free carriers. **Fig. 3** shows the band bending and surface field in an n-type GaAs wafer. When a femtosecond laser pulse excites the semiconductor, photo-induced carriers will start drifting again under the built-in electric field, and therefore emit THz radiation just like what happens in the PC antenna. Since the surface field of a n-type GaAs is in opposite direction with a p-type GaAs, the THz pulse generated from p-type GaAs will have an opposite polarity compared to the one from n-type GaAs. The THz radiation amplitude depends on the optical pump pulse incident angle. The amplitude reaches the maximum as the angle approaches the Brewster angle.

Photo-Dember Effect

With the semiconductor materials that have small or no built-in field effect, THz radiation can still be generated due to photo-Dember effect (Zhang and Xu, 2010). When a semiconductor material surface is illuminated by a femtosecond laser pulse with photon energy higher than the material band gap, a large amount of free electron hole pairs are generated with an inhomogeneous distribution. The asymmetric distribution of the carriers on the surface causes them to diffuse toward the inside of the bulk material. Due to the fact that electrons and holes have different mobility, an ultrafast spatial separation between the electrons and holes is formed, and results in the radiation of THz pulses. InAs has been considered as an important semiconductor THz emitter due to its high electron mobility. When the THz generation process from InAs was studied, it was found that both the THz radiation from n-type and p-type InAs have the same polarity, identical to the THz pulse polarity from n-type GaAs, which cannot be explained by the mechanism of built-in surface field THz generation, but can be addressed by photo-Dember effect (Gu *et al.*, 2002). Actually, both effects exist in semiconductor materials when excited by ultrafast laser pulses. But the dominating effect depends on the semiconductor band gap structure and the property of the laser pulses.

Accelerated Free Electrons

Free-Electron Lasers

The free-electron lasers are outstanding broadband THz sources featuring ultra intense THz radiation brightness, normally several orders of magnitude higher than the laser-driven table-top THz sources. It has broad applications from chemical and biological imaging to driving nonlinear effects in the matter with THz radiation.

A relativistic electron bunch under acceleration emits cone-shaped radiation pattern in the direction of the electrons' velocity (Lee, 2010). There are many different ways to generate radiation by accelerating/decelerating the electron beam: bending the beam in a strong magnetic field (coherent synchrotron radiation and coherent edge radiation, CSR and CER), passing the beam through a small aperture (coherent diffraction radiation, CDR), or passing the beam from vacuum into a medium (coherent transition radiation, CTR, or *bremsstrahlung*) (Wu *et al.*, 2013). **Fig. 4** presents a schematic demonstration of the coherent THz radiation from a linear particle accelerator (LINAC). An ultrafast femtosecond pulse releases an ultrashort electron bunch from an electron source, which normally is either a photocathode electron gun or a semiconductor surface. In the accelerator, the electron bunch gains relativistic energy, 10–100 MeV, and the THz radiation can be generated through smashing the electron bunch onto a metal target to rapidly decelerate the electrons, or bending the electron trajectory by a magnetic field.

Many accelerator-based THz source facilities have been constructed worldwide, including the energy-recovered linac (ERL) in the Jefferson Laboratory (Carr *et al.*, 2002), the Source Development Lab (SDL) of the Brookhaven National Laboratory (Shen *et al.*, 2007), the Linac Coherent Light Source (LCLS) and the Facility for Advanced Accelerator Experimental Tests (FACET) of the SLAC National Accelerator Laboratory (Wu *et al.*, 2013), Berliner Elektronenspeicherring-Gesellschaft für Synchrotronstrahlung (BESSY) and Deutsches Elektronen Synchrotron (DESY) in Germany (Abo-Bakr *et al.*, 2002). The strongest coherent broadband THz radiation electric field from accelerator-based facilities has been generated by the SLAC National Accelerator Laboratory. The peak electric field at a THz focus has reached 4.4 GV m^{-1} (0.44 V $Å^{-1}$) (Wu *et al.*, 2013). Notice that an electric field in the V $Å^{-1}$ regime is comparable to the dipole field of an ionic bond, and can lead to many research topics involving the interaction between material and extreme THz radiation (Dhillon *et al.*, 2017).

Nonlinear Crystals

Optical Rectification

Optical Rectification (OR) is a second order nonlinear effect that is widely exploited for broadband THz pulse generation. Compared with PC antennae, OR crystals do not require application of high voltage on the material, making it one of the most

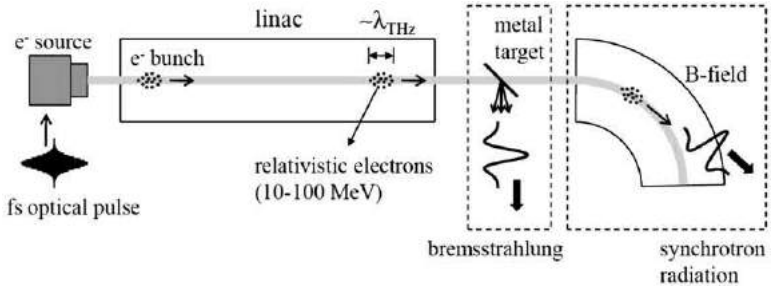


Fig. 4 Coherent THz radiation from relativistic electrons in a linear accelerator (linac). Reprinted from Lee, Y.S., 2010. Principles of Terahertz Science and Technology. New York: Springer, p. 103.

efficient and convenient intense coherent broadband table-top THz sources (Hafez *et al.*, 2016; Auston *et al.*, 1984b; Hu *et al.*, 1990).

When an intense optical beam passes through a nonlinear material, electric polarization P induced in the material can be expressed as a power series in the field strength E of the optical beam (Boyd, 2008):

$$P = \varepsilon_0(\chi^{(1)}E + \chi^{(2)}EE + \chi^{(3)}EEE + \chi^{(4)}EEEE + \ldots) \tag{2}$$

where ϵ_0 is the permittivity of free space, $\chi^{(1)}$ is the linear susceptibility, the quantities $\chi^{(2)}$ and $\chi^{(3)}$ are known as the second- and third- order nonlinear optical susceptibilities respectively. Among the different orders of nonlinear polarizations, the second order polarization components include:

$$\begin{aligned} P(2\omega_1) &= \varepsilon_0\chi^{(2)}E_1^2(SHG) \\ P(2\omega_2) &= \varepsilon_0\chi^{(2)}E_2^2(SHG) \\ P(\omega_1+\omega_2) &= 2\varepsilon_0\chi^{(2)}E_1E_2(SFG) \\ P(\omega_1-\omega_2) &= 2\varepsilon_0\chi^{(2)}E_1E_2^*(DFG) \\ P(0) &= 2\varepsilon_0\chi^{(2)}\left(E_1E_1^* + E_2E_2^*\right)(OR) \end{aligned} \tag{3}$$

ω_1,ω_2, E_1 and E_2 are respectively the frequencies and the electric fields of the two distinct frequency components of the incident beam. Here different polarization components are labeled by their corresponding physical process, such as second-harmonic generation (SHG), sum-frequency generation (SFG), difference-frequency generation (DFG), and optical rectification (OR) that we shall focus on in this section.

In essence, OR is the generation of a quasi-DC polarization in a nonlinear material when it has been passed through by an intense optical beam (see **Fig. 5(e)**), which results in a DC electric field proportional to the beam intensity. It can be understood as a second order phenomenon that is the reverse process of the electro-optic effect (Dexheimer, 2007). If the light beam that induces OR is a pulse instead of a continuous wave, the electric field generated by OR will become a time varying function related with the pulse envelope, thus radiate electromagnetic waves (Zhang and Xu, 2010). The far field radiation electric field should be proportional to the second derivative of OR induced time varying electric field:

$$E_{THz}(t) \propto \frac{\partial^2 P(0,t)}{\partial t^2} = \chi^{(2)}\frac{\partial^2 I(t)}{\partial t^2} \tag{4}$$

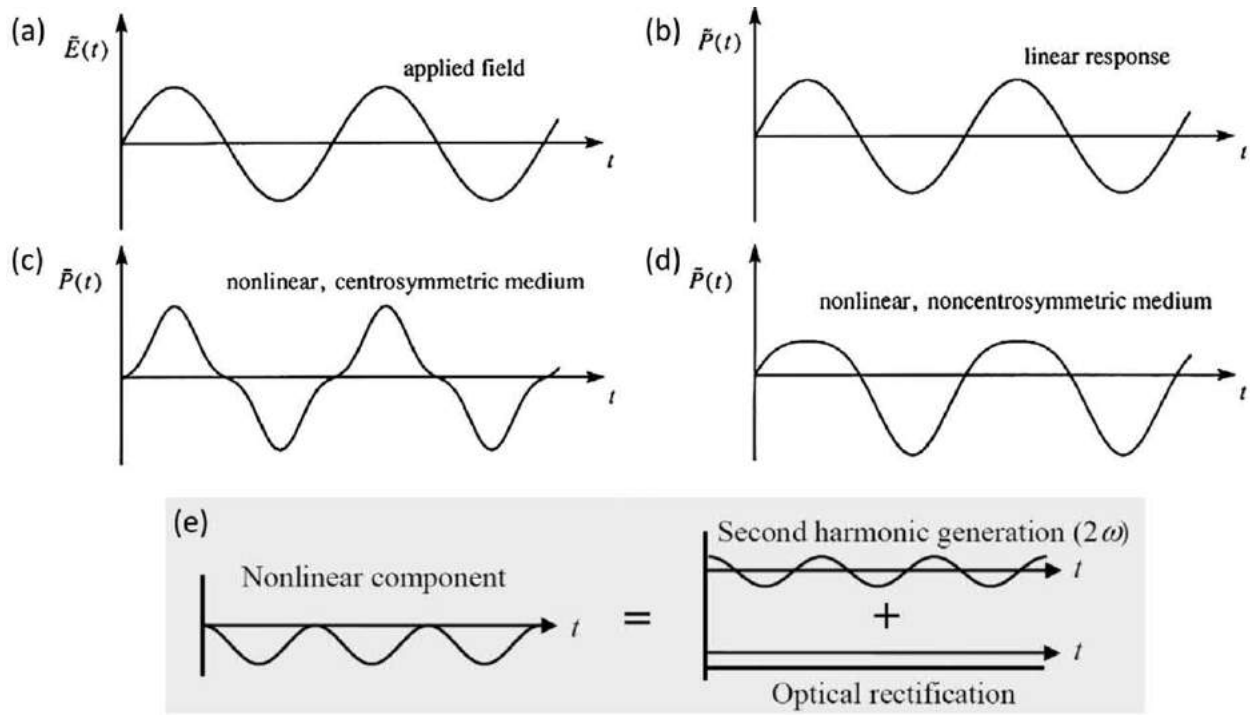


Fig. 5 (a–d) Waveforms associated with the atomic response (Reprinted from Boyd, R.W., 2010. Nonlinear Optics. Elsevier Science, p. 44) and (e) the decomposition of nonlinear response in a noncentrosymmetric medium. Reprinted from Lee, Y.S., 2010. Principles of Terahertz Science and Technology. New York: Springer, p. 76.

Here, $I(t)$ is the time dependent optical beam intensity. When the optical pulse duration is in picosecond or sub-picosecond level, the radiation frequency is within THz range.

Due to the fact that the second order nonlinear susceptibility $\chi^{(2)}$ vanishes in the centrosymmetric medium (Boyd, 2008), OR requires the material to have a noncentrosymmetric structure, such as the zinc-blende structure (see **Fig. 6**). The typical zinc-blende crystals that have been used in THz generations and detections include ZnTe, GaAs, GaP, etc.

Other than the susceptibility that is based on the crystal structure, the THz radiation efficiency, waveform and bandwidth depend on many factors, including laser beam properties, phase matching conditions, the crystal thickness, orientation, absorption and dispersion, etc.

In a nonlinear process such as optical rectification, one of the most crucial factors is the phase matching condition, which requires the energy and momentum conservation of the participating electromagnetic waves. For OR THz generation process, it can be described as follows:

$$\begin{cases} \omega_{O1} - \omega_{O2} = \Omega_{\mathrm{THz}} \\ k_{O1} - k_{O2} = k_{THz} \end{cases} \tag{5}$$

where ω_{O1} and ω_{O2} are the two frequency components that participate in the OR, k_{O1} and k_{O2} are their corresponding wave vectors. If we divide the first equation with the second one, we get:

$$\frac{\partial \omega_O}{\partial k_O} = \frac{\Omega_{THz}}{k_{THz}} \tag{6}$$

$$v_{G,O} = v_{Ph,THz} \tag{7}$$

where $v_{G,O}$ is the group velocity of the optical pulse, and $v_{Ph,THz}$ is the phase velocity of THz. Simply put, when the optical pulse group velocity equals the phase velocity of THz wave, the phase matching condition of OR THz generation is satisfied. That is why this condition is also called velocity-matching condition.

ZnTe is the most popular crystal for THz generation, because in this material the group velocity of femtosecond laser pulse around 800 nm (Ti: Sapphire laser center wavelength) matches very well with phase velocity of THz wave. For the optical wavelength $\lambda = 812$ nm and the THz frequency 1.69 THz, $n_{gr}(812\ \mathrm{nm}) = n_T(1.69\ \mathrm{THz}) = 3.22$ in ZnTe (Lee, 2010). **Table 1** lists the optical wavelengths for OR THz generation velocity-matching at 2 THz in several typical zinc-blende crystals.

If a crystal system is highly symmetric, many of its second order nonlinear susceptibility tensor elements vanish. Among the nonvanishing ones, only a few of them are independent. For example, ZnTe has the crystal class of $\bar{4}3m$, which has three nonvanishing contracted matrix (Boyd, 2008) elements and only one is independent:

$$d_{il} = \begin{pmatrix} 0 & 0 & 0 & d_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & d_{14} \end{pmatrix} \tag{8}$$

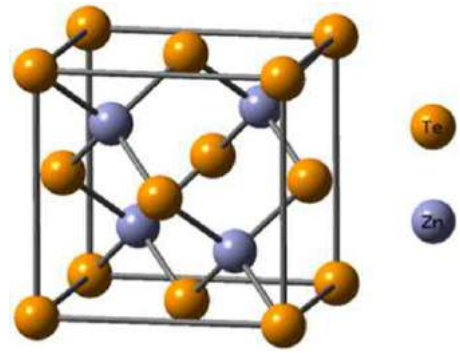


Fig. 6 ZnTe crystal structure. Reprinted from Kurban, M., Erkoç, Ş., 2016. Mechanical properties of CdZnTe nanowires under uniaxial stretching and compression: A molecular dynamics simulation study. Computational Materials Science 122, 295–300.

Table 1 Optical wavelength for velocity-matching in zinc-blende crystals at 2 THz

	ZnTe	*CdTe*	*GaP*	*InP*	*GaAs*
Wavelength (μm)	0.8	0.97	1.0	1.22	1.35

Source: From Lee, Y.S., 2010. Principles of Terahertz Science and Technology. New York: Springer, p. 90.

When a linear polarization optical beam is passing through the crystal, the OR induced polarization P can be expressed as follows:

$$\begin{pmatrix} P_x \\ P_y \\ P_z \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & d_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & d_{14} \end{pmatrix} \begin{pmatrix} E_x^2 \\ E_y^2 \\ E_z^2 \\ 2E_yE_x \\ 2E_zE_x \\ 2E_zE_y \end{pmatrix} \tag{9}$$

where x,y,z scales represent the [100], [010] and [001] crystal axis, respectively.

For a normal incidence case of a (110) cut ZnTe crystal, as shown in **Fig. 7(a)**, a linearly polarized incident beam is propagating along the [110] axis and the polarization direction has an angle θ with [001] axis, then the THz radiation intensity can be expressed as (Lee, 2010):

$$I_{THz}(\theta) = \frac{3}{4} I_{THz}^{max} \sin^2\theta(4 - 3\sin^2\theta) \tag{10}$$

where the I_{THz}^{max} is achieved when $\theta = \sin^{-1}\sqrt{\frac{2}{3}}$. **Fig. 7(b)** plots the THz radiation intensity as a function of θ.

The spectral bandwidth of the THz generation from electro-optic (EO) crystals is mainly limited by the absorption at THz frequency of the materials. From 5–10 THz, the absorption is normally dominated by the transverse-optical (TO) phonon resonances, whereas at lower frequencies, the absorption is attributed to more complicated combination of second-order phonon processes. ZnTe, for example, has a strong fundamental TO phonon resonance at 5.32 THz. At 1.6 and 3.7 THz, there are two major absorption bands as a mixture of LO-LA (longitudinal-optical and -acoustical) differences at the X and L points of the Brillouin zone (Schall *et al.*, 2001). **Fig. 8** shows the measured (solid line) absorption coefficient of ZnTe at THz frequency as well

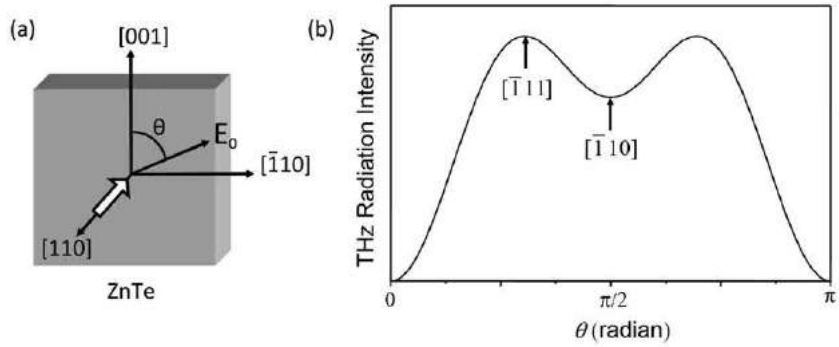


Fig. 7 (a) A linearly polarized optical wave is incident on a (110) ZnTe crystal with normal angle, θ is the angle between the optical field and the [001] axis; (b) THz radiation intensity from ZnTe as a function of θ. Modified and reprinted from Lee, Y.S., 2010. Principles of Terahertz Science and Technology. New York: Springer, p. 76.

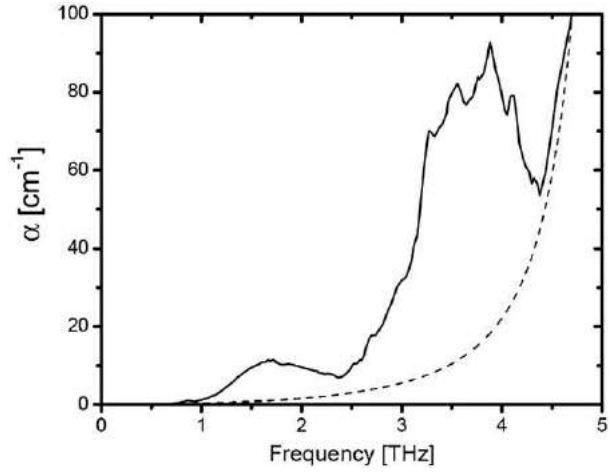


Fig. 8 Measured (solid line) power absorption coefficient (cm^{-1}) for ZnTe crystal compared with the calculated (dashed line) absorption for the TO-phonon line at room temperature. Reprinted from Lee, Y.S., 2010. Principles of Terahertz Science and Technology. New York: Springer, p. 90.

Table 2 Lowest TO-phonon lines of EO crystals

	ZnTe	*CdTe*	*GaP*	*InP*	*GaAs*	*GaSe*	*$LiNbO_3$*	*$LiTaO_3$*
ν_{TO} (THz)	5.3[a]	4.3[a]	11[b]	9.2[c]	8.1[d]	6.4[d]	7.7[e]	4.2[f]

[a]Schall *et al.* (2001).
[b]Leitenstorfer *et al.* (1999).
[c]Debernardi (1998).
[d]Kuroda *et al.* (1987).
[e]Seiji *et al.* (2002).
[f]Seiji *et al.* (2003).
Source: From Lee, Y.S., 2010. Principles of Terahertz Science and Technology. New York: Springer, p. 91.

Table 3 Properties of common materials for OR (Wu and Zhang, 1996)

Crystal	*EO coefficient (pmV^{-1})*	*Index of refraction*	*THz index of refraction*	*THz absorption coefficient (cm^{-1})*
ZnTe	$r_{41}=4.0$ (0.633 μm)	2.85 (0.8 μm)	~3.17	1.3
$LiNbO_3$	$r_{33}=30.9$ $r_{51}=32.6$ (0.633 μm)	$n_o=2.29$, $n_e=2.18$ (0.633 μm)	$n_o\sim6.8$, $n_e\sim4.98$	16
$LiTaO_3$	$r_{33}=r_{51}=30.5$ (0.82 μm)	$n_o=2.176$, $n_e=2.18$ (0.633 μm)	$n_o\sim6.5$, $n_e\sim6.4$	46
CdTe	$r_{41}=4.5$ (1.00 μm)	2.84 (0.8 μm)	~3.23	4.8
DAST	$r_{11}=160$ (0.82 μm)	$n_o=2.46$, $n_e=1.70$ (0.820 μm)	~2.4	150
GaSe	1.7 (0.8 μm)	2.85 (0.8 μm)	~3.72	0.07
GaAs	$r_{41}=1.43$ (1.15 μm)	3.61 (0.886 μm)	~3.4	0.5

Source: From Hafez, H.A., Chai, X., Ibrahim, A. *et al.*, 2016. Intense terahertz radiation and their applications. Journal of Optics 18, 093004.

as the calculated absorption (dashed line) due to the TO phonon (Gallot *et al.*, 1999). **Table 2** lists the lowest TO phonon lines of some commonly used EO crystals (Schall *et al.*, 2001; Leitenstorfer *et al.*, 1999; Debernardi, 1998; Kuroda *et al.*, 1987; Seiji *et al.*, 2002, 2003).

Tilted Pulse Front

Although ZnTe is one of the most popular EO crystals for THz generation, as shown in **Table 3**, its EO coefficient is relatively small compared with some other crystals, such as inorganic EO crystal $LiNbO_3$ and organic crystal 4-N-methylstilbazolium tosylate (DAST) (Hafez *et al.*, 2016). This section focuses on the tilted pulse front technique associated with THz generation from $LiNbO_3$, and THz generation from DAST and other organic crystals will be introduced in the next section.

Yang *et al.* (1971) first demonstrated the THz generation by OR with picosecond laser pulses, using $LiNbO_3$ crystal as the excited material. However, the conventional OR method of sending the pump beam at normal incidence to the crystal and generating THz radiation in the forward direction does not work efficiently with $LiNbO_3$, due to the large mismatch between the optical group velocity and the THz phase velocity in the material. The optical group refractive index is $n_O=2.3$, and the THz refractive index is $n_T=5.2$ (Lee, 2010). In order to conquer this mismatch, Hebling *et al.* (2002) proposed the pulse front tilting technique, which later became a standard procedure for THz generation from $LiNbO_3$ crystal.

In analogy to the Cherenkov radiation (Auston *et al.*, 1984b), in $LiNbO_3$ crystal, a femtosecond laser beam with considerably smaller beam size than the THz wavelength can be seen as a point source moving faster than the THz radiation, as shown in **Fig. 9(a)**. If the optical beam size is not negligible compared to THz wavelength, when the pulse front is aligned with the Cherenkov cone, like in **Fig. 9(b)**, the optical pulse front will propagate with the THz radiation at the same speed. Therefore, the velocity matching is fulfilled. This angle θ_c can be calculated using the optical group and THz phase refractive indices in $LiNbO_3$ as following:

$$\theta_c = cos^{-1}\left(\frac{v_T}{v_O}\right) = cos^{-1}\left(\frac{n_O}{n_T}\right) \cong 64^\circ \tag{11}$$

The tilt of the optical pulse front can be achieved with a diffractive grating (Hebling *et al.*, 2008). After the grating, an imaging system can be used to image the beam area on the grating onto the emitting surface of $LiNbO_3$ crystal. **Fig. 10** shows a typical experimental setup for pulse front tilting THz generation.

$LiNbO_3$ based pulse front tilting setup is one of the most intense table-top ultrafast THz sources. Hirori *et al.* (2011) reported single cycle THz pulse with amplitude exceeding 1 $MVcm^{-1}$ at room temperature. Huang *et al.* (2013) demonstrated that with cryogenically cooled $LiNbO_3$, the THz generation energy conversion efficiency can be increased about 3 times compared to room temperature crystal. Lange *et al.* (2014) have used 1.5 $MVcm^{-1}$ intense THz field generated by cryogenically cooled $LiNbO_3$ to excite extremely nonperturbative nonlinearities in GaAs.

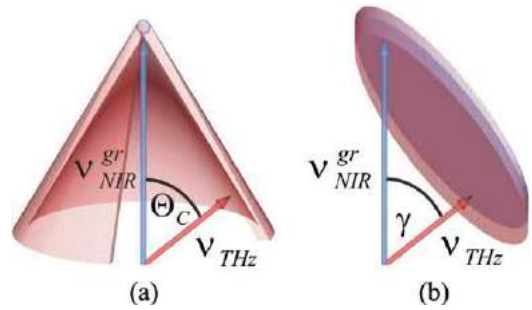


Fig. 9 (a) Cherenkov radiation: THz radiation is emitted as a cone with an angle θ_c. The velocity matching condition is satisfied if the excitation beam size is very small. (b) Pulse front tilting: velocity matching is automatically satisfied without upper limit of the excitation beam size, γ is the angle between the infrared beam velocity and the THz radiation velocity. Reprinted from Hebling, J., Almási, G., Kozma, I.Z., Kuhl, J., 2002. Velocity matching by pulse front tilting for large-area THz-pulse generation. Optics Express 10, 1161–1166.

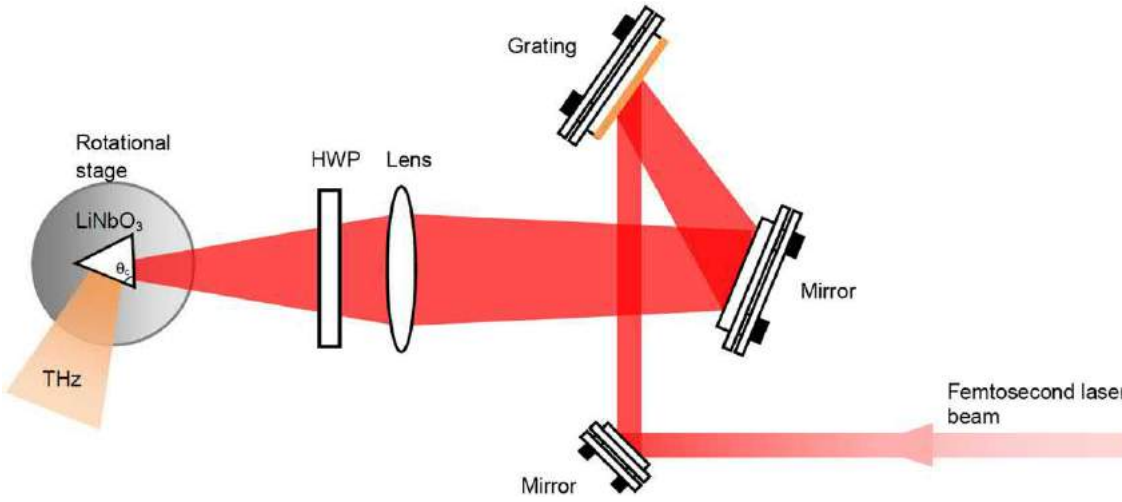


Fig. 10 Experimental setup for pulse tilting technique. HWP: Half Wave Plate.

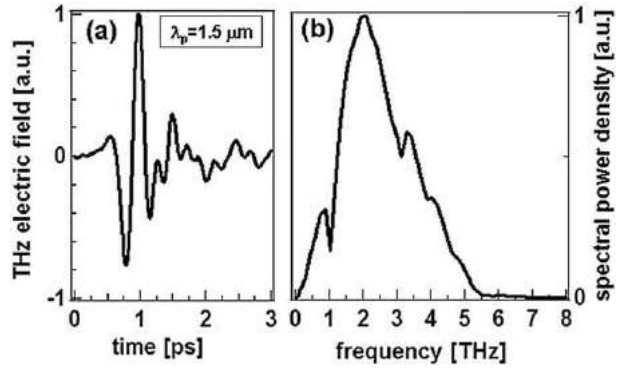


Fig. 11 (a) THz pulse from DAST. The corresponding THz spectrum is plotted in (b). Reprinted from Hauri, C.P., Ruchert, C., Vicario, C., Ardana, F., 2011. Strong-field single-cycle THz pulses generated in an organic crystal. Applied Physical Letters 99, 161116.

Organic Crystals

Organic crystals have recently become a great interest for ultra intense THz pulse generation.Zhang *et al.* (1992) first demonstrated the THz pulse generation from 4-N-methylstilbazolium tosylate (DAST). In 2011, a high quality THz beam generated from DAST with maximum electric field of 1.35 $MVcm^{-1}$ was demonstrated (Hauri *et al.*, 2011), as shown in **Fig. 11**. However, although the THz spectrum extended up to 5 THz, a TO phonon absorption of DAST can be clearly seen at 1.1 THz. To avoid this phonon absorption, other than cooling the crystal to suppress the vibration modes, alternative organic crystals such as

2-(3-(4- hydroxystyryl)-5,5-dimethylcyclohex-2-enylidene)malononitrile (OH1) and 4-N,N-dimethylamino-40-N0-methyl-stilbazolium 2,4,6-trimethylbenzenesulfonate (DSTMS) can be used. In 2014, Shalaby *et al.* generated an extremely bright THz bullet with peak field up to 83 $MVcm^{-1}$ (Shalaby and Hauri, 2015), which holds the current record of THz generation from an EO crystal. Meanwhile, the demanding conditions for growing good organic crystals in large dimension as well as the low damage threshold of organic crystals bring challenges to their wide use as THz sources.

Air-Plasma

Single-Color Method

Filamentation is the process by which a high-intensity beam self-focuses through nonlinear processes and collapses. This plasma channel stabilizes the beam at very small diameters (30 to 100 μm) and maintains high intensities over ranges much longer than the Rayleigh length of a traditional, geometrically focused beam. **Fig. 12** shows a laser induced filamentation in the air.

Intense broadband THz generation from ionized gas drew a great amount of attention ever since it was first demonstrated byHamster *et al.* (1993). In this work, strong femtosecond laser pulses were focused into a Helium gas target, generating strong emission of THz pulses with a conversion efficiency of less than 10^{-6}. The mechanism behind the THz radiation is the Ponderomotive force at the focus of the laser beam. The Ponderomotive force is experienced by a charged particle in an inhomogeneous oscillating electro-magnetic field, pushing the particle toward the area of low optical intensity regardless of the sign of the charge (Morales and Lee, 1974). However, since the ions are many orders of magnitude heavier compared to electrons, these forces generate a large density difference between ionic and electronic charges, and this charge separation results in a powerful electromagnetic transient, as schematically shown in **Fig. 13**.

The following relation between the emitted THz power from a one-color plasma and the excitation laser parameters was proposed byHamster *et al.* (1994):

$$P_{THz} \propto \left(\frac{W}{R_0}\right)^2 \left(\frac{\lambda}{\tau}\right)^4 \tag{12}$$

where W is the laser pulse energy, R_0 is the $1/e^2$ radius of the laser beam at the focus, λ is laser wavelength and τ is the pulse duration. Therefore, this equation predicts a strong dependence of the THz power on the wavelength and the pulse duration of the excitation laser. In the presence of strong bias field, the generation efficiency can be improved to comparable to that of THz radiation from semiconductor surface (Löffler *et al.*, 2000).

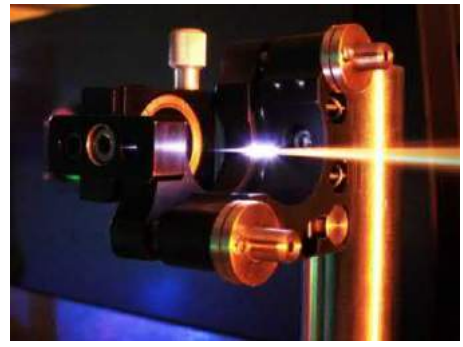


Fig. 12 Laser-induced air plasma (center bright line) emits an intense THz field.

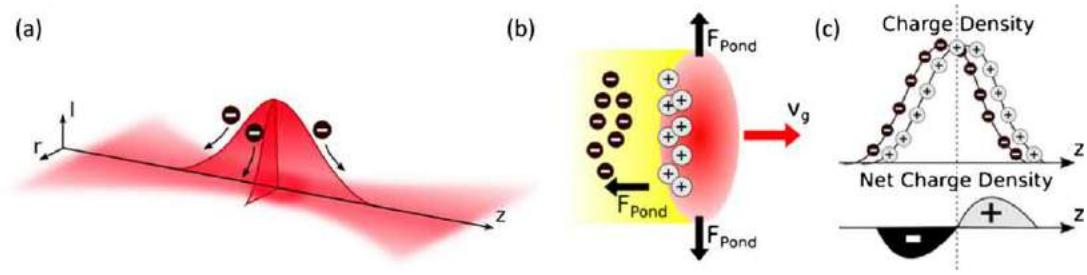


Fig. 13 (a) Laser propagates along *z* direction, the Ponderomotive force created by the gradient of the electric field drives free electrons. (b) Ponderomotive force drives electrons away in the plasma, leave the heavy ions relatively stationary. (c) Net charge density creates electrical dipole along *z* direction (Liu *et al.*, 2015).

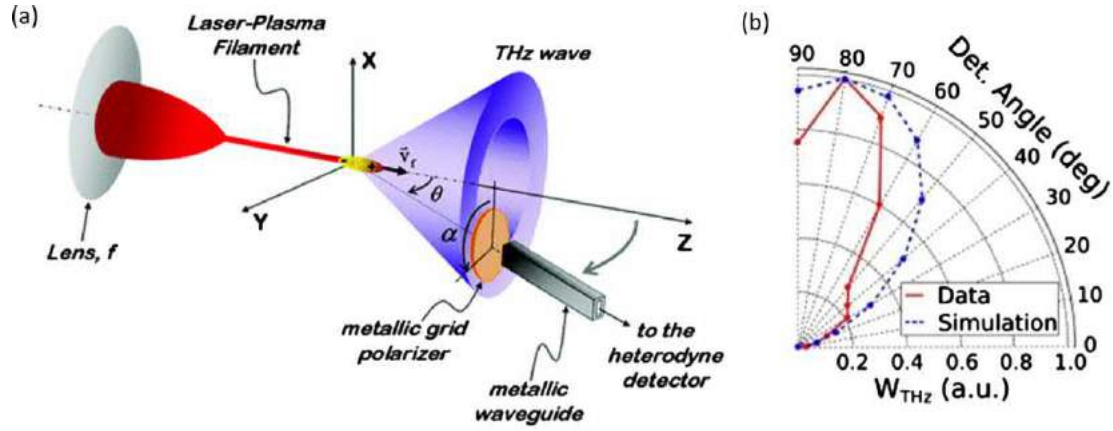


Fig. 14 (a) The experimental setup for measuring the THz radiation pattern from a laser-induced filamentation (Reprinted from D'Amico,C., Houard, A., Franco, M., *et al.*, 2007. Conical forward THz emission from femtosecond-laser-beam filamentation in air. Physical Review Letters 98, 235002). (b) THz radiation pattern from a microplasma measured (red solid line) and calculated from Eq. (13) (blue dotted line). Reprinted from Buccheri, F., Zhang, X.-C., Terahertz emission from laser-induced microplasma in ambient air. Optica 2, 366–369.

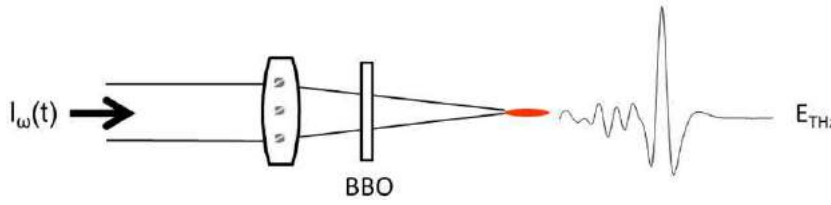


Fig. 15 Schematic demonstration of experimental setup for THz generation from two-color laser induced air plasma.

At the focus of the laser beam, the ionization front is moving in the wake of the laser pulse at light velocity, therefore generate THz radiation in a Cherenkov radiation cone shape in the forward direction, as shown in **Fig. 14(a)**. The angular energy density distribution of the THz radiation can actually be expressed as (Amico *et al.*, 2008):

$$f(\omega,\theta,L)=\frac{\sin^2\theta}{(1-\cos\theta)^2}\sin^2\left(\frac{L\omega}{2c}(1-\cos\theta)\right) \tag{13}$$

where ω is the radiated THz frequency, θ is the emission angle, L is the longitudinal length of the plasma column, and c is the speed of light. As we can see, the THz radiation angular distribution has something to do with the plasma length. To explore the extreme case of THz radiation from "microplasma," where the plasma length is at μm scale, Buccheri *et al.* demonstrated that with such a small length of plasma, the THz radiation cone angle opens up to about 80 degrees (Buccheri and Zhang, 2015), as shown in **Fig. 14(b)**.

Two-Color Method

All the THz generation schemes from gas plasma mentioned in last section use only one color laser pulses as the excitations. However, the mixing of laser pulses (fundamental ω) and their second harmonics (SH) 2ω to create the plasma has shown to provide enhancement of THz generation efficiency by several orders of magnitude (Cook and Hochstrasser, 2000). **Fig. 15** illustrates a basic experimental setup for the two-color air plasma generation method. The SH of the laser pulse is usually generated by using a type-I β-barium borate (BBO) crystal. The THz radiation intensity is maximized when the fundamental and SH polarizations are parallel, and is almost negligible when they are perpendicular (Kress *et al.*, 2004; Xie *et al.*, 2006).

THz radiation from two-color laser induced air plasma can be phenomenologically described as Four Wave Mixing (FWM) process, which is a third order nonlinear optical process. The model can be expressed as:

$$E_{THz}(t)\propto\chi^{(3)}E_{2\omega}(t)E_{\omega}^{*}(t)E_{\omega}^{*}(t) \tag{14}$$

where E_{THz}, E_{ω} and $E_{2\omega}$ are respectively the THz, the fundamental and the second harmonic (SH) electric field and $\chi^{(3)}$ is an effective third-order nonlinear susceptibility of the plasma filament. This relation has been experimentally confirmed by Xie *et al.* (2006) (see **Fig. 16**). It indicates that above the ionization threshold, the generated THz field is proportional to intensity of the fundamental laser and is also proportional to the square root of the second harmonic laser.

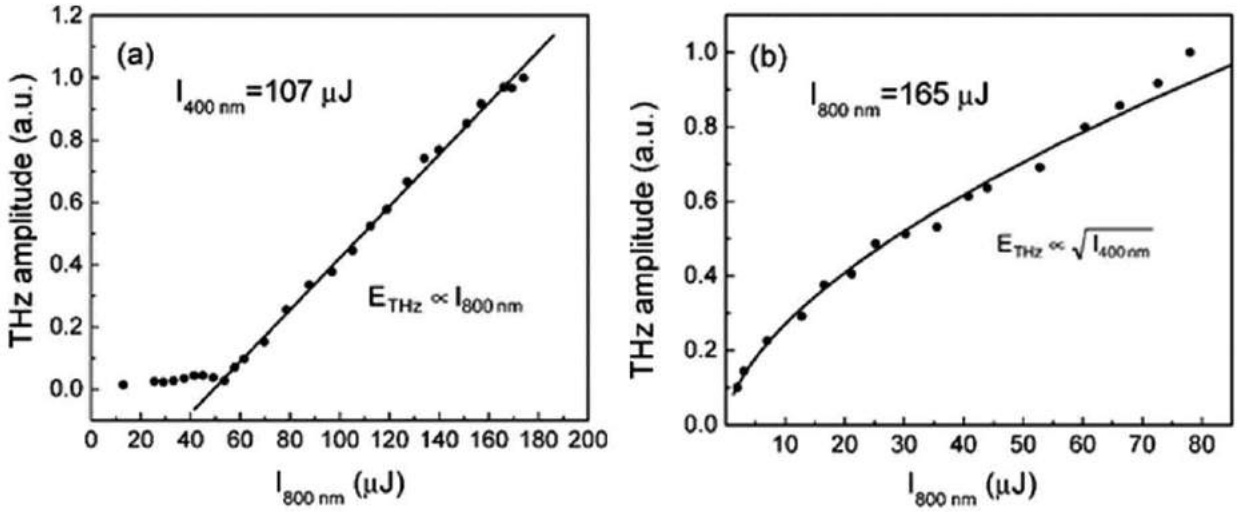


Fig. 16 (a) THz field as a function of laser intensities of 800 nm laser beam (a) and 400 nm laser beam (b) in THz wave generation by combination of 800 and 400 nm fs laser beams. Reprinted from Xie, X. Dai, J., Zhang, X.-C., 2006. Coherent control of THz wave generation in ambient air. Physical Review Letters 96, 075005.

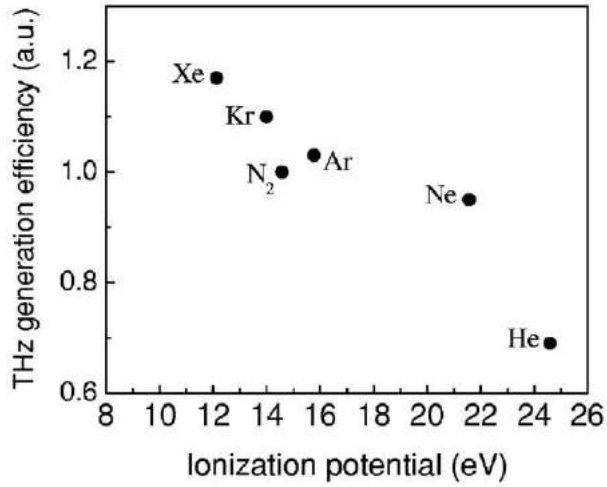


Fig. 17 Terahertz wave generation efficiency of six inert gases vs ionization potential at 20 Torr. Reprinted from Chen, Y., Yamaguchi, M., Wang, M., Zhang, X.-C., 2007. Terahertz pulse generation from noble gases. Applied Physics Letters 91, 251116.

Other than the ambient air, noble gases can also be used in the two-color plasma THz generation scheme. As shown in **Fig. 17**, the generation efficiencies from various noble gases highly depend on their ionization potentials (Chen *et al.*, 2007). The strongest THz emitter among them appeared to be Xenon.

The generated THz field from two-color air plasma can be modified by changing the phase between the fundamental pulses and the SH pulses. In the FWM model, this can be expressed as:

$$E_{THz}(t) \propto \chi^{(3)} E_{2\omega}(t) E_{\omega}^{*}(t) E_{\omega}^{*}(t) \cos((\varphi)) \tag{15}$$

where (φ) denotes the phase difference between ω and 2ω. Due to the dispersion of the two colors in air, the phase shift between these two color beams varies as they propagate. Therefore, the phase (φ) can be easily controlled by shifting the β-BBO longitudinally (Kress *et al.*, 2004). However, due to the laser intensity change toward the focus and the spatial limit between the focusing optic and the focus, simply shifting the β-BBO cannot control the phase (φ) precisely or significantly enough for certain purposes, such as remote generation of THz (Dai *et al.*, 2011). In those cases, phase compensators including a pair of silica wedges were exploited in order to fine tune the phases and polarizations of the ω and 2ω beams (Dai *et al.*, 2009). **Fig. 18** lists two types of phase compensators, which respectively control the two color beams in one beam path or in two separate paths. With the help of phase compensator, the generated THz field strength as well as the pulse polarity can be modified explicitly, as shown in **Fig. 19**

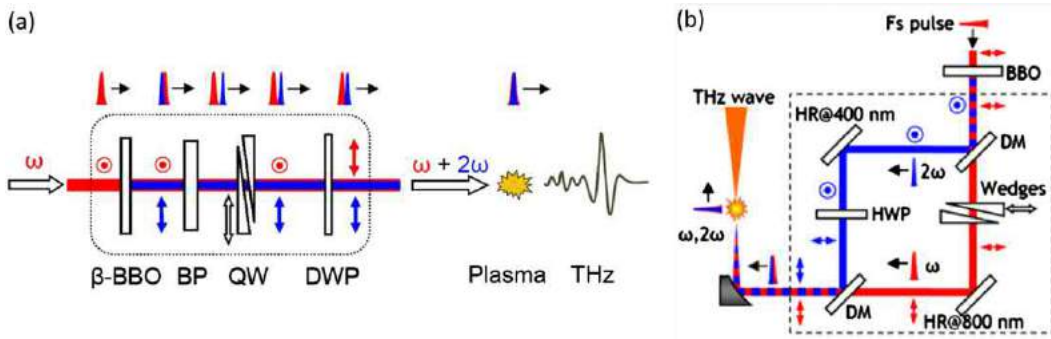


Fig. 18 (a) Inside the dashed line is the in-line PC (phase compensator). (b) Schematic illustration of the PC incorporated with a wedge pair: DM used to separate or recombine ω and 2ω beams; HWP used to control the polarization of the 2ω beam. DWP, dual-wavelength waveplate; BP, birefringent plate (α-BBO); QW, quartz wedges; DM, dichroic mirror. Reprinted from Dai, J., Karpowicz, N., Zhang, X.C., 2009. Coherent polarization control of terahertz waves generated from two-color laser-induced gas plasma. Physical Review Letters 103, 023001 and Dai, J., Liu, J., Zhang, X.-C., 2011. Terahertz wave air photonics: Terahertz wave generation and detection with laser-induced gas plasma. IEEE Journal of selected topics in Quantum Electronics 17, 183.

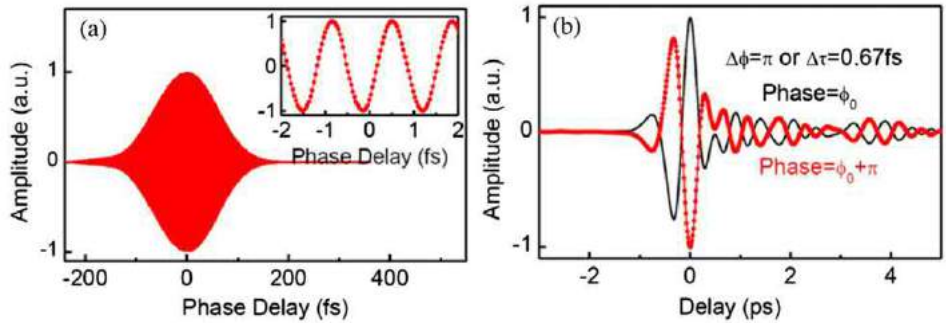


Fig. 19 (a) Phase scan obtained by continuously changing the insertion of the fused silica wedge pair while keeping the delay line for the THz waveform at the peak THz electric field. (b) Two THz waveforms with opposite polarities due to π phase shift induced by changing the relative insertion of the wedge pair in the PC. $\Delta(\varphi)$ is the phase change on relative phase difference between the two colors to generate these two THz waveforms, and $\Delta\tau$ is the corresponding time delay. Reprinted from Dai, J., Liu, J., Zhang, X.-C., 2011. Terahertz wave air photonics: Terahertz wave generation and detection with laser-induced gas plasma. IEEE Journal of selected topics in Quantum Electronics 17, 183.

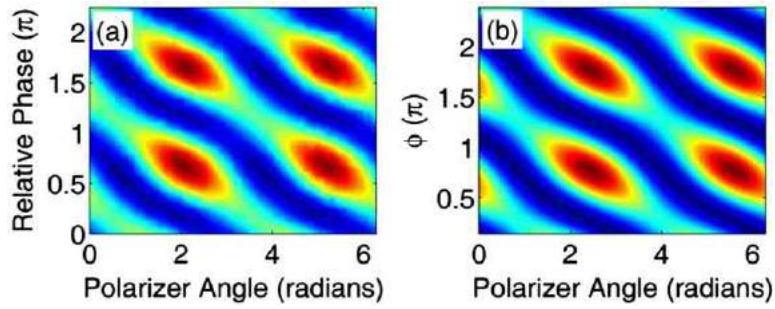


Fig. 20 THz intensity vs. THz polarizer angle and the relative phase between the ω and 2ω pulses with left-handed circularly polarized ω pulse and elliptically polarized 2ω pulse. (a) the experimental results, and (b) the corresponding simulation result. Reprinted from Dai, J., Karpowicz, N., Zhang, X.C., 2009. Coherent polarization control of terahertz waves generated from two-color laser-induced gas plasma. Physical Review Letters 103, 023001.

(Dai *et al.*, 2011). Further examination on how the various pump pulse polarizations affect the THz generation field was also enabled due to the phase compensator technique, as shown in **Fig. 20** (Dai *et al.*, 2009).

With ultra-short femtosecond laser pulses, an ultra-broadband THz wave with spectrum extending up to 30 THz can be achieved (Clough *et al.*, 2012; Kim *et al.*, 2008). Such a broad bandwidth requires broadband THz detection methods, such as THz-air-biased-coherent-detection (ABCD) (Dai *et al.*, 2011) or interferometric detection (Kim *et al.*, 2008), to reflect the complete

spectral properties of the THz source. **Fig. 21** compares the THz radiation from air-plasma detected by ZnTe crystal and THz-ABCD method. In 2014, maximum THz fields from two-color laser induced plasma reaching 8 MV/cm has been demonstrated (Oh *et al.*, 2014), covering the spectrum from 0.1–10 THz.

Although many experiments support the FWM model as the principle mechanism for THz generation from two-color laser induced air-plasma, the third order nonlinearity susceptibility $\chi^{(3)}$ of bond or free electrons is too small to explain the high intensity of THz radiation. Therefore, a transient photocurrent model was developed to explain coherent terahertz emission from air excited by a symmetry-broken laser field composed of ω and 2ω laser pulses (Kim *et al.*, 2007, 2008). In this model, a nonvanishing transverse plasma current is produced when the bound electrons are stripped off the ions by an asymmetric laser field, e.g. a mixture of two-color laser field with a proper relative phase. This photocurrent transient, occurring on the timescale of the photoionization, can thus produce electromagnetic radiation at THz frequencies. The electron dynamics right after the ionization were treated classically by Kim *et al.* (2007, 2008), whereas a full quantum mechanical simulation was reported by Karpowicz and Zhang (Dai *et al.*, 2009).

For elongated two-color filamentation, the THz radiation forms a cone shape due to off-axis phase matching (You *et al.*, 2012). The THz yield and angular distribution depend highly on the filament dimensions, and the dephasing length l_d, over which the THz radiation polarity remains the same. l_d can be expressed as

$$l_d = \left(\frac{\lambda}{2}\right)(n_\omega - n_{2\omega})^{-1} \tag{16}$$

where λ is the wavelength at ω, n_ω and $n_{2\omega}$ are respectively the refractive index at frequency ω and 2ω. For filament with electron density of $\sim 10^{16}$ cm^{-3} in ambient air, the dephasing length is about 22 mm.

As shown in **Fig. 22**, for a filament comparable to or shorter than the dephasing length, THz waves generated along the filament have both positive polarity and negative polarity. Those components interfere with each other constructively or destructively in the

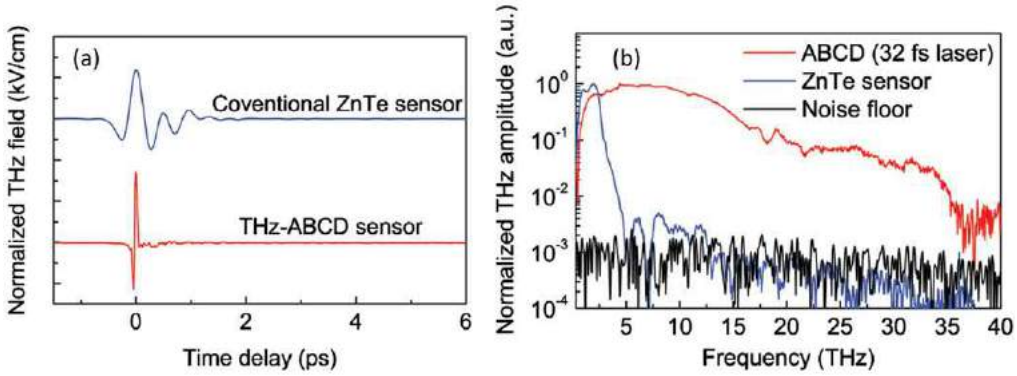


Fig. 21 (a) Time-resolved THz signals generated and detected using dry nitrogen gas as compared to conventional EO crystal detection in ZnTe. The probe beam for air detection has energy of 85 μJ and pulse duration of 32 fs. (b) Corresponding spectra after Fourier transform. Reprinted from Clough, B., Dai, J., Zhang, X.-C., 2012. Laser air photonics: Beyond the terahertz gap. Materials Today 15, 50–58.

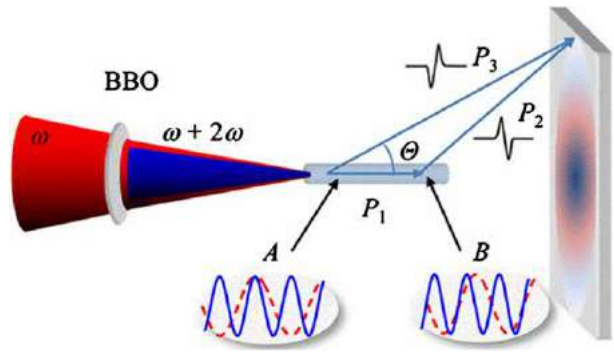


Fig. 22 Schematic of THz emission from a long, two-color laser-induced filament. The phase difference between 800 nm (dashed red curves) and 400 nm (solid blue curves) pulses along the filament results in a periodic oscillation of microscopic current amplitude and polarity. The resulting far-field THz radiation is determined by interference between the waves emitted from the local sources along the filament. P_1, P_2 and P_3 are respectively the optical path along the different directions as shown in the figure, θ is the angle between P_3 and P_1. Reprinted from You, Y.S., Oh, T.I., Kim, K.Y., 2012. Off-axis phase-matched terahertz emission from two-color laser-induced plasma filaments. Physical Review Letters 109, 183902.

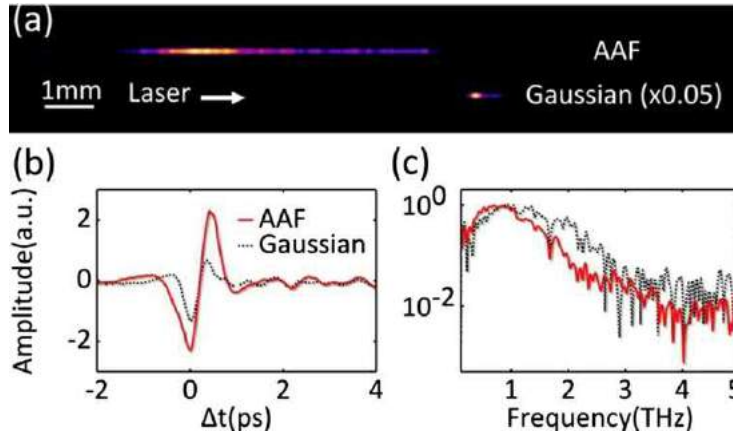


Fig. 23 (a) Fluorescence image comparison between an AAF beam plasma and a Gaussian plasma (intensity reduced by 20 times), both generated with pump pulse energy of 0.5 mJ. (b) Measured THz waveforms and (c) their corresponding normalized spectra generated by the two plasmas in (a). Reprinted from Liu, K., Koulouklidis, A.D., Papazoglou, D.G., Tzortzakis, S., Zhang, X.-C., 2016. Enhanced terahertz wave emission from air-plasma tailored by abruptly autofocusing laser beams. Optica 3, 605.

Table 4 List of peak THz electric field strength ranges from different sources. LA-PC, large-aperture photoconductive antenna; LA-EO, large-aperture electro-optic crystal

THz sources	*1–10 kVcm^{-1}*	*10–100 kVcm^{-1}*	*0.1–1 MVcm^{-1}*	*1–10 MVcm^{-1}*	*10–100 MVcm^{-1}*
LA-PC (Ropagnol *et al.*, 2016)		331 kVcm^{-1}			
Free electron lasers (Wu *et al.*, 2013)					44 MVcm^{-1}
LA-EO (Blanchard *et al.*, 2007)		230 kVcm^{-1}			
$LiNbO_3$ (Lange *et al.*, 2014)			1.5 MVcm^{-1}		
Organic crystal (Shalaby and Hauri, 2015)					83 MVcm^{-1}
Air Photonics (Oh *et al.*, 2014)				8 MVcm^{-1}	

far field, depending on their path difference which is related to the radiation angle, forming a donut shape far field radiation profile. You *et al.* (2012) also found out that the generation yield increases almost linearly with the plasma filament length, so one can increase the THz generation by simply extending the plasma length. Recently, different methods to tailor the filamentation toward more intense THz generation have been reported, including concatenating two plasmas into one (Manceau *et al.*, 2010) and using an abruptly autofocusing (AAF)optical beam (Liu *et al.*, 2016) (**Fig. 23**).

Summary

After introducing the main broadband THz pulse emitters, in this section we would like to present a chart comparing the maximum THz radiation amplitude of each source as of 2017 (Zhang *et al.*, 2017). Notice the large-aperture (LA) EO crystals listed do not include $LiNbO_3$-based tilted pulse front setup or organic crystals (**Table 4**).

See also: Strong-Field Terahertz Excitations in Semiconductors. Terahertz Lasers

References

Abo-Bakr, M., Feikes, J., Holldack, K., Wüstefeld, G., Hübers, H.W., 2002. Steady-state far-infrared coherent synchrotron radiation detected at BESSY II. Physical Review Letters 88, 254801.

Amico, C.D., Houard, A., Akturk, S., *et al.*, 2008. Forward THz radiation emission by femtosecond filamentation in gases: Theory and experiment. New Journal of Physics 10, 013015.

Auston, D.H., Cheung, K.P., Smith, P.R., 1984a. Picosecond photoconducting Hertzian dipoles. Applied Physics Letters 45, 284–286.

Auston, D.H., Cheung, K.P., Valdmanis, J.A., Kleinman, D.A., 1984b. Cherenkov radiation from femtosecond optical pulses in electro-optic media. Physical Review Letters 53, 1555–1558.

Blanchard, F., Razzari, L., Bandulet, H.C., *et al.*, 2007. Generation of 1.5 μJ single-cycle terahertz pulses by optical rectification from a large aperture ZnTe crystal. Optics Express 15, 13212–13220.

Boyd, R.W., 2008. Nonlinear Optics. Elsevier Science.
Buccheri, F., Zhang, X.-C., 2015. Terahertz emission from laser-induced microplasma in ambient air. Optica 2, 366–369.
Carr, G.L., Martin, M.C., Mckinney, W.R., *et al.*, 2002. High-power terahertz radiation from relativistic electrons. Nature 420, 153–156.
Chen, Y., Yamaguchi, M., Wang, M., Zhang, X.-C., 2007. Terahertz pulse generation from noble gases. Applied Physics Letters 91, 251116.
Clough, B., Dai, J., Zhang, X.-C., 2012. Laser air photonics: Beyond the terahertz gap. Materials Today 15, 50–58.
Cook, D.J., Hochstrasser, R.M., 2000. Intense terahertz pulses by four-wave rectification in air. Optics Letters 25, 1210.
Dai, J., Karpowicz, N., Zhang, X.C., 2009. Coherent polarization control of terahertz waves generated from two-color laser-induced gas plasma. Physical Review Letters 103, 023001.
Dai, J., Liu, J., Zhang, X.-C., 2011. Terahertz wave air photonics: Terahertz wave generation and detection with laser-induced gas plasma. IEEE Journal of selected topics in Quantum Electronics 17, 183.
Debernardi, A., 1998. Phonon linewidth in III-V semiconductors from density-functional perturbation theory. Physical Review B 57, 12847–12858.
Dexheimer, S.L., 2007. Terahertz Spectroscopy: Principles and Applications. Taylor & Francis.
Dhillon, S.S., Vitiello, M.S., Linfield, E.H., *et al.*, 2017. The 2017 terahertz science and technology roadmap. Journal of Physics D: Applied Physics 50, 043001.
Gallot, G., Zhang, J., Mcgowan, R.W., Jeon, T.-I., Grischkowsky, D., 1999. Measurements of the THz absorption and dispersion of ZnTe and their relevance to the electro-optic detection of THz radiation. Applied Physics Letters 74, 3450–3452.
Gu, P., Tani, M., Kono, S., Sakai, K., Zhang, X.-C., 2002. Study of terahertz radiation from InAs and InSb. Journal of Applied Physics 91, 5533–5537.
Hafez, H.A., Chai, X., Ibrahim, A., *et al.*, 2016. Intense terahertz radiation and their applications. Journal of Optics 18, 093004.
Hale, P.J., Madeo, J., Chin, C., *et al.*, 2014. 20 THz broadband generation using semi-insulating GaAs interdigitated photoconductive antennas. Optics Express 22, 26358–26364.
Hamster, H., Sullivan, A., Gordon, S., Falcone, R.W., 1994. Short-pulse terahertz radiation from high-intensity-laser-produced plasmas. Physical Review E 49, 671–677.
Hamster, H., Sullivan, A., Gordon, S., White, W., Falcone, R., 1993. Subpicosecond, electromagnetic pulses from intense laser-plasma interaction. Physical Review Letters 71, 2725.
Hauri, C.P., Ruchert, C., Vicario, C., Ardana, F., 2011. Strong-field single-cycle THz pulses generated in an organic crystal. Applied Physics Letters 99, 161116.
Hebling, J., Almási, G., Kozma, I.Z., Kuhl, J., 2002. Velocity matching by pulse front tilting for large-area THz-pulse generation. Optics Express 10, 1161–1166.
Hebling, J., Yeh, K.-L., Hoffmann, M.C., Bartal, B., Nelson, K.A., 2008. Generation of high-power terahertz pulses by tilted-pulse-front excitation and their application possibilities. Journal of the Optical Society of America B 25, B6–B19.
Hirori, H., Doi, A., Blanchard, F., Tanaka, K., 2011. Single-cycle terahertz pulses with amplitudes exceeding 1 MV/cm generated by optical rectification in $LiNbO_3$. Applied Physics Letters 98, 091106.
Huang, S.-W., Granados, E., Huang, W.R., *et al.*, 2013. High conversion efficiency, high energy terahertz pulses by optical rectification in cryogenically cooled lithium niobate. Optics Letters 38, 796–798.
Hu, B.B., Zhang, X.C., Auston, D.H., Smith, P.R., 1990. Free-space radiation from electro-optic crystals. Applied Physics Letters 56, 506–508.
Kim, K.Y., Glownia, J.H., Taylor, A.J., Rodriguez, G., 2007. Terahertz emission from ultrafast ionizing air in symmetry-broken laser fields. Optics Express 15, 4577–4584.
Kim, K.Y., Taylor, A.J., Glownia, J.H., Rodriguez, G., 2008. Coherent control of terahertz supercontinuum generation in ultrafast laser–gas interactions. Nature Photonics 2, 605.
Kress, M., Ler, T.L., Eden, S., Thomson, M., Roskos, H.G., 2004. Terahertz-pulse generation by photoionization of air with laser pulses composed of both fundamental and second-harmonic waves. In: Optics Letters, 29. . p. 1120.
Kuroda, N., Ueno, O., Nishina, Y., 1987. Lattice-dynamical and photoelastic properties of GaSe under high pressures studied by Raman scattering and electronic susceptibility. Physical Review B 35, 3860–3870.
Lange, C., Maag, T., Hohenleutner, M., *et al.*, 2014. Extremely nonperturbative nonlinearities in GaAs driven by atomically strong terahertz fields in gold metamaterials. Physical Review Letters 113, 227401.
Lee, Y.S., 2010. Principles of Terahertz Science and Technology. New York: Springer.
Leitenstorfer, A., Hunsche, S., Shah, J., Nuss, M.C., Knox, W.H., 1999. Detectors and sources for ultrabroadband electro-optic sampling: Experiment and theory. Applied Physics Letters 74, 1516–1518.
Liu, K., Buccheri, F., Zhang, X.-C., 2015. THz science and technology of micro-plasma. Physics (Chinese Wuli) 44, 497–502.
Liu, K., Koulouklidis, A.D., Papazoglou, D.G., Tzortzakis, S., Zhang, X.-C., 2016. Enhanced terahertz wave emission from air-plasma tailored by abruptly autofocusing laser beams. Optica 3, 605.
Löffler, T., Jacob, F., Roskos, H.G., 2000. Generation of terahertz pulses by photoionization of electrically biased air. Applied Physics Letters 77, 453–455.
Manceau, J.M., Massaouti, M., Tzortzakis, S., 2010. Strong terahertz emission enhancement via femtosecond laser filament concatenation in air. Optics Letters 35, 2424–2426.
Morales, G.J., Lee, Y.C., 1974. Ponderomotive-force effects in a nonuniform plasma. Physical Review Letters 33, 1016–1019.
Oh, T.I., Yoo, Y.J., You, Y.S., Kim, K.Y., 2014. Generation of strong terahertz fields exceeding 8 MV/cm at 1 kHz and real-time beam profiling. Applied Physics Letters 105, 041103.
Ropagnol, X., Khorasaninejad, M., Raeiszadeh, M., *et al.*, 2016. Intense THz pulses with large ponderomotive potential generated from large aperture photoconductive antennas. Optics Express 24, 11299–11311.
Schall, M., Walther, M., Uhd Jepsen, P, , 2001. Fundamental and second-order phonon processes in CdTe and ZnTe. Physical Review B 64, 094301.
Seiji, K., Hideaki, K., Seizi, N., Mitsuo Wada, T., 2003. Dielectric properties of ferroelectric lithium tantalate crystals studied by terahertz time-domain spectroscopy. Japanese Journal of Applied Physics 42, 6238.
Seiji, K., Naoki, T., Hideaki, K., Mitsuo Wada, T., Seizi, N., 2002. Terahertz time domain spectroscopy of phonon-polaritons in ferroelectric lithium niobate crystals. Japanese Journal of Applied Physics 41, 7033.
Shalaby, M., Hauri, C.P., 2015. Demonstration of a low-frequency three-dimensional terahertz bullet with extreme brightness. Nature Communications 6, 5976.
Shen, Y., Watanabe, T., Arena, D.A., *et al.*, 2007. Nonlinear cross-phase modulation with intense single-cycle terahertz pulses. Physical Review Letters 99, 043901.
Tani, M., Matsuura, S., Sakai, K., Nakashima, S.-I., 1997. Emission characteristics of photoconductive antennas based on low-temperature-grown GaAs and semi-insulating GaAs. Applied Optics 36, 7853–7859.
Wu, Q., Zhang, X.C., 1996. Ultrafast electro-optic field sensors. Applied Physics Letters 68, 1604–1606.
Wu, Z., Fisher, A.S., Goodfellow, J., *et al.*, 2013. Intense terahertz pulses from SLAC electron beams using coherent transition radiation. Review of Scientific Instruments 84, 022701.
Xie, X., Dai, J., Zhang, X.-C., 2006. Coherent control of THz wave generation in ambient air. Physical Review Letters 96, 075005.
Yang, K.H., Richards, P.L., Shen, Y.R., 1971. Generation of far-infrared radiation by picosecond light pulses in $LiNbO_3$. Applied Physics Letters 19, 320–323.
You, Y.S., Oh, T.I., Kim, K.Y., 2012. Off-axis phase-matched terahertz emission from two-color laser-induced plasma filaments. Physical Review Letters 109, 183902.
Zhang, X.C., Hu, B.B., Darrow, J.T., Auston, D.H., 1990. Generation of femtosecond electromagnetic pulses from semiconductor surfaces. Applied Physics Letters 56, 1011–1013.
Zhang, X.C., Ma, X.F., Jin, Y., *et al.*, 1992. Terahertz optical rectification from a nonlinear organic crystal. Applied Physics Letters 61, 3080–3082.
Zhang, X.C., Shkurinov, A., Zhang, Y., 2017. Extreme terahertz science. Nature Photonics 11, 16–18.
Zhang, X.-C., Xu, J., 2010. Introduction to THz Wave Photonics. New York: Springer.

Terahertz Detectors

Antoni Rogalski, Military University of Technology, Warsaw, Poland

Introduction

Terahertz (THz) range of electromagnetic spectrum still presents a challenge for both electronic and photonic technologies and is often described as the final unexplored area of spectrum. This radiation is frequently treated as the spectral region within frequency range $(v) \approx 0.1$–10 THz.

The past 20 years have seen a revolution in THz systems, as advanced materials research provided new and higher-power sources, and the potential of THz for advanced physics research and commercial applications was demonstrated. Numerous recent breakthroughs in the field have pushed THz research into the centre stage. As examples of milestone achievements can be included the development of THz time-domain spectroscopy (TDS), THz imaging, and high-power THz generation by means of nonlinear effects (Siegel, 2002; Bründermann *et al.*, 2012; Saeedkia, 2013; Sizov and Rogalski, 2010). Research involved with THz technologies is now receiving increasing attention, and devices exploiting this wavelength band are set to become increasingly important in diverse range of human activity applications (e.g., security, biological, drugs and explosion detection, gases fingerprints, imaging, etc.). Nowadays, the THz technology is also of much use in fundamentals science, such as nanomaterials science and biochemistry. This is based on the fact that THz frequencies correspond to single and collective excitations in nanoelectronic devices and collective dynamics in biomolecules.

General Classification of Terahertz Detectors

The majority of detectors can be classified in two broad categories: photon detectors and thermal detectors.

Photon Detectors

In photon detectors the radiation is absorbed within the material by interaction with electrons either bound to lattice atoms or to impurity atoms or with free electrons. The radiation can be also absorbed by electrons localized in the quantum wells or in the minibands of superlattices. The observed electrical output signal results from the changed electronic energy distribution. The photon detectors show a selective wavelength dependence of response per unit incident radiation power.

Depending on the nature of the interaction, the class of photon detectors is further sub-divided into different types. The most important are: intrinsic detectors, extrinsic detectors, and photoemissive (Schottky barriers). Different types of detectors are described in detail in the monograph *Infrared Detectors* (Rogalski, 2011). **Fig. 1** shows spectral detectivity characteristics of different types of detectors.

Photodetectors that utilize excitation of an electron from the valence to conduction band are called intrinsic detectors. Those which operate by exciting electrons into the conduction band or holes into the valence band from impurity states within the band (impurity-bound states in energy gap, quantum wells or quantum dots), are called extrinsic detectors. In comparison with intrinsic photoconductivity, the extrinsic photoconductivity is far less efficient because of limits in the amount of impurity that can be introduced into semiconductor without altering the nature of the impurity states. Intrinsic detectors are most common at the short wavelengths, below 20 μm.

A key difference between intrinsic and extrinsic detectors is that extrinsic detectors require much cooling to achieve high sensitivity at a given spectral response cutoff in comparison with intrinsic detectors. Low-temperature operation is associated with longer-wavelength sensitivity in order to suppress noise due to thermally induced transitions between close-lying energy levels. The long wavelength cutoff can be approximated as $T_{\max} = 300\ \text{K}/\lambda_c\ [\mu\text{m}]$, where λ_c is the cut-off wavelength.

Thermal Detectors

The second class of detectors is composed of thermal detectors – their operation principles are briefly described in **Table 1**. In a thermal detector the incident radiation is absorbed to change the material temperature, and the resultant change in some physical property is used to generate an electrical output. The detector is suspended on lags, which are connected to the heat sink (see figures inside of **Table 1**). The signal does not depend upon the photonic nature of the incident radiation. Thus, thermal effects are generally wavelength independent; the signal depends upon the radiant power (or its rate of change) but not upon its spectral content. Attention is directed toward three approaches which have found the greatest utility in infrared technology, namely, bolometers, pyroelectric, and thermoelectric effects. In pyroelectric detectors a change in the internal electrical polarization is measured, whereas in the case of thermistor bolometers a change in the electrical resistance is measured.

 doi:10.1016/B978-0-12-803581-8.09641-7

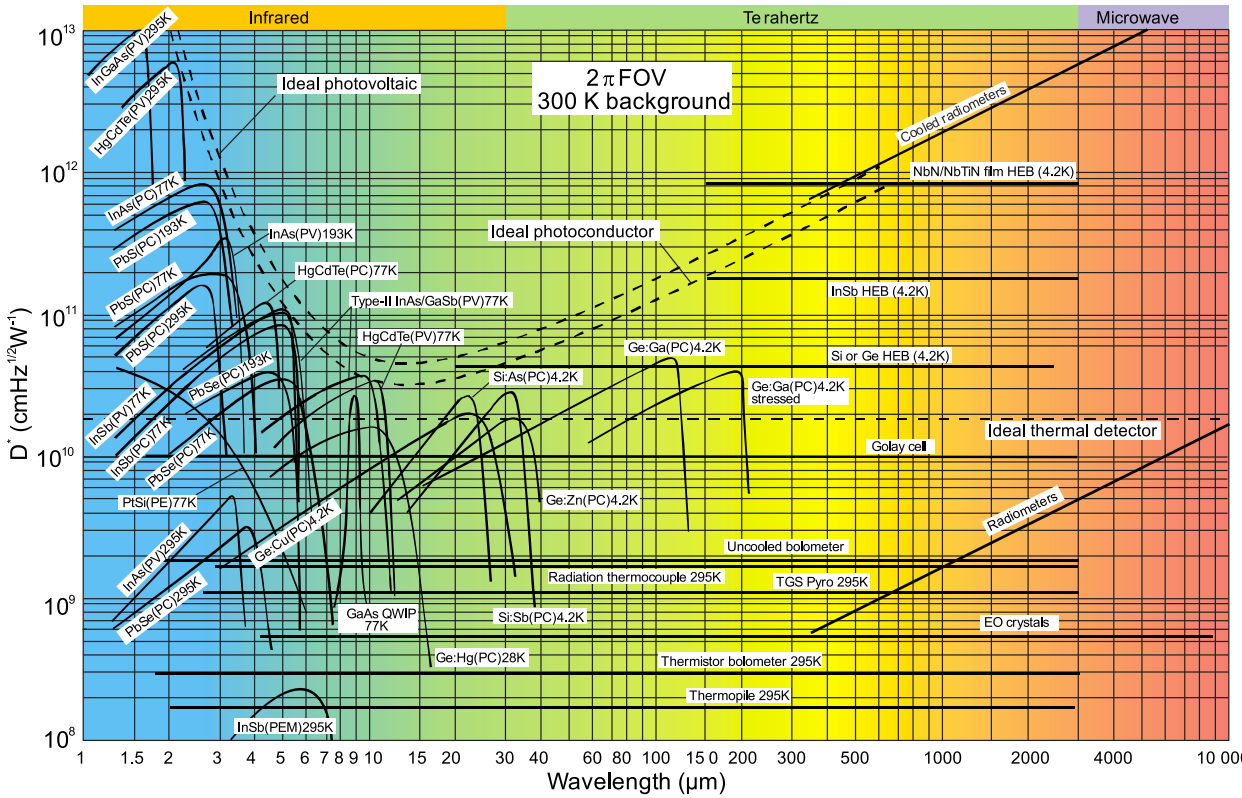


Fig. 1 Comparison of the D^* of various available detectors when operated at the indicated temperature. Chopping frequency is 1000 Hz for all detectors except the thermopile (10 Hz), thermocouple (10 Hz), thermistor bolometer (10 Hz), Golay cell (10 Hz) and pyroelectric detector (10 Hz). Each detector is assumed to view a surrounding hemisphere (2π field of view) at a temperature of 300 K. Theoretical curves for the background-limited D^* (dashed lines) for ideal photovoltaic and photoconductive detectors and thermal detectors are also shown. PC, photoconductive detector; PEM, photoelectromagnetic detector; PV, photovoltaic detector; HEB, hot electron bolometer.

Bolometers may be divided into several types. The most commonly used are the metal, the thermistor, and the semiconductor bolometers. A fourth type is the superconducting bolometer. This bolometer operates on a conductivity transition in which the resistance changes dramatically over the transition temperature range.

The key trade-off with respect to conventional thermal detectors is between sensitivity and response time. The detector sensitivity is often expressed by noise equivalent temperature (*NEDT*) represented by the temperature change, for incident radiation, that gives an output signal equal to the rms noise level. The thermal conductance is an extremely important parameter, since the noise equivalent temperature difference (*NEDT*) is proportional to $(G_{th})^{1/2}$, but the thermal response time of the detector, τ_{th}, is inversely proportional to G_{th}. Therefore, a change in thermal conductance due to improvements in material processing techniques improves sensitivity at the expense of time response.

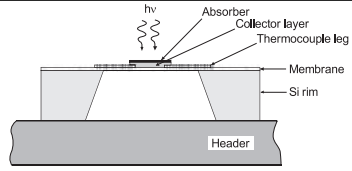
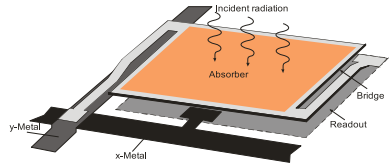
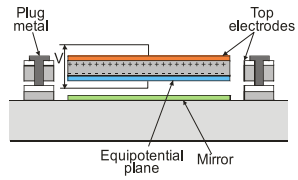
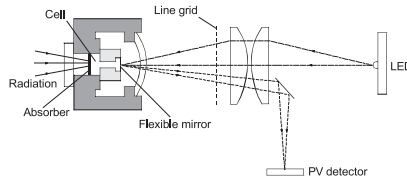
Detectors for Room Temperature THz Imaging

Particular attention in development of THz imaging systems is devoted to the realization of sensors with a large potential for real-time imaging while maintaining a high dynamic range and room-temperature operation. CMOS process technology is especially attractive due to their low price tag for industrial, surveillance, scientific, and medical applications. However, CMOS THz imagers developed thus far have mainly operated single detectors based on lock-in technique to acquire raster-scanned imagers with frame rates on the order of minutes. With this mind, much of recent developments are directed towards three types of focal plane arrays (FPAs):

- Schottky barrier diodes (SBDs) compatible with the CMOS process,
- field effect transistors (FETs) relying on plasmonic rectification phenomena, and
- adaptation of infrared bolometers to the THz frequency range.

An important issue for a FPA is pixel uniformity. It appears however, that the production of monolithically-integrated detector arrays encounters so many technological problems that the device-to-device performance variations and even the percentage of non-functional detectors per chip tend to be unacceptably high.

Table 1 Thermal detectors

Mode of operation	*Schematic of detector*	*Operation and properties*
Thermopile		The thermocouple is usually a thin, blackened flake connected thermally to the junction of two dissimilar metals or semiconductors. Heat absorbed by the flake causes a temperature rise of the junction, and hence a thermoelectric electromotive force is developed which can be measured. Although thermopiles are not as sensitive as bolometers and pyroelectric detectors, they will replace these in many applications due to their reliable characteristics and good cost/performance ratio. Thermocouples are widely used in spectroscopy.
Bolometer Metal Semiconductor		The bolometer is a resistive element constructed from a material with a very small thermal capacity and large temperature coefficient so that the absorbed radiation produces a large change in resistance. The change in resistance is like the photoconductor; however, the basic detection mechanisms are different. In the case of a bolometer, radiant power produces heat within the material, which in turn produces the resistance change. There is no direct photon-electron interaction.
Superconductor Hot electron		Initially, most bolometers were the thermistor type made from oxides of manganese, cobalt, or nickel. At present microbolometers are fabricated in large format arrays for thermal imaging applications. Some extremely sensitive low-temperature semiconductor and superconductor bolometers are used in THz region.
Pyroelectric detector		The pyroelectric detector can be considered as a small capacitor with two conducting electrodes mounted perpendicularly to the direction of spontaneous polarization. During incident of radiation, the change in polarization appears as a charge on the capacitor and a current is generated, the magnitude of which depends on the temperature rise and the pyroelectrical coefficient of the material. The signal, however, must be chopped or modulated. The detector sensitivity is limited either by amplifier noise or by loss-tangent noise. Response speed can be engineered making pyroelectric detectors useful for fast laser pulse detection, however with proportional decrease in sensitivity.
Golay cell		The Golay cell consists of an hermetically sealed container filled with gas (usually xenon for its low thermal conductivity) and arranged so that expansion of the gas under heating by a photon signal distorts a flexible membrane on which a mirror is mounted. The movement of the mirror is used to deflect a beam of light shining on a photocell and so producing a change in the photocell current as the output. In modern Golay cells the photocell is replaced by a solid state photodiode and light emitting diode is used for illumination. The performance of the Golay cell is only limited by the temperature noise associated with the thermal exchange between the absorbing film and the detector gas; consequently, the detector can be extremely sensitive with $D^* \approx 3 \times 10^9$ cm $Hz^{1/2}$ W^{-1}, and responsivities of 10^5 to 10^6 V/W. The response time is quite long, typically 15 ms.

The performance of monolithically integrated detector arrays with room temperature THz detectors is summarized in **Table 2**. SPDs respond to the THz electric field and usually generate an output current or voltage through a quadratic term in their current-voltage characteristics. In general, the noise equivalent power (*NEP*) of SBD and FET detectors is better than that of Golay cells and pyroelectric detectors around 300 GHz. Both the pyroelectric and the bolometer FPAs with detector response times in the millisecond time range are not suited for heterodyne operation. FET detectors are clearly capable in heterodyne detection with improving sensitivity. Diffraction aspects predicts FPAs for higher frequencies (0.5 THz and above) and in conjunction with large f/# optics.

Table 2 Parameter of some uncooled THz detectors

Device type	*Electrical responsivity (V/W)*	*Conditions*	*NEP (W/Hz$^{1/2}$)*
Schottky diodes			
ErAs/InGaAlAs spiral planar antenna	–	Zero bias, 639 GHz	4.0×10^{-12} NEDT = 120 mK
InGaAs log-spiral antenna	~200 for system estimate 10^3 intrinsic for the diode	0.8 THz	5.0×10^{-12}
VDI Model: WR2.8 ZBD	1500	260–400 GHz	2.7×10^{-12}
VDI Model:WR1.5ZBD	750	500–750 GHz	5.1×10^{-12}
VDI Model: WR1.0 ZBD	200	750–1100 GHz	20×10^{-12}
VDI Model:WR0.65 ZBD	100	1100–1700 GHz	40×10^{-12}
Bolometers			
$Hg_{0.8}Cd_{0.2}Te$ HEB	0.30 at 17 mV bias, 36 GHz	Room temperature	2.2×10^{-9} for 17 mV bias, 35 GHz
	96 for 0.89 THz, 13 mV bias		7.4×10^{-9} for 0.89 THz, 12 mV bias
SixGey:H	170	0.934 THz, uncooled	0.2×10^{-9}
Vanadium oxide	-	Uncooled	320×10^{-12} at 4.3 THz, 9×10^{-13} @ 7.5-14 μm
Niobum film	21	3.6 mA bias, 1 kHz mod, 300K	1.10×10^{-10}
Ti, antenna-coupled microbolometer	-	10 kHz chop, 1.04 mA bias, 300K	1.5×10^{-11}
Nb_5N_6	400	0.4 mA bias, >10 kHz	9.8×10^{-12}
Vanadium oxide array	1.5×10^4	1 V bias, 130 μm, uncooled	2.00×10^{-10}
Nb, polymide, antenna coupled	450	<1 THz	1.5×10^{-11}
Al/Nb; antenna coupled	85	1 kHz mod, 1.6 mA bias	2.5×10^{-11}
Fee-standing Nb bridge antenna coupled	210 (average over 10 devices)	650 GHz	12.5×10^{-12}
Pyroelectrics			
Philips P5219 deuterated L-alanine TGS	321	10 Hz mod; amplifier with gain of 4.8, 91 GHz	3.1×10^{-8}
QMC instr	18,300	10 Hz mod;	4.4×10^{-10}
	1200	1.89 THz, <20 Hz mod	
$LiTaO_3$	-	530 GHz, Melectron Model SPH-45	2.0×10^{-9}
Golay cells			
Tydex Golay Cell GC-1X	100 000	21 Hz chper	1.4×10^{-10}
Microtech Instruments	10000	12.5 Hz chopper	10×10^{-8}
Micro-array, layer by layer, polimer membranes over Si	-	30 Hz mod 105 GHz	300×10^{-9}
Tydex Golay cell, 6-mm-diameter diamond window		10 Hz mod	7.0×10^{-10}
CMOS-based and plasma detectors			
BiCMOS SiGe, 0.25 μm HBT	Current R_i 1 A/W at 0.7 THz	3 × 5 array, chopper 125 kHz	50×10^{-12} at 0.7 THz
BiCMOS SiGe, 0.25 μm NMOS	Voltage R_v 80 kV/W at 0.6 THz	3 × 5 array, chopper 16 kHz	300×10^{-12} at 0.6 THz
CMOS SiGe, 65 nm NMOS	Voltage R_v 140 kV/W at 0.87 THz	32 × 32 array, chopper 5 kHz	100×10^{-12} at 0.87 THz
CMOS SiGe, 65 nm NMOS	Voltage R_v 0.8 kV/W at 1 THz	3 × 5 array, chopper 1 kHz	66×10^{-12} at 1 THz
CMOS-SBD, 130 nm	Voltage R_v 0.323 kV/W at 0.28 THz	4 × 4 array, chopper 1 kHz	29×10^{-12} at 0.28 THz
CMOS-SBD, 65 nm	–	1 element; 1 MHz mod	42×10^{-12} at 0.86 THz
CMOS, 150 nm, NMOS	Voltage R_v at 4.1 THz	1 element	133×10^{-12} at 4.1 THz
InGaAs HEMT	Voltage R_v 23 kV/W at 200 GHz	1 element	0.5×10^{-12} at 200 GHz
Asymmetric dual-grating gate InGaAs HEMT	Voltage R_v 6.4 kV/W at 1.5 THz	1 element	50×10^{-12} at 1.5 THz

Below, a short description of different kinds of uncooled THz detectors is presented (Brown and Segovia-Vargas, 2015).

Schottky Barrier Diodes

In spite of achievements of other kind of detectors for THz waveband, the Schottky barrier diodes (SBDs) are among the basic elements in THz technologies. They are used either in direct detection and as nonlinear elements in heterodyne receiver mixers

operating in temperature range of 4–300 K. The cryogenically cooled SBDs were used in mixers preferably in 1980s and early 1990s and then they have been replaced widely by superconductor-insulator-superconductor (SIS) or hot electron bolometer (HEB) mixers, in which mixing processes are similar to that observed in SBDs, but, e.g., in SIS structures the rectification process is based on quantum-mechanical photon-assisted tunnelling of quasiparticles (electrons). The nonlinearity of SBD *I–V* characteristic (the current increases exponentially with the applied voltage) is the prerequisite for mixing to occur.

Historically first Schottky-barrier structures were pointed contacts of tapered metal wires (e.g., a tungsten needle) with a semiconductor surface (the so-called crystal detectors). Due to limitation of whisker technology, such as constraints on design and repeatability, starting in the 1980s, the efforts were made to produce planar Schottky diodes with air-bridge fingers (see **Fig. 2(a)**). This design has been the most important steps toward a practical Schottky diode mixer for THz frequency applications, with several thousand diodes on a single chip and where parasitic losses such as the series resistance and the shunt capacitance are minimized. To achieve good performance at high frequencies the diode area should be small. Reducing junction area one reduces junction capacitances to increase operating frequency. But at the same time one increases the series resistance.

Using advanced technology, the diodes are integrated with many passive circuit elements (impedance matching, filters, and waveguide probes) onto the same substrate. By improving the mechanical arrangement and reducing loss, the planar technology is pushed well beyond 300 GHz up to several THz. For example, **Fig. 2(b)** shows photographs of a bridged four-Schottky diodes' chip arrayed in a balanced configuration to increase power handling.

Recently, an alternative method of Schottky barrier formation has been elaborated by molecular beam epitaxy (MBE) in-situ deposition of a semimetal (ErAs) on semiconductor (InGaAs/InAlAs on InP substrates) to reduce the imperfections that give rise to excess low-frequency noise, particularly l/f noise (Brown *et al.*, 2006). Excellent *NEP* performance for this III–V semiconductor SBD has been reported (1.4 pW/Hz$^{1/2}$ at 100 GHz). By using interband tunnelling, a heterojunction backward diode demonstrated 49.7 kV/W responsivity and 0.18 pW/Hz *NEP* at 94 GHz (Zhang *et al.*, 2011). More recently Han *et al.* (2012) have demonstrated fully functional CMOS imager operating near or in the sub-millimeter-wave frequency range (see **Fig. 3**). The 4 × 4 array increases the imaging speed by 4–8 times, due to fewer mechanical scan steps.

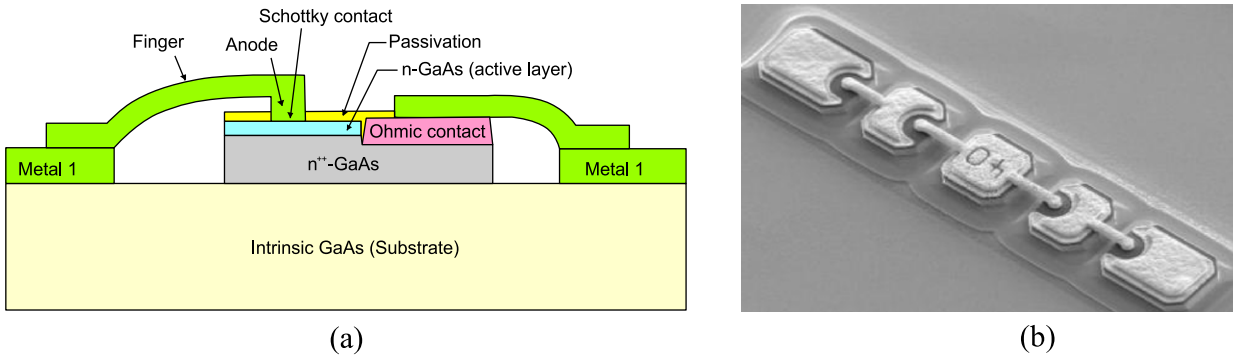


Fig. 2 Gaas Schottky barrier diode: (a) schematic of a planar diode and (b) a four-diode chip array.

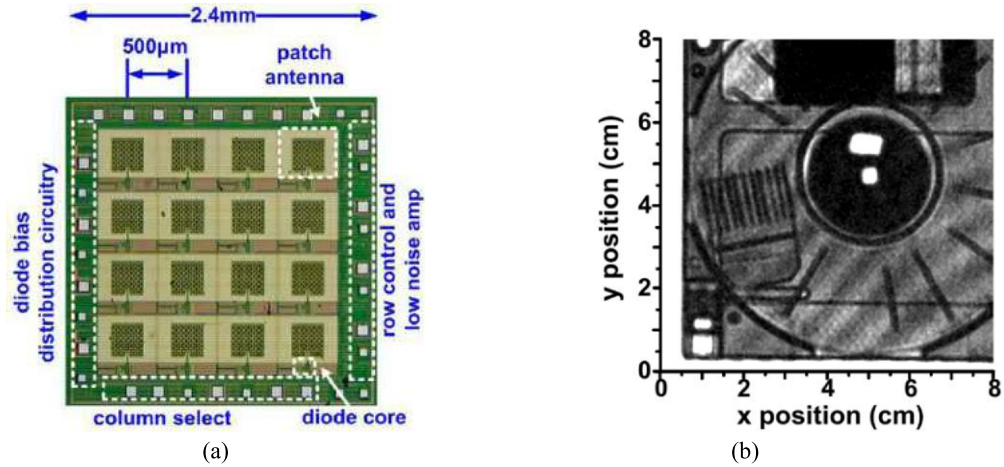


Fig. 3 CMOS SBD 280-GHz imager: die photos of the array (a) and an image of music greeting card obtained. After Han, R., Zhang, Y., Kim, Y., *et al.*, 2012. 280 GHz and 860 GHz image sensors using Schottky-barrier diodes in 0.13 μm digital CMOS. IEEE International Solid-State Circuits Conference, 254–256.

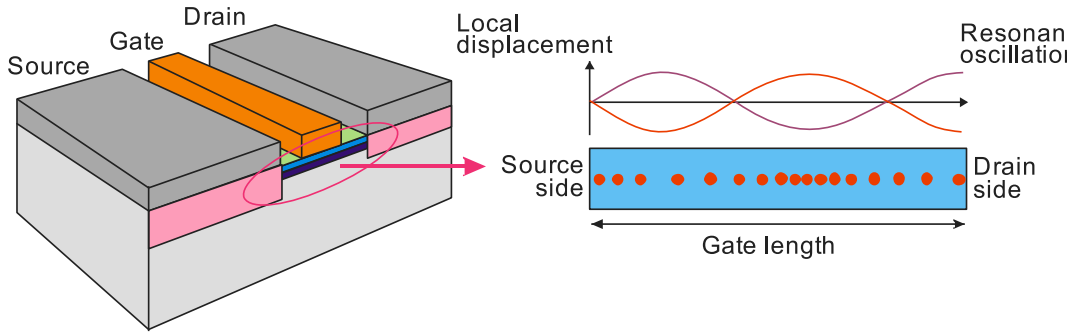


Fig. 4 Plasma oscillations in a field effect transistor.

Field Effect Transistor and CMOS-Based Detectors

The use field effect transistors (FETs) as detectors of THz radiation was first proposed by Dyakonov and Shur (1993) on the basis of formal analogy between the equations of the electron transport in a gated two-dimensional transistor channel and those of shallow water, or acoustic waves in music instruments. **Fig. 4** schematically shows the resonant oscillation of plasma waves in gated region of FET.

The detection by FETs is due to nonlinear properties of the transistor, which lead to the rectification of the ac current induced by the coming radiation. As a result, a photoresponse appears in the form of a dc voltage between source and drain. This voltage is proportional to the radiation intensity (photovoltaic effect). A big improvement in sensitivity can be obtained by adding a proper antenna or a cavity coupling. The FETs can be used both for resonant (resonant case of high electron mobility, when plasma oscillation modes are excited in the channel) and non-resonant (broadband) THz detection in the case of low mobility, where plasma oscillations are overdamped (Knap and Dyakonov, 2013).

The large-scale interest in using FETs as THz detectors started around 2004 after the first experimental demonstration of sub-THz and THz detection in silicon-CMOS FETs. Two years later it was shown that Si-CMOS FETs can reach a *NEP* value competitive with the best conventional room temperature THz detectors. At present the advantages of Si-CMOS FET technology (room temperature operation, very fast response time, easy on-chip integration with read-out electronics and high reproducibility) lead to the straight-forward array fabrication (Knap *et al.*, 2014).

Recently, the first CMOS FPA used to capture transmission-mode THz video streams in real-time without the need for raster scanning and source modulation has been fabricated. A camera with 32×32 pixel array fully integrated in a 65-nm CMOS process technology has been demonstrated (Al Hadi *et al.*, 2012) (see right side of **Fig. 5**). Each 80-μm array pixel consists of a differential on-chip ring antenna coupled to NMOS direct detector operated well-beyond its cut-off frequency. The camera chip has been packed together with a 41.7-dBi silicon lens in a $5 \times 5 \times 3$ cm^3 camera module. In continuous-wave illumination the camera achieves an responsivity of 100–200 kV/W and a total *NEP* of 10–20 $nW/Hz^{1/2}$ up to 500 fps at 856 GHz.

Microbolometers

An impressive promising technology is also coming from commercially available microbolometer arrays. Adaptation of infrared microbolometers to the THz frequency range after the successful demonstration of active THz imaging in 2006 (Lee *et al.*, 2006) entailed that in the period 2010–2011 three different companies/organizations announced cameras optimized for the >1-THz frequency range: NEC (Japan) (Oda, 2010), INO (Canada) (Bolduc *et al.*, 2011) and Leti (France) (Nguyen *et al.*, 2012). The number of vendors is expected to increase soon.

Different designs of THz bolometer pixels have been proposed. NEC's pixel is divided into two parts (see left side of **Fig. 5**): a silicon readout integrated circuit (ROIC) in the lower part, and a suspended microbridge structure in the upper part. The microbridge has a two-storied structure. The bottom is composed of a diaphragm and two legs, while the top (eaves) structure is formed on the diaphragm to increase the sensitive area and fill factor. The diaphragm and the eaves absorb THz radiation. The diaphragm is composed of VO_x bolometer thin film, SiN_x passivation layers and TiAlV electrodes, while the eaves structure is composed of SiN_x layer and TiAlV thin film THz absorption layer.

A schematic of one Leti pixel of an amorphous silicon microbolometer array is shown in top centre of **Fig. 5**. The 50-μm pitch is associated with quasi-double-bowtie antennas to a thermometer microbridge structure derived from the standard IR bolometer. The membrane is suspended over the substrate by arms and pillars. In order to enhance the antenna gain, an equivalent quarter-wavelength resonant cavity is realized under antennas with an 11-μm thick SiO_2 layer deposited over the metallic reflector. To ensure electric contact between the bolometer pillars and CMOS metal upper contacts, the vias are etched through an 11-μm cavity and then metalized.

Fig. 6 summarizes the *NEP* values for bolometer FPAs fabricated by three vendors. The FPAs optimized for 2–5 THz exhibit impressive *NEP* values below 100 $pW/Hz^{1/2}$. It can be seen that wavelength dependence of *NEP* is quite flat below 200 μm. Further improvement of performance is possible by increasing number of pixels, modification of antenna design while preserving pixel pitch, ROIC and technological stack.

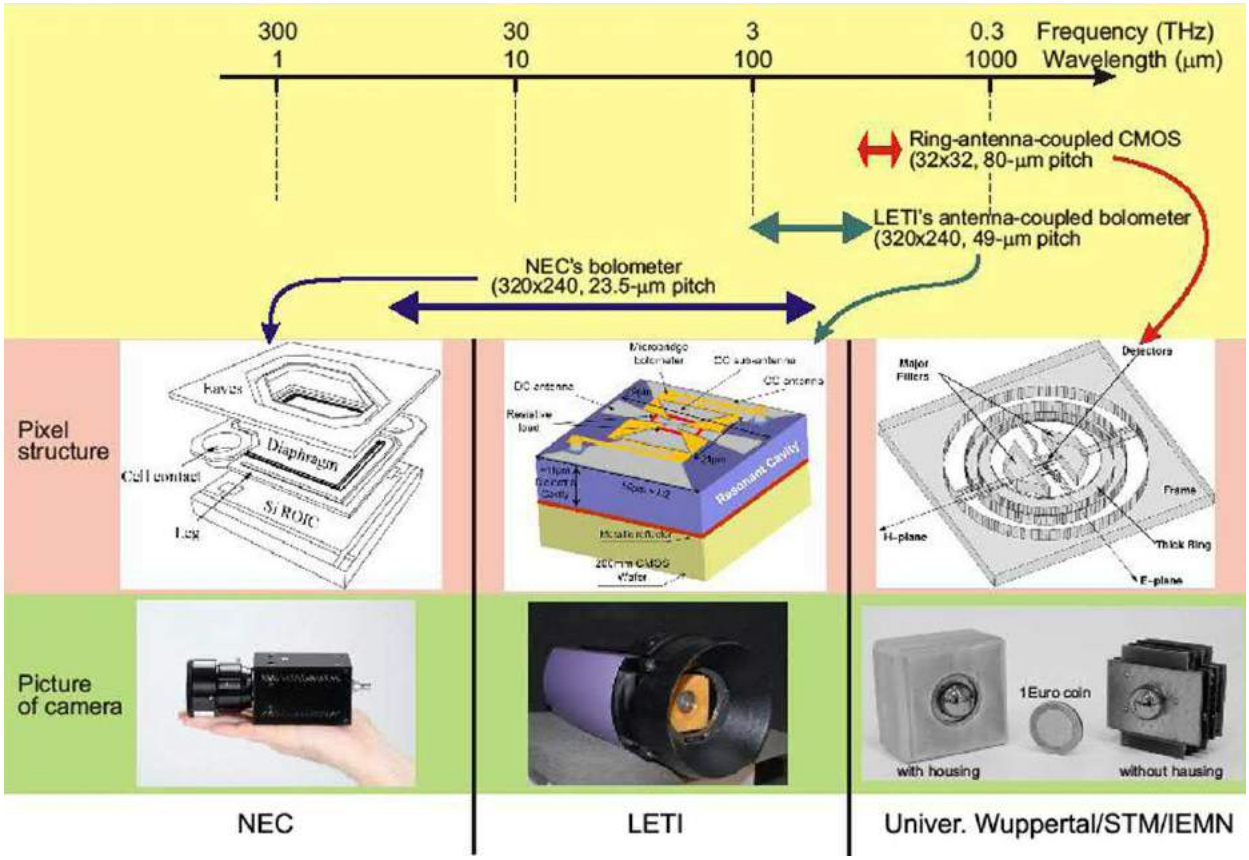


Fig. 5 Development status of uncooled THz focal plane arrays.

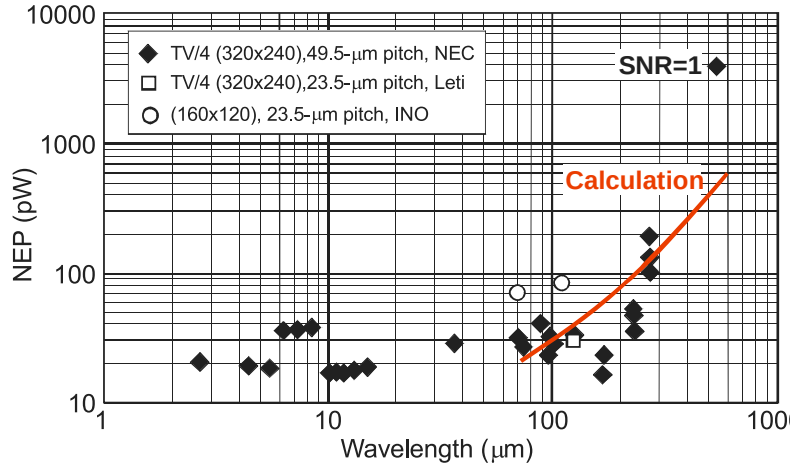


Fig. 6 Spectral dependence of *NEP* for bolometer THz FPAs.

Extrinsic Semiconductor Detectors

Research and development of extrinsic IR photodetectors have been ongoing for more than 50 years. In the 1950s and 1960s, germanium could be made purer than silicon. Today, the problems with producing pure Si have been largely solved. Si has several advantages over Ge; for example, three orders of magnitude higher impurity solubilities are attainable, hence thinner detectors with better spatial resolution can be fabricated from silicon. Si has a lower dielectric constant than Ge, and the related device technology of Si has now been more thoroughly developed, including contacting methods, surface passivation, and mature MOS and CCD technologies. Moreover, Si detectors are characterized by superior hardness in nuclear radiation environments.

For wavelengths longer than 40 μm there are no appropriate shallow dopants for silicon; therefore, germanium devices are still of interest for very long wavelengths. Germanium photoconductors have been used in a variety of infrared astronomical experiments, both airborne and space-based at wavelength ranging from 3 to more than 200 μm. Very shallow donors, such as Sb, and acceptors, such as B, In, or Ga, provide cut-off wavelengths in the region of 100 μm. **Fig. 7** shows the spectral response of the extrinsic germanium and silicon photoconductors.

Ge:Ga photoconductors are the best low background photon detectors for the wavelength range from 40 to 120 μm. Application of uniaxial stress along the [100] axis of Ge:Ga crystals reduces the Ga acceptor binding energy, extending the cutoff wavelength to ≈240 μm. At the same time, the operating temperature must be reduced to less than 2 K.

The standard planar hybrid architecture, commonly used to construct near and mid-infrared focal-plane arrays (Rogalski, 2011), is not suitable for far IR detectors where readout glow, lack of efficient heat dissipation, and thermal mismatch between the detector and the readout could potentially limit their performance. Usually, the far-infrared arrays have a modular design with many modules stacked together to form a 2-dimensional array.

The Infrared Astronomical Satellite (IRAS), the Infrared Space Observatory (ISO), and for the far-infrared channels the Spitzer-Space Telescope (Spitzer) have all used bulk germanium photoconductors. In the Spitzer mission a 32 × 32-pixel Ge:Ga unstressed array was used for the 70-μm band, while the 160 μm band had a 2 × 20 array of stressed detectors. The detectors are configured in the so-called Z-plane to indicate that the array has substantial size in the third dimension.

An innovative integral field spectrometer, called the Field Imaging Far-Infrared Line Spectrometer (FIFI-LS) was developed for the Herschel Space Observatory and SOFIA – see **Fig. 8**. To accomplish this, the instrument has two 16 × 25 Ge:Ga arrays, unstressed for the 45–110 μm range and stressed for the 110–210 μm range.

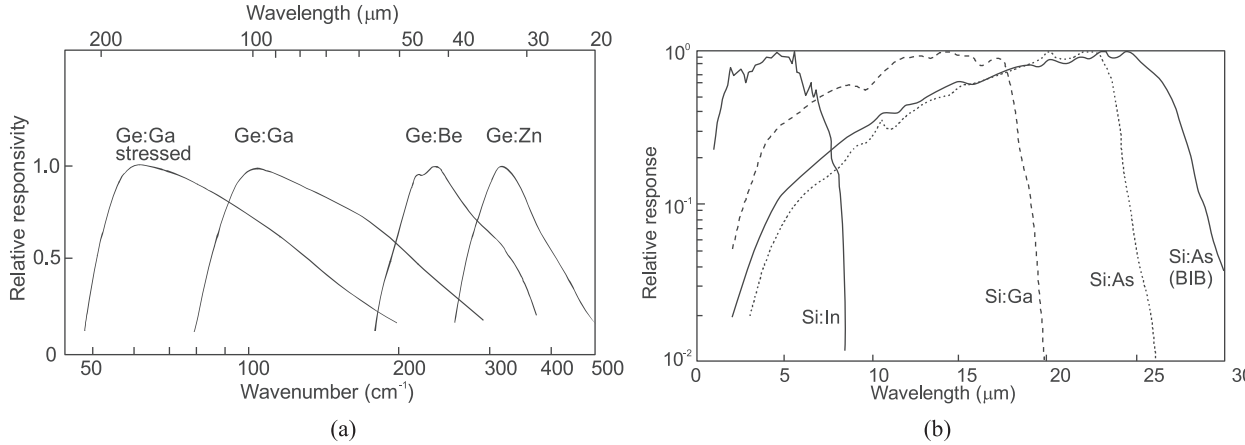


Fig. 7 Relative spectral response of some germanium (a) and silicon (b) extrinsic photoconductors. For comparison also response of Si:As BIB is shown.

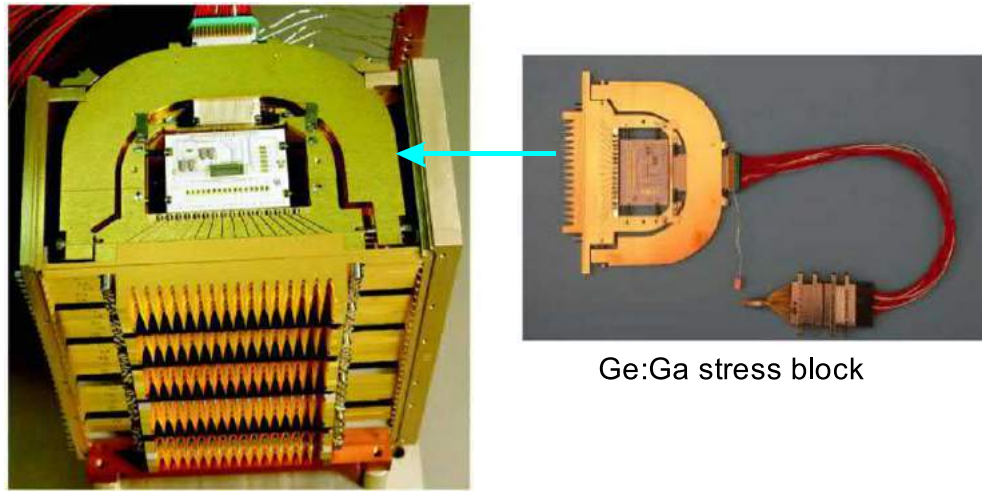


Fig. 8 PACS photoconductor focal plane array. The 25 stressed and low-stress modules of PACS instrument (corresponding to 25 spatial pixels) in the red and blue arrays are integrated into their housing. Available at: http://fifi-ls.mpg-garching.mpg.dr/detector.html.

The Photodetector Array Camera and Spectrometer (PACS) is one of the three science instruments on ESA's far infrared and sub-millimeter observatory – Herschel Space Laboratory. Apart from two Ge:Ga photoconductor arrays, it employs two filled silicon bolometer arrays with 16×32 and 32×64 pixels, respectively, to perform integral-field spectroscopy and imaging photometry in the 60–210 μm wavelength regime. Median *NEP* values are 8.9×10^{-18} W/Hz$^{1/2}$ for the stressed and 2.1×10^{-17} W/Hz$^{1/2}$ for the unstressed detectors, respectively. The detectors are operated at ~ 1.65 K. The readout electronics is integrated into the detector modules – each linear module of 16 detectors is read out by a cryogenic amplifier/multiplexer circuit in CMOS technology but operates at temperature 3–5 K.

Blocked Impurity Band Detectors

One of the major problems in the design of extrinsic photoconductors is that the doping concentration is driven by conflicting requirements: the doping concentration needs to be as high as possible to get high photon-absorption coefficients (the doping concentration is limited by hopping conduction induced by direct transfer of charge carriers from one impurity site to the next in heavily-doped semiconductors). In contrast a low doping concentration is also desirable to achieve a low electrical conductivity which in turn reduces Johnson noise.

In 1979 M. Petroff and D. Stapelbroeck, working at the Rockwell International Science Centre, invented what they referred to as Blocked-Impurity Band (BIB) detectors. These detectors were developed to provide significantly reduced nuclear radiation sensitivity and improved performance compared to extrinsically doped silicon photoconductors. BIBs have some excellent properties which make them extremely useful for astronomical applications.

BIB detectors overcome the limitation of the doping density present in a standard extrinsic photoconductor by placing a thin intrinsic (undoped) silicon blocking layer between a heavily doped active layer and a planar contact. The active region of detector structure, usually based on epitaxially grown n-type material, is sandwiched between a higher doped degenerate substrate electrode and an undoped blocking layer. Doping of the active layer is high enough for the onset of an impurity band in order to display a high quantum efficiency for impurity ionization (in the case of Si:As BIB, the active layer is doped to $\approx 5 \times 10^{17}$ cm^{-3}). The device exhibits a diode-like characteristic, except that photoexcitation of electrons takes place between the donor impurity and the conduction band.

The design of BIB detectors offers a number of advantages over conventional extrinsic photoconductors: the high absorption coefficient of the absorbing layer means that detectors with comparatively small active volumes can be made, providing low susceptibility to cosmic rays without compromising quantum efficiency. Also, due to the heavy doping of the active layer, the impurity band increases in width, therefore effectively decreasing the energy gap between the impurity band and the conduction band.

The main application of BIB arrays today is for ground- and space-based far-infrared astronomy – Si-As BIB performance are gathered in **Table 3**. The arrays should be operated under the most uniform possible conditions, in the most benign and constant environment possible. Array performance is strongly affected by background levels. Extrinsic silicon arrays for high background applications are less developed than that for low background applications.

The largest extrinsic infrared detector arrays are manufactured for astronomy owing to investments from NASA and the National Science Foundation. At present Raytheon Vision Systems (RVS), DRS Technologies, and Teledyne Imaging Sensors (formerly Rockwell Scientific Company) supply the majority of IR arrays used in astronomy, and between them the most important are BIB detector arrays. Impressive progress has been achieved especially in Si:As BIB array technology with formats as large as 2048×2048 and pixels as small as 18 μm.

Table 3 Performance of Si:As BIB FPAs fabricated in several formats for both ground and space based applications

Parameter	*Si:As BIB*	*Phenix*	*MIRI*	*Aquarius-1k*
Application/Users	Ground-based telescopes, ESO, Univer. of Tokyo	Space telescopes, JAXA	Space telescopes, JAXA, NASA	Ground-based telescopes, ESO, Univer. of Arizona
Format	320×240	1024×1024, 2048×2048	1024×1024	1024×1024
Pixel size	50 μm	25 μm	25 μm	30 μm
ROIC type	DI	SFD	SFD	SFD
Fill factor	≥95%	≥95%	≥98%	≥98%
ROIC input referred noise	$<1000e^-_{RMS}$	$6–20e^-_{RMS}$	<10–30e	Low gain $<1000e^-_{RMS}$ High gain $<100e^-_{RMS}$
Integration capacity	7 or 20×10^6 e$^-$	3×10^6 e$^-$	2×10^5 e$^-$	1 or 11×10^6 e$^-$
Max. frame rates	100–500 Hz	0.1 Hz	0.1 Hz	120 Hz
Number of outputs	16 or 32	4	4	16 or 64
Packaging	LCC	LCC	Module	Module – 2 side buttable

DI, direct injection; e$^-$, electron; LCC, leadless chip carrier; RMS, root mean square; SFD, source follower per detector.

Source: Mills, R., Beuville, E., Corrales, E., *et al*., 2011. Evolution of large format impurity band conductor focal plane arrays for astronomy applications. Proceedings of SPIE 8154, 81540R.

Semiconductor Bolometers

Cooled silicon bolometers demonstrate broadband and nearly flat spectral response in the 1–3000 μm wavelength range. They are easier to fabricate with high operability, good uniformity, and lower cost, but has low operating temperature (4.2–0.3 K). During fabrication of bolometers, their area, operating temperature, thermal time constant, and thermal conductance are adjusted to meet the specific design requirements. The present day technology exists to produce arrays of hundreds of pixels that are operated in many experiments including NASA Pathfinder ground based instruments, and balloon experiments.

The thermistors are typically fabricated by lithography on membranes of Si or SiN. The impedance is selected to a few MΩ to minimize the noise in JFET amplifiers operated at about 100 K. Limitation of this technology is assertion of thermal mechanical, and electrical interface between the bolometers at 100–300 mK and the amplifiers at ≈100 K. Usually, JFET amplifiers are sited on membranes which isolate them so effectively that the environment remains at much lower temperatures (about 10 K) – see **Fig. 9**. In addition, the equipment at 10 K is itself thermally isolated from nearby components at 0.1–0.3 K.

In bolometer metal films that can be continuous or patterned in a mesh absorb the photons. The patterning is designed to select the spectral band, to provide polarization sensitivity, or to control the throughput. Different bolometer architectures are used. In close-packed arrays and spider web, the pop-up structures or two-layer bump bonded structures are fabricated.

At present high-performance bolometer arrays for the far IR and sub-mm spectral ranges are available. For example, the Herschel/PACS instrument uses a 2048-pixel array of bolometers and is an alternative to JFET amplifiers (Mills *et al.*, 2011). The architecture of this array is vaguely similar to the direct hybrid mid-infrared arrays, where one silicon wafer is patterned with bolometers, each in the form of a silicon mesh.

Pair Braking Photon Detectors

One of the methods of photon detection consists in using superconducting materials. If the temperature is far below the transition temperature, T_c, most of electrons in them are banded into Cooper pairs. Photons with energies exceeding the binding Cooper pair energies in the superconductor, 2Δ (each electron must be supplied an energy Δ), can break these pairs producing quasiparticles (electrons) (see inset of **Fig. 10(a)**). When the bias voltage is increased to the gap voltage, the Cooper pairs on one side of the junction can break up into two quasiparticles, which then tunnel to the other side of the junction before recombining, resulting in a sharp increase in current. This process resembles the interband absorption in semiconductors, with the energy gap equal 2Δ, when the photons are absorbed and electron-hole pairs are created.

Several structures of pair braking detectors which use different ways to separate quasiparticles from Cooper pairs have been proposed. Among them are: superconductor-insulator-superconductor (SIS) and superconductor-insulator-normal metal (SIN) detectors and mixers, radio frequency (RF) kinetic inductance detectors, and superconducting quantum interference device

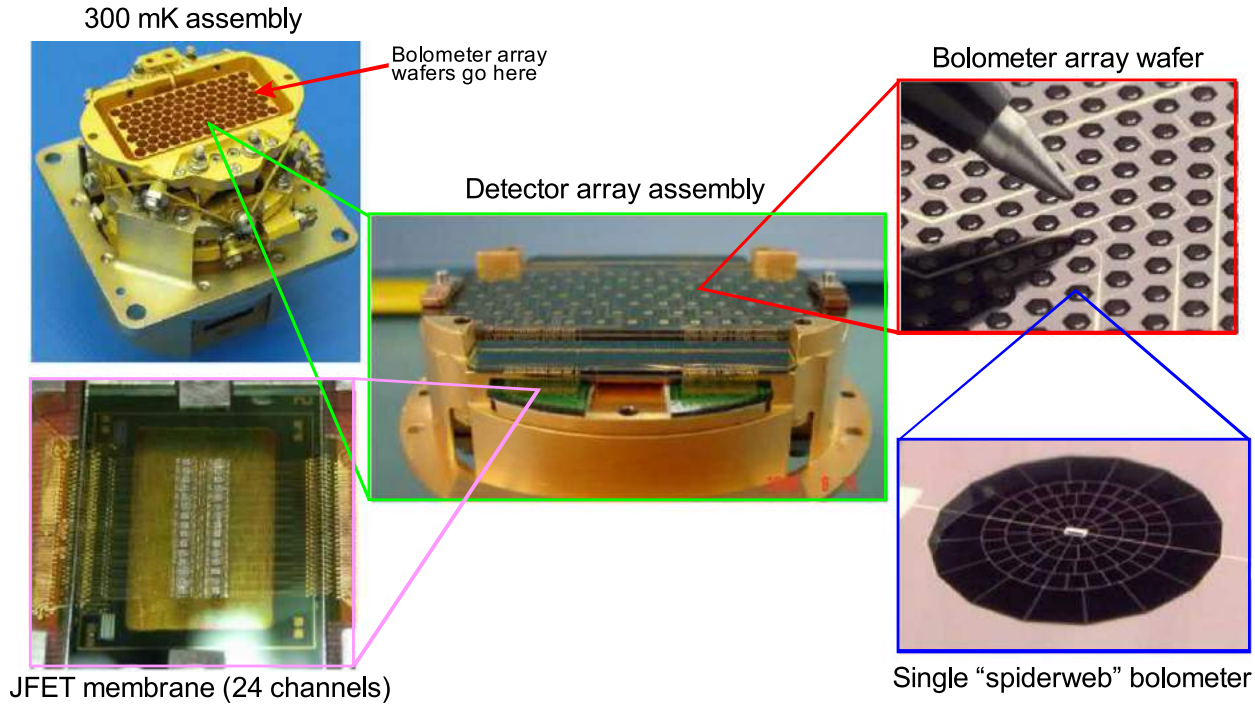


Fig. 9 Bolometer array of the Spectral and Photometric Imaging Receiver (SPIRE). Available at: http://herschel.jpl.nasa.gov/spireInstrument.shtml.

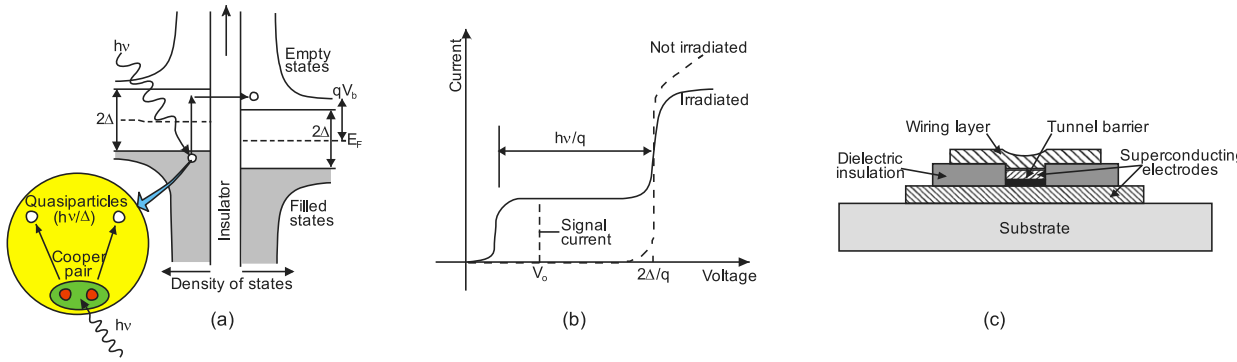


Fig. 10 SIS junction: (a) energy diagram with applied bias voltage and illustration of photon assisted tunnelling, (b) current-voltage characteristic of a non-irradiated and irradiated barrier (the intensity of the incident radiation is measured as an excess of the current at a certain bias voltage V_o – schematic creation of quasiparticle is shown inside (a)), and (c) a cross section of a typical SIS junction.

(SQUID) kinetic inductance detectors. Superconducting detectors offer many benefits: outstanding sensitivity, lithographic fabrication, and large array sizes, especially through the development of multiplexing techniques. The basics physics of these devices and progress in their developments are described in Zmuidzinas and Richards (2004).

The SIS detector is a sandwich of two superconductors separated by a thin ($\approx$20 Å) insulator, what is schematically shown in **Fig. 10(c)**. Nb and NbTiN are almost exclusively used as superconductors for the electrodes. For a standard junction process, the base electrode is 200-nm sputtered Nb, the tunnel barrier is made using a thin 5-nm sputtered Al layer which is either thermally oxidized (Al_2O_3) or plasma nitridized (AlN). The counterelectrode is 100-nm sputtered Nb or reactively sputtered NbTiN. Typical junction areas are about 1 μm^2.

SIS tunnel junctions are mainly used as mixers in heterodyne type mm and sub-mm receivers, because of their strong nonlinear I–V characteristic. They can be also used as direct detection detectors. Operating temperature of SIS junctions is below 1 K; typically $T \leq 300$ mK. However, up to now SIS detectors are difficult to integrate into large arrays.

Progress in development of large-format, high-sensitivity focal plane arrays is especially promising with two detector technologies: transition-edge superconducting (TES) bolometers (see next section) and microwave kinetic inductance detectors (MKIDs) based on different principles of superconductivity. Multiple instruments are currently in development based on arrays up to 10,000 detectors using both time-domain multiplexing (TDM) and frequency-domain multiplexing (FDM) with superconducting quantum interference devices (SQUIDs) (Bock, 2009). Both sensors show potential to realize the very low $\sim 10^{-20}$ W/$Hz^{1/2}$ sensitivity needed for space-borne spectroscopy.

A MKID is essentially a high-Q resonant circuit made out of either superconducting microwave transmission lines or a lumped element *LC* resonator (fabricated from thin aluminium and niobium films). In the first case a meandered quarter-wavelength strip of superconducting material is coupled by means of a coupling capacitance to a coplanar waveguide through-line used for excitation and readout. Lumped element are instead created from an *LC* series resonant circuit inductively coupled to a microstrip feed line placed in a high frequency resonant circuit (see **Fig. 11**). Photons hitting an MKID break Cooper pairs, which changes the surface impedance of the transmission line or inductive element producing a number of quasiparticles. This causes the resonant frequency and quality factor to shift an amount proportional to the energy deposited by the photon. The amplitude (c) and phase (d) of a microwave excitation signal sent through the resonator. The change in the surface impedance of the film following a photon absorption event pushes the resonance to lower frequency and changes its amplitude. The energy of the absorber photon can be determined from the degree of phase and amplitude shift. The readout is almost entirely at room temperature and can be highly multiplexed; in principle hundreds or even thousands of resonators could be read out on a single feedline (Day *et al.*, 2003; Mazin, 2009).

Superconducting HEB and TES Detectors

Among superconducting detectors used for terahertz downconversion, hot-electron-bolometer (HEB) mixers have attracted the attention. Their low local-oscillator (LO) power consumption (less than 1 μW), near-quantum-limited noise performance, and ease of fabrication have placed them above competing technologies in the quest for implementation of large-format heterodyne arrays.

In principle, HEB is quite similar to the transition-edge sensor (TES) bolometer, where small temperature changes caused by the absorption of incident radiation strongly influence resistance of biased sensor near its superconducting transition. The main difference between HEBs and ordinary bolometers is the speed of their response. High speed is achieved by allowing the radiation power to be directly absorbed by the electrons in the superconductor, rather than using a separate radiation absorber and allowing the energy to flow to the superconducting TES via phonons, as ordinary bolometers do. After photon absorption, a single electron initially receives the energy $h(v)$, which is rapidly shared with other electrons, producing a slight increase in the electron temperature. In the next step, the electron temperature subsequently relaxes to the bath temperature through emission of phonons.

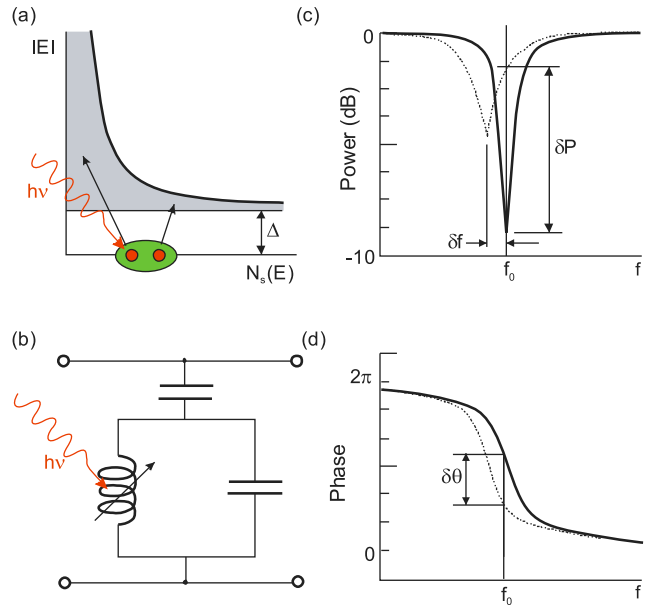


Fig. 11 An illustration of the operational principle behind a MKID.

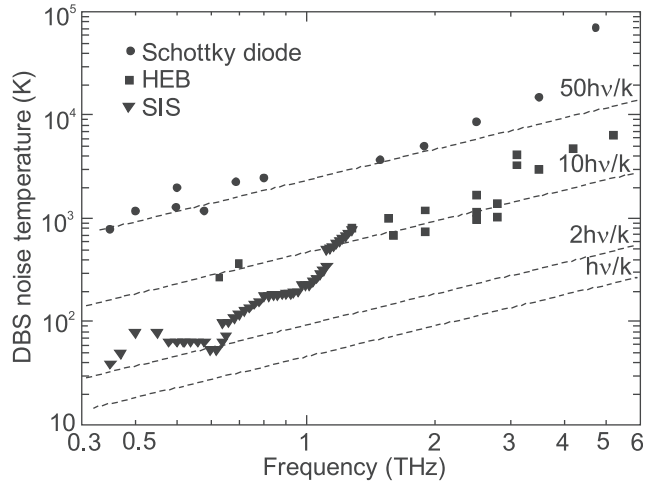


Fig. 12 DSB noise temperature of Schottky diode mixers, SIS mixers, and HEB mixers operated in terahertz spectral band. Reproduced from Hübers, H.-W., 2008. Terahertz heterodyne receivers. IEEE Journal of Selected Topics in Quantum Electronics 14, 378–391.

In comparison with TES, the thermal relaxation time of the HEB's electrons can be made fast by choosing a material with a large electron-phonon interaction. The development of superconducting HEB mixers has lead to the most sensitive systems at frequencies in the terahertz region, where the overall time constant has to be a few tens of picoseconds. These requirements can be realized with a superconducting microbridge made from NbN, NbTiN, or Nb on a dielectric substrate (Gol'tsman *et al.*, 2005).

Generally, in terahertz receivers, the noise of a mixer is quoted in terms of a single-sideband (SSB), T^{SSB}, or double-sideband (DSB), T^{DSB}, mixer noise temperature. The DSB noise temperatures achieved with Schottky diode mixers, SIS mixers, and HEB mixers operated in terahertz spectral band are presented in **Fig. 12** (Hübers, 2008). The noise temperature of SBD receivers has essentially reached a limit of about 50 $h\nu/k$ in frequency range below 3 THz. Above 3 THz, there occurs a steep increase, mainly due to increasing losses of the antenna and reduced performance of the diode itself.

Fig. 13(a) presents an example cross-section view of NbN mixer chip. About 150-nm thick Au spiral structure is connected to the contacts pads. The superconducting NbN film extends underneath the contact layer/antenna. The central area of a mixer chip shown in **Fig. 13(b)** is manufactured from a 3.5-nm thick superconducting NbN film on a high resistive Si substrate (Gol'tsman *et al.*, 2005). The active NbN film area is determined by the dimensions of the 0.2-μm gap between the gold contact pads. The NbN microstrip is integrated with a planar antenna patterned as log-periodic spiral.

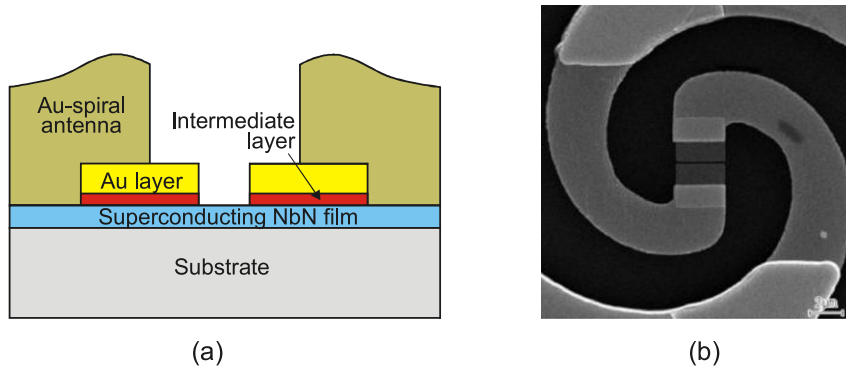


Fig. 13 NbN HEB mixer chip: (a) cross section view and (b) SEM micrograph of the central area of mixer.

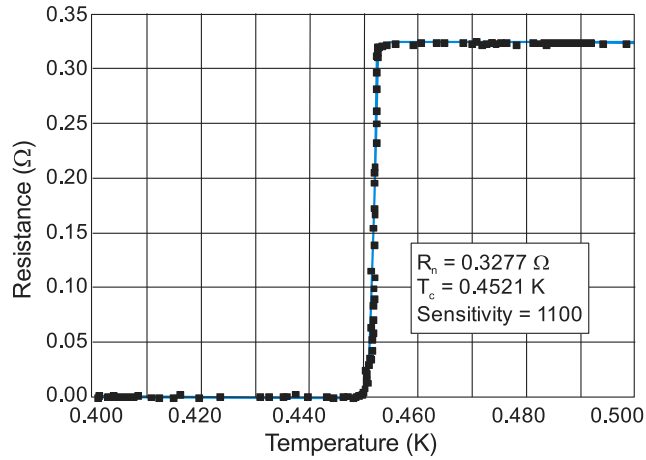


Fig. 14 Resistance vs. temperature for a high-sensitivity TES Mo/Au bilayer with superconducting transition at 444 mK.

The name of the transmission-edge sensor (TES) bolometer is derived from its thermometer, which is based on thin superconducting films held within transition region, where it change from the superconducting to the normal state over a temperature range of a few milliKelvin (see **Fig. 14**) (Benford and Moseley, 2000). Changes in temperature transition can be set by using a bilayer film consisting of a normal material and a layer of superconductor (e.g., thin Mo/Au, Mo/Cu, Ti/Au, etc.). Such design enables diffusion of the Cooper pairs from the superconductor into the normal metal and makes it weakly superconducting – this process is called the proximity effect. As a result, the transition temperature is lowered relative to that for the pure superconducting film ($T<200$ mK). Thus in principle, the TES bolometers are quite similar to the HEBs.

TES bolometers are superior to current-biased particle detectors in terms of linearity, resolution, and maximum count rate. At present, theses detectors can be applied for THz photons counting because of their high sensitivity and low thermal time constant. Membrane isolated TES bolometers are capable of reaching a phonon $NEP \approx 4 \times 10^{-20}$ W/Hz$^{1/2}$. The current generation of sub-orbital experiments largely rely on TES bolometers. Important feature of this sensor is that it can operate in wide spectral band, between the radio and gamma rays.

The most ambitious example of TES bolometer array is that used in the submillimetre camera SCUBA (Submillimetre Common-User Bolometer Array) – 2 with 10,240 pixels (SCUBA-2, 2006). The camera operated at wavelengths of 450 and 850 μm has been mounted on the James Clerk Maxwell Telescope in Hawaii. Each SCUBA-2 array is made of four side-buttable sub-arrays, each with 1280 (32×40) transition-edge sensors.

Conclusions

The THz detectors will receive increasing importance in a very diverse range of applications including detection of biological and chemical hazardous agents, explosive detection, building and airport security, radio astronomy and space research, biology and medicine. The future sensitivity improvement of THz instruments will come with the use of large format arrays with readouts in the focal plane to provide the vision demands of high resolution spectroscopy. One of promising solution is room temperature THz imaging achieved with low-cost CMOS detectors and microbolometer.

References

Al Hadi, R., Sherry, H., Grzyb, J., *et al*., 2012. A 1 k-pixel video camera for 0.7–1.1 terahertz imaging applications in 65-nm CMOS. IEEE Journal of Solid-State Circuits 47, 2999–3012.

Benford, D.J., Moseley, S.H. 2000. Superconducting transition edge sensor bolometer arrays for submillimeter astronomy. In: Proceedings of the International Sympsium on Space and THz Technology. Available at: www.eecs.umich.edu/~jeast/benford_2000_4_1.pdf.

Bock, J.J., 2009. Superconducting detector arrays for far-infrared to mm-wave astrophysics. Available at: http://cmbpol.uchicago.edu/depot/pdf/white-paper-j-bock.pdf.

Bolduc, M., Terroux, M., Tremblay, B., *et al*., 2011. Noise-equivalent power characterization of an uncooled microbolometer-based THz imaging camera. Proceedings of SPIE 8023.80230C-1–10.

Brown, E.R., Segovia-Vargas, D., 2015. Principles of THz direct detection. In: Carpintero, G., Garcia Muñoz, L.E., Hartnagel, L., Preu, S., Räisänem, A.V. (Eds.), Semiconductor Terahertz Technology: Devices and Systems at Room Temperature Operation. Wiley, pp. 212–253.

Brown, E.R., Young, A.C., Zimmerman, J., Kazemi, H., Gossard, A.C., 2006. High-sensitivity, quasi-optically-coupled semimetal-semiconductor detectors at 104 GHz. Proceedings of SPIE 6212, 621205.

Bründermann, E., Hübers, H.-W., Kimmitt, M.F., 2012. Terahertz Techniques. Heidelberg: Springer-Verlag.

Day, P., LeDuc, H.G., Mazin, B.A., Vayonakis, A., Zmuidzinas, J., 2003. A broadband superconducting detector suitable for use in large arrays. Nature 425, 817–821.

Dyakonov, M., Shur, M.S., 1993. Shallow water analogy for a ballistic field effect transistor: new mechanism of plasma wave generation by the dc current. Physical Review Letters 71, 2465–2468.

Gol'tsman, G.N., Vachtomin, Yu.B., Antipov, S.V., *et al*., 2005. NbN phonon-cooled hot-electron bolometer mixer for terahertz heterodyne receivers. Proceedings of SPIE 5727, 95–106.

Han, R., Zhang, Y., Kim, Y., *et al*., 2012. 280 GHz and 860 GHz image sensors using Schottky-barrier diodes in 0.13 μm digital CMOS. IEEE International Solid- State Circuits Conference. 254–256.

Hübers, H.-W., 2008. Terahertz heterodyne receivers. IEEE Journal of Selected Topics in Quantum Electronics 14, 378–391.

Knap, W., Coquillat, D., Dyakonova, N., *et al*., 2014. Terahertz plasma field effect transistors. In: Perenzoni, M., Paul, D.J. (Eds.), Physics and Applications of Terahertz Radiation. Dordrecht: Springer, pp. 77–100.

Knap, W., Dyakonov, M.I., 2013. Field effect transistors for terahertz applications. In: Saeedkia, D. (Ed.), Handbook of Terahertz Technology. Cambridge: Woodhead Publishing, pp. 121–155.

Lee, A.W.M., Williams, B.S., Kumar, S., Hu, Q., Reno, J.L., 2006. Real-time imaging using a 4.3-THz quantum cascade laser and a 320×240 microbolometer focal-plane array. IEEE Photonics Technology Letters 18, 1415–1417.

Mazin B.A., 2009. Microwave kinetic inductance detectors: The first decade. The Thirteenth International Workshop on Low Temperature Detectors-LTD13. AIP Conference Proceedings 1185 (1), 135–142.

Mills, R., Beuville, E., Corrales, E., *et al*., 2011. Evolution of large format impurity band conductor focal plane arrays for astronomy applications. Proceedings of SPIE 8154.81540R-1–81540R-10.

Nguyen, D.-T., Simoens, F., Ouvrier-Buffet, J.-L., Meilhan, J., Coutaz, J.-L., 2012. Broadband THz uncooled antenna-coupled microbolometer array – electromagnetic design, simulations and measurements. IEEE Transactions on Terahertz Science and Technology 2, 299–305.

Oda, N., 2010. Uncooled bolometer-type terahertz focal-plane array and camera for real-time imaging. Comptes Rendus Physique 11, 496–509.

Rogalski, A., 2011. Infrared Detectors, second ed. Boca Raton: CRC Press.

Saeedkia, D. (Ed.), 2013. Handbook of Terahertz Technology for Imaging, Sensing and Communications. Oxford: Woodhead Publishing Limited.

SCUBA-2, 2006. The Royal Observatory. Available at: http://www.roe.ac.uk/ukatc/projects/scubatwo/.

Siegel, P.H., 2002. Terahertz technology. IEEE Transactions on Microwave Theory and Techniques 50, 910–928.

Sizov, F., Rogalski, A., 2010. THz detectors. Progress in Quantum Electronics 34, 278–347.

Zhang, Z., Rajavel, R., Deelman, P., Fay, P., 2011. Sub-micro area heterojunction backward diode millimeter-wave detectors with 0.18 $pW/Hz^{1/2}$ noise equivalent power. IEEE Microwave and Wireless Components Letters 21, 267–269.

Zmuidzinas, J., Richards, P.L., 2004. Superconducting detectors and mixers for millimeter and submillimeter astrophysics. Proceedings of IEEE 92, 1597–1616.

Gravitational Wave Detection

Marie-Anne Bizouard, Paris-Saclay University, Orsay, France
Nelson Christensen, Carleton College, Northfield, MN, United States and University of Côte d'Azur, Nice, France

Introduction

In the later part of the 19th century, Albert Michelson performed extraordinary experiments that shook the foundations of the physics world. Michelson's precise determination of the speed of light was an accomplishment that takes great skill to reproduce today. Edward Morley teamed up with Michelson to measure the velocity of Earth with respect to the aether. The interferometer that they constructed was exquisite, and through amazing experimental techniques the existence of the aether was disproved. The results of Michelson and Morley led to a revolution in physics, and provided evidence that helped Albert Einstein to develop the general theory of relativity. Now the Michelson interferometer has provided dramatic confirmation of Einstein's theory of general relativity through the direct detection by the Laser Interferometer Gravitational-Wave Observatory (LIGO) of gravitational waves, and the observation of black holes (Abbott *et al.*, 2016a,b).

An accelerating electric charge produces electromagnetic radiation – light. It should come as no surprise that an accelerating mass produces gravitational light, namely, gravitational radiation (or gravitational waves). In 1888 Heinrich Hertz had the luxury to produce and detect electromagnetic radiation in his laboratory. There will be no such luck with gravitational waves because gravity is an extremely weak force.

Albert Einstein postulated the existence of gravitational waves in 1916, and Taylor and Weisberg (1989) indirectly confirmed their existence through observations of the orbital decay of the binary pulsar $1913+16$ system. The direct detection of gravitational waves has been difficult, and has literally taken decades of tedious experimental work to accomplish. The only possibility for producing detectable gravitational waves comes from extremely massive objects accelerating up to relativistic velocities. The gravitational waves that have been detected so far have come from the coalescence of binary black hole systems. For example, GW150914 was produced by the merger of a 29 $M_\odot$ black hole and a 36 $M_\odot$ black hole some 1.3×10^9 light-years away. The total energy radiated in gravitational waves was equivalent to 3 $M_\odot c^2$, with a peak luminosity of 3.6×10^{56} ergs/s.

Other possibly detectable gravitational wave sources are also astrophysical: supernovae, pulsars, neutron star binary systems, newly formed black holes, or even the Big Bang. The observation of these types of events would be extremely significant for contributing to knowledge in astrophysics and cosmology. Gravitational waves from the Big Bang would provide unique information of the universe at its earliest moments. Observations of core-collapse supernovae will yield a gravitational snapshot of these extreme cataclysmic events. Pulsars are neutron stars that can spin on their axes at frequencies up to hundreds of Hertz, and the signals from these objects will help to decipher their characteristics. Gravitational waves from the final stages of coalescing binary neutron stars could help to accurately determine the size of these objects and the equation of state of nuclear matter; they would also help to explain the mechanism that produces short gamma ray bursts. The observation of black hole formation from these binary systems, and the ringdown of the newly formed black hole as it approaches a perfectly spherical shape, would be the *coup de grâce* for the debate on black hole existence, and the ultimate triumph for general relativity.

Advanced LIGO (Aasi *et al.*, 2015) and Advanced Virgo (Acernese *et al.*, 2015) are second generation interferometric gravitational wave detectors. Initial LIGO and Virgo conducted observations from 2002 through 2010. Advanced LIGO and Advanced Virgo will ultimately have better sensitivities, by a factor of 10, over their initial designs. They will search for gravitational waves from 10 Hz up to a few kilohertz. Their target sensitivities will allow them to observe signals from the coalescence of binary neutron star systems ($1.4\ M_\odot$ – $1.4\ M_\odot$) out to distances of 200 Mpc for Advanced LIGO and 150 Mpc for Advanced Virgo. The mergers of more massive binary black holes systems will extend much farther.

Electromagnetic radiation has an electric field transverse to the direction of propagation, and a charged particle interacting with the radiation will experience a force. Similarly, gravitational waves will produce a transverse force on massive objects, a tidal force. Explained via general relativity it is more accurate to say that gravitational waves will deform the fabric of spacetime. Just like electromagnetic radiation there are two polarizations for gravitational waves. Let us imagine a linearly polarized gravitational wave propagating in the z-direction, $h(z,t)=h_{0+}e^{i(kz-\omega t)}$. The fabric of space is stretched due to the strain created by the gravitational wave. Consider a length L_0 of space along the x-axis. In the presence of the gravitational wave the length oscillates like

$$L(t)=L_0+\frac{h_{0+}L_0}{2}\cos(\omega t)$$

Hence, there is a change in its length of

$$\Delta L_x=\frac{h_{0+}L_0}{2}\cos(\omega t)$$

A similar length L_0 of the y-axis oscillates, like

$$\Delta L_y=-\frac{h_{0+}L_0}{2}\cos(\omega t)$$

 doi:10.1016/B978-0-12-803581-8.09516-3

One axis stretches, while the perpendicular one contracts, and then vice versa, as the wave propagates through. Consider the relative change of the lengths of the two axes (at $t=0$),

$$\Delta L = \Delta L_x - \Delta L_y = h_{0+} L_0$$

or

$$h_{0+} = \frac{\Delta L}{L_0}$$

So the amplitude of a gravitational wave is the amount of strain that it produces on spacetime. The other gravitational wave polarization (h_{0x}) produces a strain on axes 45 degree from (x, y). Imagine some astrophysical event produces a gravitational wave that has amplitude h_{0+} on Earth; in order to detect a small distance displacement ΔL one should have a detector that spans a large length L_0. The first gravitational wave observed by LIGO had an amplitude of $h \sim 10^{-21}$ with a frequency at peak gravitational wave strain of 150 Hz (Abbott *et al.*, 2016a). The magnitude of a gravitational wave falls off as $1/r$, so it will be impossible to observe events that are too far away. However, when the detectors' sensitivity is improved by a factor of n, the rate of signals should grow as n^3 (the increase of the observable volume of the universe). This is because the gravitational wave detectors to be discussed below measure signals from all directions; they cannot be pointed, but reside in a fixed position on the surface of the Earth.

A Michelson interferometer, with arms aligned along the x and y axes, can measure small phase differences between the light in the two arms. Therefore this type of interferometer can turn the length variations of the arms produced by a gravitational wave into changes in the interference pattern of the light exiting the system. This was the basis of the idea from which modern laser interferometric gravitational wave detectors have evolved. Imagine a gravitational wave of amplitude h is incident on an interferometer. The change in the arm length will be $\Delta L \sim h\, L_0$, so in order to optimize the sensitivity it is advantageous to make the interferometer arm length L_0 as large as possible. The Advanced LIGO and Advanced Virgo detectors will measure distance displacements that are of order $\Delta L \sim 10^{-18}$ m or smaller, much smaller than an atomic nucleus. The recent observation of gravitational waves has been one of the most spectacular accomplishments in experimental physics, and has been greeted with much excitement across the globe.

The history of the attempt to measure gravitational waves has been long. The realization that gravitational waves might be detectable crystalized as a result of the Conference on the Role of Gravitation in Physics, Chapel Hill, NC in 1957 (De Witt, 1957). Pirani (2009) had recently published a paper, and then gave presentation at the conference. He showed that the relative acceleration of particle pairs can be associated with the Riemann tensor. The interpretation of the Chapel Hill attendees was that nonzero components of the Riemann tensor were due to gravitational waves. Pirani, Richard Feynmann, and Hermann Bondi came up with the *sticky bead* argument (Bondi, 1957), essentially showing that gravitational waves exist and can be detected. Joe Weber, of the University of Maryland, was also at the Chapel Hill Conference, and from this inspiration he started to think about gravitational wave detection.

A few years after, in the early 1960s, Weber initiated the first experimental attempts to detect gravitational waves. Weber used a 1400 kg aluminum cylinder; a gravitational wave would excite the fundamental mechanical oscillation mode of the bar (Cho, 2016). The idea of using a Michelson interferometer to detect gravitational waves is almost as old as Weber's bar detector. In 1962 two Soviet physicists, Pustovoit and Gertsenshtein, noted that the use of a Michelson interferometer would be a possible means to detect gravitational waves over a frequency range that was broader than the Weber bars. In addition, the authors noted that the interferometers would have a sensitivity that would potentially be better than the Weber bars (Pustovoit and Gertsenshtein, 1963).

Also in the early 1960s, Weber and his student, Robert Forward, also considered using a Michelson interferometer to detect gravitational waves. After completing his PhD with Weber, Forward worked with Hughes Research Laboratories. It was at Hughes that Forward first constructed a Michelson interferometer to be used as a gravitational wave detector. Forward used earphones to listen to the motion of the interference signal (Forward, 1978). The engineering of signal extraction for modern interferometers is obviously far more complex.

At the time Forward was implementing an interferometric gravitational wave detector, Rainer Weiss at MIT produced a thorough investigation into not only how a Michelson interferometer could be used to detect gravitational waves, but also a systematic and comprehensive investigation into the noise sources that would constrain such a measurement (Weiss, 1972). However, as opposed to Forward's design where the laser beam traveled down the arms of the interferometer once, Weiss proposed a system where the laser beam would bounce back and forth multiple times in the interferometer, thereby increasing the effective arm length (and increasing the strain sensitivity of the detector). This is known as a Herriott optical delay line system. In what could be considered as the most important part of Weiss's (1972) presentation, he systematically listed and quantified the most important noise sources in an interferometric gravitational wave detector. These noise sources included amplitude noise on the laser source (including shot noise), frequency noise of the laser source, thermal noise in the masses and their suspension systems, radiation pressure noise from the laser light, seismic noise, noise due to residual gas in the vacuum system housing the interferometric detector, cosmic ray noise, gravitational-gradient noise, and residual electric and magnetic field noise. This comprehensive description of a realistic broadband interferometric gravitational wave detector initiated the experimental effort that has led to the present day LIGO and Virgo detectors.

Subsequently there was rapid activity on the construction of prototype laser interferometric gravitational wave detectors. The goal was to make prototypes that demonstrated the technology needed to construct kilometer-length interferometers. In the late 1970s, the Max Planck group in Garching, near Munich, Germany, created a 3 m arm length detector (Billing *et al.*, 1979), and

a 30 m detector in the early 1980s (Shoemaker *et al.*, 1988); this instrument was the first which showed a correspondence between noise models and performance in the spirit of Weiss's (1972) paper. Also, in the early 1980s Weiss and his team at MIT built laser interferometric gravitational wave detector with a Herriott optical delay line system and 1.5 m length arms. The Munich interferometers were also Herriott delay lines. The early 1980s also saw the construction of a 10 m prototype in Glasgow, Scotland (Ward *et al.*, 1985). The uniqueness of this system was that instead of the Herriott delay lines in the arms it used resonant optical cavities (Drever *et al.*, 1981, 1983), an idea by Ron Drever (then at Glasgow, before moving to Caltech). These Fabry–Perot cavities, to be discussed below, also have the light bounce back and forth a number of times, but in a resonant fashion within an optical cavity. A similar interferometer, albeit with 40 m arms, was then constructed at Caltech in the late 1980s (Spero, 1986). Fabry–Perot cavities were eventually incorporated into the design of LIGO and Virgo. The work on all of these prototypes were absolutely critical in establishing the technology needed for LIGO (Abramovici *et al.*, 1992). Similarly in the 1980s, the work on lasers, laser stabilization and interferometer optics in Orsay, France, plus the research on vibration isolation systems in Pisa, Italy, helped to create Virgo (Bradaschia *et al.*, 1990).

Numerous collaborations are building and operating second generation interferometers in order to detect gravitational waves. Advanced LIGO in the United States consists of two 4 km interferometers located in Livingston, Louisiana, and Hanford, Washington (Aasi *et al.*, 2015). Advanced LIGO started observations in 2015, and will be working over the coming years to achieve its design sensitivity, with the goal to reach it by 2019. The European Advanced Virgo is a 3 km interferometer near Pisa, Italy (Acernese *et al.*, 2015), and will start acquiring data in 2017, and will also be aiming for its target sensitivity in the coming years. GEO-600, a German-British collaboration, is a 600 m detector near Hanover, Germany (Affeldt *et al.*, 2014), and is currently operational. KAGRA is the Japanese 3 km interferometer that is presently under construction, and should commence observations in 2019 (Somiya, 2012). There will be a third 4 km LIGO interferometer, LIGO-India (Unnikrishnan, 2013), located in India, with the goal to be operational by 2024. All of the kilometer-length detectors will be attempting to detect gravitational waves with frequencies from 10 Hz up to a few kilohertz.

As will be described below, there are a number of terrestrial noise sources that will inhibit the performance of the interferometric detectors. The sensitivity of detection increases linearly with interferometer arm length, which implies that there could be advantages to constructing a gravitational wave detector in space. This is the goal of the laser interferometer space antenna (LISA) consortium. The plan is to deploy three satellites in a heliocentric orbit with a separation of about 2.5×10^6 km. LISA is a European Space Agency project, with a target launch date of 2034. LISA will observe gravitational waves in a frequency band from below 10^{-4} Hz to above 10^{-1} Hz. Due to the extremely long baseline, LISA is not strictly an interferometer, as most light will be lost as the laser beams expand, while traveling such a great distance. Instead, the phase of the received light will be detected and used to lock the phase of the light that is reemitted by another laser. Much of the technology needed for LISA to succeed was recently demonstrated with the LISA Pathfinder mission; in this mission the relative acceleration between two test masses was measured to be $5.2 \pm 0.1\,\mathrm{fm/s^2/\sqrt{Hz}}$ for frequencies between 0.7 and 20 mHz (Armano *et al.*, 2016).

Interferometer Configurations

The Michelson interferometer is the tool to be used to detect a gravitational wave. **Fig. 1** shows a basic optical setup. The beamsplitter and the end mirrors would be suspended by wires, and effectively free to move in the plane of the interferometer. The arms have lengths L_1 and L_2 that are roughly equal on a kilometer scale. With a laser power P and wavelength λ incident on the beamsplitter, the light exiting the dark port of the interferometer is

$$P_{\mathrm{out}} = P \sin^2\left[\frac{2\pi}{\lambda}(L_1 - L_2)\right]$$

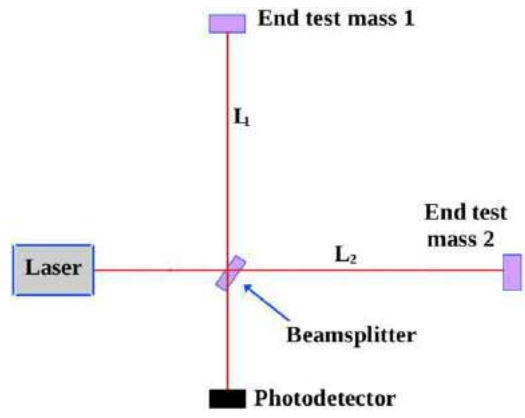


Fig. 1 A basic Michelson interferometer. The photodetector receives light exiting the dark port of the interferometer and hence the signal.

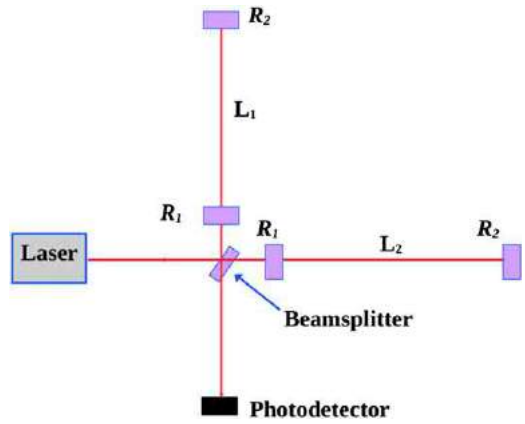


Fig. 2 A Michelson interferometer with Fabry–Perot cavities in each arm. The front cavity mirrors have reflectivity R_1, while the end mirrors have $R_2 \sim 1$. By using Fabry–Perot cavities Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) will increase the effective arm length by a factor of 140.

The interferometer operates with the condition that in the absence of excitation the light exiting the dark port is zero. This would be the case for a simple and basic interferometer. If E_0 is the amplitude of the electric field from the laser, and assuming the use of a 50–50 beamsplitter, the electric field (neglecting unimportant common phase shifts) for the light incident on the photodetector would be

$$E_{\text{out}} = E(e^{i\delta\phi_1} - e^{i\delta\phi_2}) \approx i\frac{E_0}{2}(\delta\phi_1 - \delta\phi_2) = iE_0\frac{2\pi}{\lambda}(L_1 - L_2)$$

A gravitational wave of optimal polarization normally incident upon the interferometer plane will cause one arm to decrease in length, while the other increases. The Michelson interferometer acts as a gravitational wave transducer; the stretching and squeezing of the spacetime between the mirrors results in more light exiting the interferometer dark port. The mirrors in the interferometer are suspended via fibers so that they are free to move under the influence of the gravitational wave, acting like relativistic freely falling masses.

An interferometer's sensitivity to gravitational waves increases with arm length, but geographical, physical, and financial constraints will limit the size of the arms. If there could be some way to bounce the light back and forth to increase the effective arm length it would increase the detector performance. Fabry–Perot cavities do just that. When they are on resonance they have a storage time for the light of

$$\tau_s = \frac{2L(R_1R_2)^{1/4}}{[c(1-\sqrt{R_1R_2})]}$$

where (R_1 and R_2 are power reflection coefficients). **Fig. 2** shows the system of a Michelson interferometer with Fabry–Perot cavities. This gravitational wave interferometer design was proposed in the late 1970s by Ron Drever, and subequently tested by his research group in the early 1980s (Drever *et al.*, 1981, 1983). The far mirror R_2 has a very high reflectivity ($R_2 \sim 1$) in order to ultimately direct the light back toward the beamsplitter. The front mirror reflectivity R_1 is such that LIGO's effective arm length increases from $L = 4$ km to $L \sim 560$ km. The optical properties of the mirrors of the Fabry–Perot cavities must be exquisite in order to achieve success. The mirror substrates are made via a combination of super-polishing for small scale smoothness, and then ion-beam milling for large scale uniformity. The coatings (doped tantala) for the mirrors are ion-beam sputtered, multilayer dielectrics. The mirrors for Advanced LIGO (and Advanced Virgo) were coated by Laboratoire des Matériaux Avancés (LMA, Lyon, France). Advanced LIGO's mirrors were tested, and the surface errors are between 0.08 and 0.23 nm, with absorption between 0.2 (parts per million) and 0.4 ppm. In terms of both absorbtion and scattering, the Advanced LIGO arm round-trip loss goal is less than 75 ppm. The radius of curvature for the input mirrors (R_1) is 1934 m, while for the end mirrors (R_2) it is 2245 m, with a radius of curvature spread between -1.5 and 1.0 m (Aasi *et al.*, 2015). A LIGO test mass (and therefore a Fabry–Perot mirror) can be seen in **Fig. 3**. The mirrors for Advanced Virgo have similar exquisite properties (Acernese *et al.*, 2015).

In 1888 Michelson and Morley, with their interferometer, had a sensitivity that allowed the measurement of 0.02 of a fringe, or about 0.126 rad. The Advanced LIGO interferometers during the first observing run O1 have already demonstrated a phase noise spectral density of

$$\phi(f) = 5 \times 10^{-11}\,\text{radian}/\sqrt{\text{Hz}}$$

for frequencies around 150 Hz. Assuming a 150 Hz signal with a 150 Hz bandwidth this implies a phase sensitivity of $\Delta\phi = 6.1 \times 10^{-10}$ rad. There has been quite an evolution in interferometry since Michelson's time.

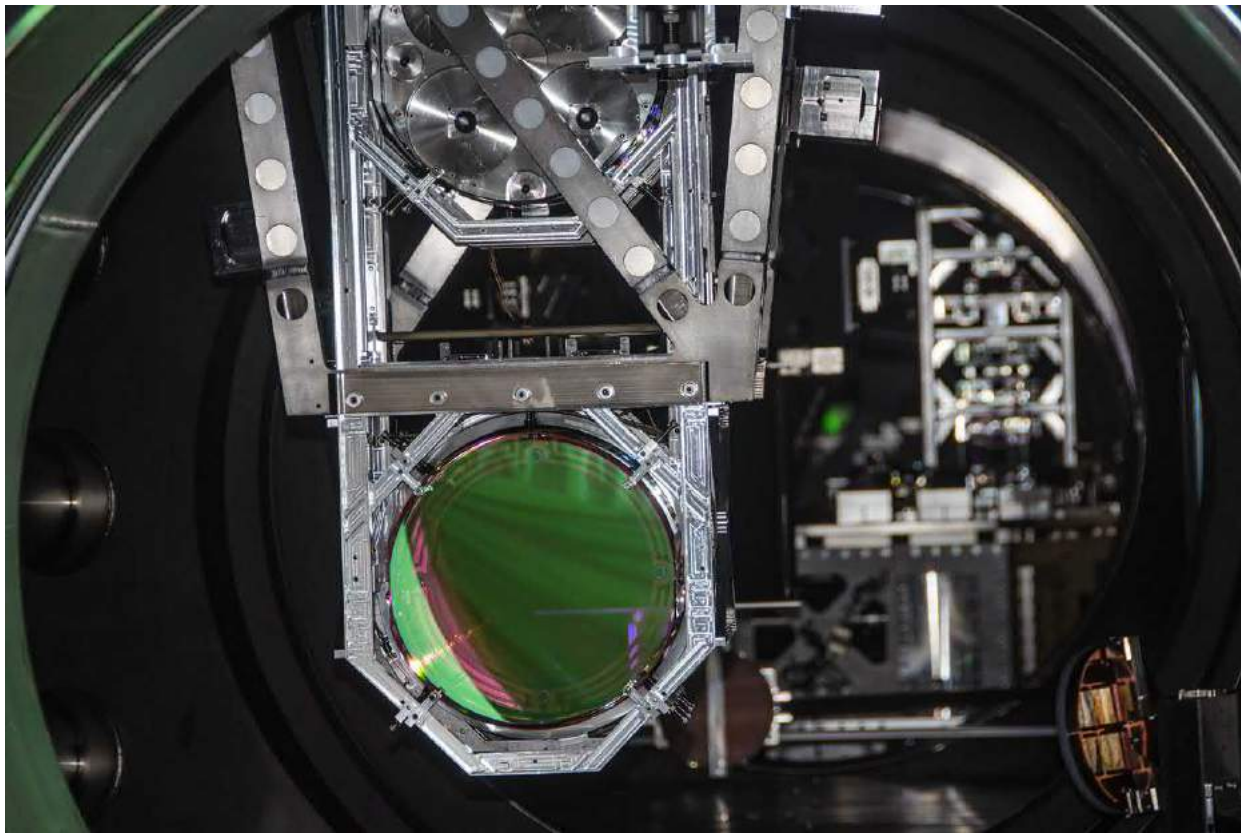


Fig. 3 A picture of the input test mass (mirror R_1) for Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) within its vibration isolation suspension system. The fused silica component is 40 kg, 34 cm in diameter, and 20 cm thick. Photograph courtesy of LIGO/Caltech/MIT.

The noise sources that inhibit the interferometer performance are discussed below. However, let us consider one's ability to measure the relative phase between the light in the two arms. The Heisenberg uncertainty relation for light with phase ϕ and photon number N is $\Delta\phi\Delta N \sim 1$. For a measurement lasting time τ using laser power P and frequency f, the photon number is $N = P\lambda\tau/hc$ (here h is Planck's constant), and with Poisson statistics describing the light $\Delta N = \sqrt{N} = \sqrt{P\lambda\tau/hc}$. Therefore

$$\Delta\phi\ \Delta N = \frac{2\pi}{\lambda}\Delta L\sqrt{P\lambda\tau/hc} = 1$$

implies that

$$\Delta L = \frac{1}{2\pi}\sqrt{hc\lambda/P\tau}$$

With more light power the interferometer can measure smaller distance displacements and achieve better sensitivity. Advanced LIGO and Advanced Virgo will use about 200 W of laser light. However, there is a nice trick one can use to produce more light circulating in the interferometer, namely power recycling (Meers, 1988). **Fig. 4** displays the power recycling interferometer design. The interferometer operates such that virtually none of the light exits the interferometer dark port, and the bulk of the light returns toward the laser. An additional mirror, R_r, in **Fig. 4**, recycles the light. The Advanced LIGO goals are to have 125 W actually impinging on the recycling mirror R_r, creating 5.2 kW upon the beamsplitter, and 750 kW within the Fabry–Perot cavities in each of the interferometer's arms. For Advanced LIGO, recycling will increase the effective light power by another factor of 42. Advanced Virgo has a similar design. The higher circulating light power therefore improves the sensitivity of these interferometric detectors. It is also interesting to note that the GEO-600 detector has been operating for years using squeezed light, namely quantum states of light, as a way to reduce the noise below the shot noise limit (Affeldt *et al.*, 2014).

There is one additional modification to the interferometer system that can further improve sensitivity, but only at a particular frequency. A further Fabry–Perot system can be made by installing what is called a signal recycling mirror (SRM); this would be mirror R_s in **Fig. 5** (Meers, 1988). Imagine the light in arm 1 of the interferometer, and that it acquires phase as the arm expands due to a gravitational wave. The traveling gravitational wave's oscillation will subsequently cause arm 1 to contract, while arm 2 expands. If the light that was in arm 1 could be sent to arm 2, while it is expanding, then the beam would acquire additional phase. This process could be repeated over and over. Mirror R_s serves this purpose, with its reflectivity defining the storage time for light in each interferometer arm. The storage time defined by the cavity formed by the SRM, R_s, and the mirror at the front of the interferometer arm cavity, R_1, determines the resonance frequency. Signal recycling will give a substantial boost to interferometer

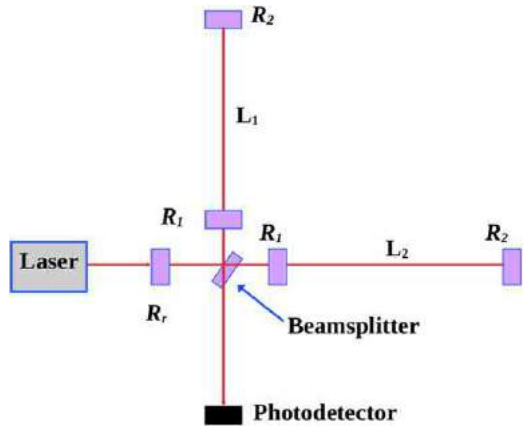


Fig. 4 A power recycled Michelson interferometer with Fabry–Perot cavities in each arm. Normally light would exit the interferometer through the light port and head back to the laser. Installation of the recycling mirror with reflectivity R_r sends the light back into the system. A Fabry–Perot cavity is formed between the recycling mirror and the first mirror (R_1) of the arms. For Advanced LIGO this strategy will increase the power circulating in the interferometer by a factor of 42.

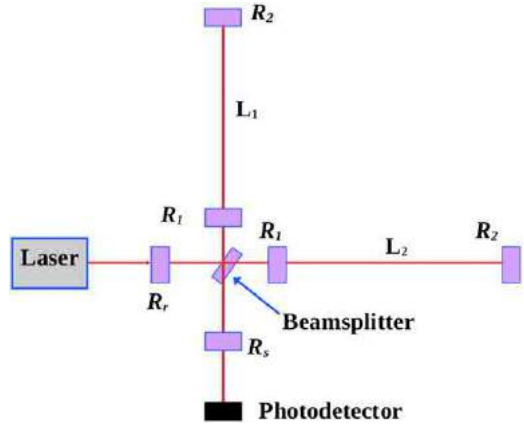


Fig. 5 A signal recycled and power recycled Michelson interferometer with Fabry–Perot cavities in each arm. Normally light containing the gravitational wave signal would exit the interferometer via the dark port and head to the photodetector. Installation of the SRM with reflectivity R_s sends the light back into the system. The phase of the light acquired from the gravitational wave will build up at a particular frequency determined by the reflectivity R_s.

sensitivity at a particular frequency, and will eventually be implemented in all the main ground-based interferometric detectors. The Advanced LIGO and Advanced Virgo interferometers are infinitely more complex than the relatively simple systems displayed in the figures of this paper.

Fig. 6(a) presents an aerial view of the LIGO site at Hanford, Washington. The magnitude of the 4 km system is apparent. Fig. 6 (b) displays the Virgo detector with its 3 km, located near Pisa, Italy.

Noise Sources and Interferometer Sensitivity

If the interferometers are to detect distance displacements less than 10^{-18} m then they must be isolated from a host of deleterious noise sources. Seismic disturbances should not shake the interferometers. Thermal excitation of components will affect the sensitivity of the detector and should be minimized. The entire interferometer must be in an adequate vacuum in order to avoid fluctuations in gas density that would cause changes in the index of refraction and hence a modification of the optical path length. The laser intensity and frequency noise must be minimized. The counting statistics of photons influences accuracy. If ever there was a detector that must avoid Murphy's law this is it; little things going wrong cannot be permitted if such small distance displacements are to be detected. The target noise sensitivity for the Advanced LIGO interferometers is displayed in Fig. 7.

In the best of all worlds the interferometer sensitivity will be limited by the counting statistics of the photons. A proper functioning laser will have its photon number described by Poisson statistics, or shot noise; if the mean number of photons arriving per unit time is N then the uncertainty is $\Delta N = \sqrt{N}$, which as noted above implies an interferometer displacement

(a)

(b)

Fig. 6 (a) Aerial view of the Laser Interferometer Gravitational-Wave Observatory (LIGO) Hanford, Washington site. The vacuum enclosure at Hanford contains the 4 km interferometer. Photograph courtesy of LIGO/Caltech/MIT. (b) Aerial view of the Virgo detector, with 3 km arms, located near Pisa, Italy. Photograph courtesy of the European Gravitational Observatory.

sensitivity of

$$\Delta L = \frac{1}{2\pi}\sqrt{\frac{hc\lambda}{P\tau}}$$

(where P is the light power impinging on the beamsplitter) or a spectral density of

$$\Delta L(f) = \frac{1}{2\pi}\sqrt{\frac{hc\lambda}{P}}$$

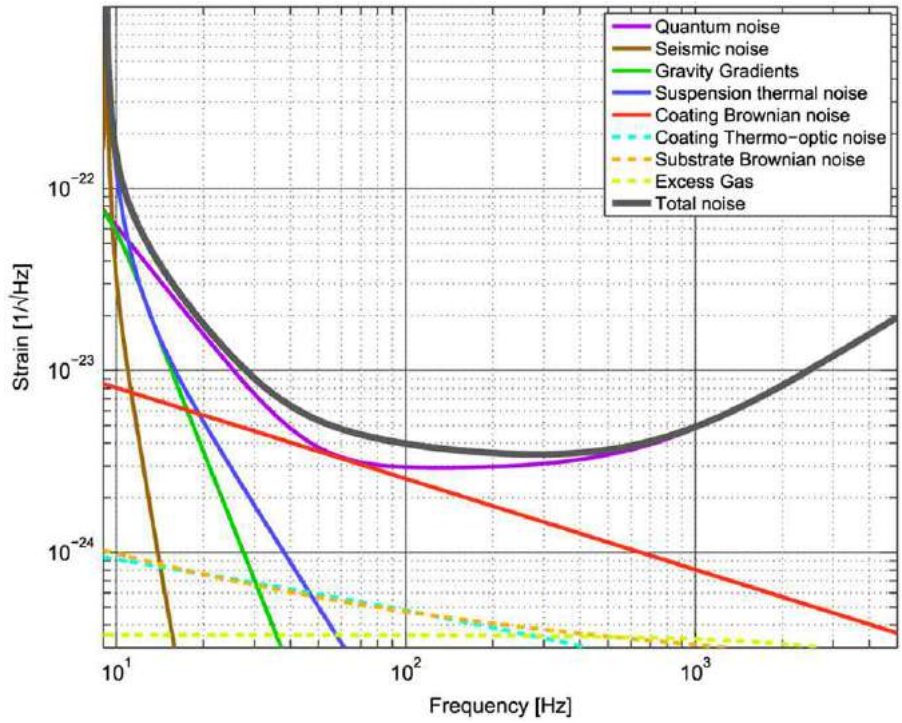


Fig. 7 The target spectral density of the noise for the Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) system. Advanced LIGO will be limited by seismic noise at low frequencies (~10 Hz), thermal noise (from the suspension system and the coatings of the mirrors) in the intermediate regime (~100 Hz). Radiation pressure will also be a dominating noise source from ~10 to ~100 Hz, while photon shot noise will be the limiting noise thereafter. Together the radiation pressure noise and the shot noise are referred to as quantum noise. Other sources of noise are also noted in the figure. This figure is from Aasi, J., Abadie, J., Abbott, B., *et al.*, 2015. Classical and Quantum Gravity 32. Available at: http://stacks.iop.org/0264-9381/32/i=7/a=074001.s; which should be consulted for a more expansive description of the limiting noise sources for Advanced LIGO.

in units of $\mathrm{m}/\sqrt{\mathrm{Hz}}$. Note also that the sensitivity increases as the light power increases. The reason for this derives from the statistics of repeated measurements. The relative lengths of the interferometer arms could be measured, once, by a photon. However, the relative positions are measured repeatedly with every photon from the laser, and the variance of the mean decreases as $\sqrt{N}$ where N is the number of measurements (or photons) involved. The uncertainty in the difference of the interferometer arm lengths is therefore inversely proportional to photon number, and hence the laser's power. In terms of strain sensitivity this would imply

$$h(f) = \frac{1}{2\pi L}\sqrt{\frac{hc\lambda}{P}}$$

This assumes the light just travels down the arm and back once. With Fabry–Perot cavities the light is stored, and the typical photon takes many trips back and forth before exiting the system. In order to maximize light power the end mirrors ($R_2 \sim 1$) and the strain sensitivity is improved to

$$h(f) = \frac{1}{4\pi\tau_s}\sqrt{\frac{\pi\hbar\lambda}{Pc}}$$

where the Fabry–Perot cavity storage time τ_s was defined above.

As the frequency of gravitational waves increases the detection sensitivity will decrease. If the gravitational wave causes the interferometer arm length to increase, then decrease, while the photons are still in the arm cavity, then the phase acquired from the gravitational wave will be washed away. This is the reason why interferometer sensitivity decreases as frequency increases, and explains the high frequency behavior seen in Fig. 7. Taking this into account, the strain sensitivity is

$$h(f) = \frac{1}{4\pi\tau_s}\sqrt{\frac{\pi\hbar\lambda}{Pc}}\left(1 + (4\pi f\tau_s)^2\right)^{1/2}$$

and f is the frequency of the gravitational wave.

If the gravitational wave is to change the interferometer arm length then the mirrors that define the arm must be free to move like freely falling masses. In systems like Advanced LIGO and Advanced Virgo, wires suspend the mirrors; each mirror is like a pendulum. The mirrors and the wires that suspend them are a monolithic-fused silica assembly, with the wires annealed and

welded to the sides of the mirrors. The pendulum itself is the first component of an elaborate vibration isolation system. Seismic noise will be troublesome for the detector at low frequencies. The spectral density of the seismic noise is about $x(f) = (10^{-9}\ m/\sqrt{\text{Hz}})(10\ \text{Hz}/f)^2$ for $f > 10$ Hz (the low frequency observational limit for Advanced LIGO and Advanced Virgo) (Saulson, 1994). A simple pendulum, by itself, acts as a motion filtering device. Above its resonance frequency a pendulum filters motion with a transfer function like $T(f) \propto (f_0/f)^2$, where f_0 is the resonant frequency for the pendulum. The mirrors for Advanced LIGO will actually be suspended by a four-stage pendulum system. The various gravitational wave detector collaborations have different vibration isolation designs. The mirrors in these interferometers are suspended in elaborate vibration isolation systems, which may include multiple pendulums, isolation stacks, and isolated optical tables. For example, for many years super-attenuators have made Virgo the most sensitive gravitational wave detector in the low frequency regime (below $\sim$40 Hz) (Acernese *et al.*, 2015). Active feedback is used on some parts of the isolation system to control seismic noise below $\sim$10 Hz. Seismic noise will be the limiting factor for interferometers seeking to detect gravitational waves in the vicinity of $\sim$10 Hz, as can be seen in the sensitivity curve presented in **Fig. 7**.

Due to the extremely small distance displacements that these systems are trying to detect it should come as no surprise that thermal noise is a problem. This noise enters through a number of components in the system. The two most serious thermal noise sources are the wires suspending the mirrors in the pendulum, and the mirrors themselves, especially the optical coatings on the mirror surfaces. Consider the wires; there are a number of notes at which they can oscillate (i.e., violin modes). At temperature T each mode will have energy of $k_B T$, but distributed over a band of frequencies determined by the quality factor (or Q) of the material. Low-loss (or high-Q) materials work best; for the violin modes of the wires there will be much noise at particular frequencies (in the hundreds of hertz). For the Advanced LIGO mirrors the first violin mode is at 510 Hz, while the vertical stretching mode of the wires is at $\sim$9 Hz.

The best sensitivity for Advanced LIGO and Advanced Virgo occurs around $\sim$100 Hz. The limiting source of noise in this region (along with quantum noise) is due to Brownian noise in the optical coatings on the mirror surfaces. There is a tremendous amount of on-going research to try and reduce the mechanical dissipation in the optical coatings. The Japanese detector KAGRA, which is currently under construction, will have its mirrors and the bottom parts of its suspension system cooled to 20K (Somiya, 2012) in order to reduce thermal noise.

The frequency noise of the laser can couple into the system to produce length displacement noise in the interferometer. With arm lengths of $\sim$4 km, it will be impossible to hold the length of the two arms absolutely equal. The slightly differing arm spans will mean that the light sent back from each of the two Fabry–Perot cavities will have slightly differing phases. As a consequence, great effort is made to stabilize the frequency of the light entering the interferometer. The Advanced LIGO laser can be seen in **Fig. 8**. The primary laser is a nonplanar ring-oscillator (NPRO). This beam is then amplified to 35 W with a medium power oscillator, and then up to 220 W with a high power oscillator; see Aasi *et al.* (2015) for more details. For Advanced LIGO, the laser is locked and held to a specific frequency by use of signals from a reference cavity, a mode cleaner cavity, and the interferometer.

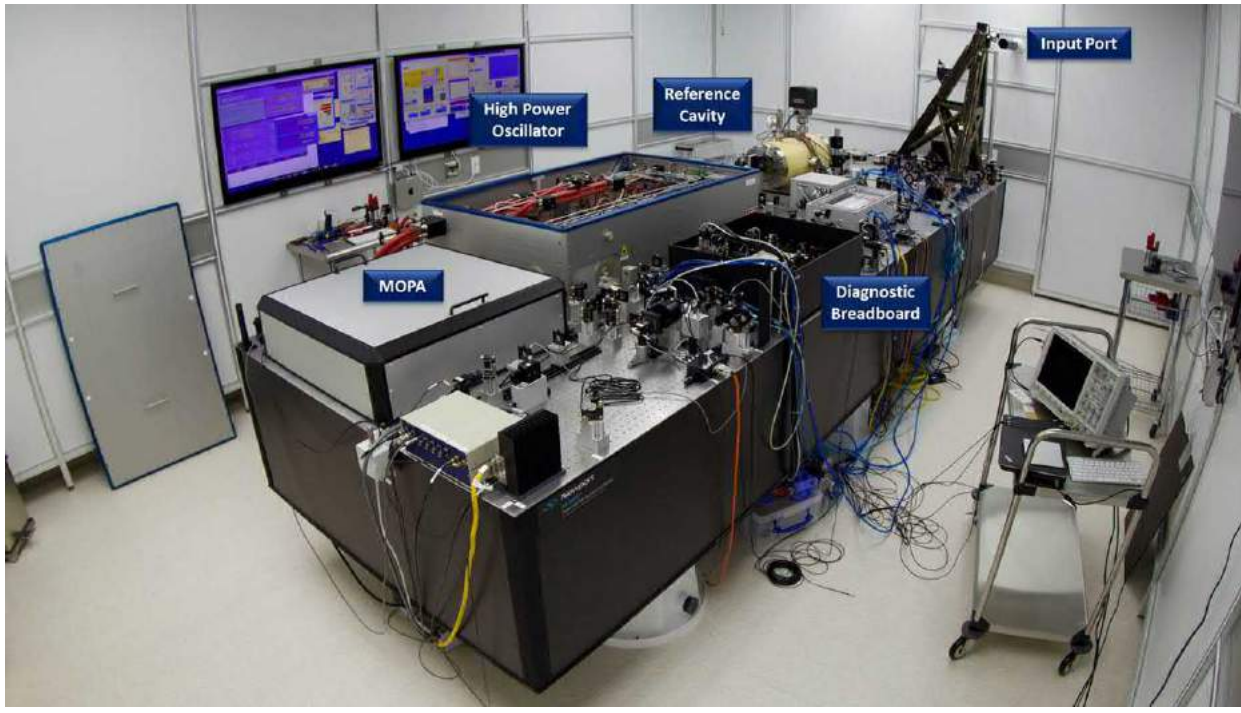


Fig. 8 The Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) laser system. This is a multistage Nd:YAG system that can deliver 200 W. Photograph courtesy of LIGO/Caltech/MIT.

For low frequency stabilization the temperature of the NPRO is adjusted. At intermediate frequencies adjustment is made by signals to a piezoelectric transducer within the NPRO cavity. At high frequencies the noise is reduced with the use of an electro-optic crystal. The Advanced LIGO lasers currently have a frequency noise of $1 \times 10^{-6}\text{Hz}/\sqrt{\text{Hz}}$ at 100 Hz; this requirement is needed by Advanced Virgo too.

It is also important to worry about the stability of the laser power for the interferometric detectors. The goal is to be quantum noise limited at frequencies within the observational frequency band. The Nd:YAG power amplifiers used are pumped with an array of laser diodes, so the light power is controlled through feedback to the laser diodes. The Advanced LIGO requirements for the fluctuations on the power P are $\Delta P/P < 2 \times 10^{-9}/\sqrt{\text{Hz}}$ at 10 Hz; Advanced Virgo's requirements are similar. In these laser interferometric gravitational wave detectors, the spatial quality of the light is ensured through the use of an input mode cleaning cavity. Advanced LIGO uses an isosceles triangular array of mirrors with the two base mirrors separated by 0.465 m and the third mirror displaced by 16.24 m. The length of the input mode cleaner for Advanced Virgo is 143.424 m. The optical system for Advanced LIGO is displayed in **Fig. 9**. Aside from the laser and the phase modulator, the entire optical system is an ultra-high vacuum. Note that at the output of the interferometer there is a SRM. Given the reflectivity of the mirror, and the phase of the light when arriving at there, it is possible to enhance the gravitational wave signal at a particular frequency by feeding it back into the interferometer for enhancement. The output mode cleaner is used to improve the spatial quality of the output beam, and remove modulation frequency sidebands, before the photodetection.

The target light powers for Advanced LIGO are displayed in **Fig. 9**. When Advanced LIGO attains its target sensitivity there will be 750 kW within the Fabry–Perot cavities. Advanced Virgo's light powers will be similar. With such a large amount of power the

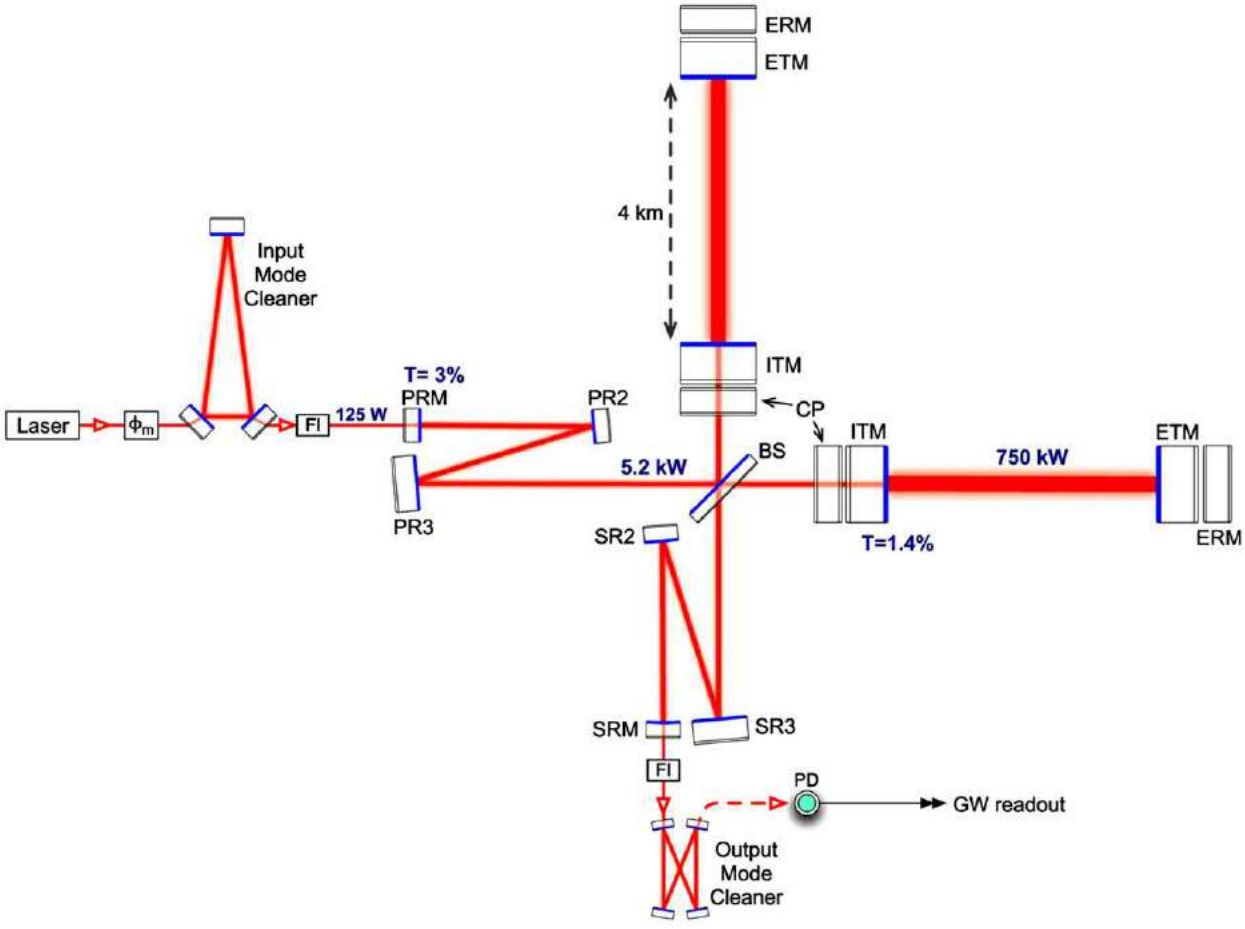


Fig. 9 The Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) optical system. The laser (~200 W) light propagates from the stabilized laser through a phase modulator (ϕ_m) to the input mode cleaner, then through a Faraday isolator (FI) to the power recycling mirror (PRM). The folding mirrors, PR2 and PR3, direct the light to the beamsplittler (BS) input of the interferometer. Note that approximately 125 W of light impinges upon the power recycling mirror, resulting in 5.2 kW at the input port to the beamsplitter. The Fabry–Perot cavities are formed with the input test mass (ITM), which is coupled to a compensation plate (CP), and the end test mass (ETM) which is coupled to a end reaction mass (ERM). The Fabry–Perot cavities will contain 750 kW of light power. Note too that the output signal from the interferometer can itself be recycled and amplified at specific frequencies, dependent on the reflectivity of the SRM; SR2 and SR3 are folding mirrors. The output beam also has its spatial features cleaned with the output mode cleaner before the light falls upon a photodetector (PD). The figure is from Aasi, J., Abadie, J., Abbott, B., *et al.*, 2015. Classical and Quantum Gravity 32. Available at: http://stacks.iop.org/0264-9381/32/i=7/a=074001.s.

mirrors will actually get heated and lightly change their shape. Ring heaters encircle the input test masses and the end test masses, and are used to correct the shape of the masses. A CO_2 laser beam is also sent onto the surface of the compensation plate to provide further corrections to the thermal lens of the input test mass. The compensation plate serves as a reaction mass for the input test mass in its isolation suspension system; the same is true for the end reaction mass with respect to the end test mass.

Parametric instabilities are another consequence of high power operation for Advanced LIGO and Advanced Virgo. This is an interaction between a mechanical oscillation mode of the mirror and higher order optical modes via light scattering. This can be a nonlinear process and can prevent the interferometers from operating at higher powers. Advanced LIGO has observed parametric instabilities (Evans *et al.*, 2015). To eliminate the excited modes one can heat the mirror to slightly change its shape, thereby changing the mechanical oscillation mode with respect to an excited optical mode. Advanced LIGO uses electrostatic actuators to move the masses; the actuators can also be used to damp excited mechanical modes.

The immense number of photons, coupled with the fact that the photon arrival times are random (Poisson statistics) means that radiation reaction noise will be important. This can be seen in the low frequency component of the quantum noise in **Fig. 7**. The low frequency radiation reaction noise, plus the high frequency shot noise, combine to create the total quantum noise in the interferometer. At high frequencies the shot noise decreases with laser power, while at low frequencies radiation reaction noise and the basic interferometer's quantum noise is a tradeoff between these two effects. While this quantum noise seems to be an unavoidable noise source, quantum states of light (namely squeezed states of light) can reduce this noise. This reduction of noise was demonstrated by initial LIGO (Aasi *et al.*, 2013) where squeezed light reduced the noise below the shot noise level for frequencies above 150 Hz; for higher frequencies a 2.15 dB (28%) reduction in the shot noise was observed. The GEO-600 gravitational wave detector now uses squeezed light continuously, and has achieved an impressive 3.7 dB reduction in the shot noise level (Affeldt *et al.*, 2014). The use of quantum states of light is one of the ways that Virgo and LIGO hope to reduce their noise in the years to come.

Gravitational Wave Detection GW150914

On September 14, 2015, at 09:50:45 UTC a gravitational wave was detected directly for the first time. The gravitational wave was first observed at the LIGO Livingston Observatory (Louisiana), and then 7 ms later at the LIGO Hanford Observations (Washington). An on-line signal search algorithm identified the signal in 3 min. An off-line examination of the data using a template-based search for compact binary coalescence signals identified the gravitational wave with a signal-to-noise ratio of 24

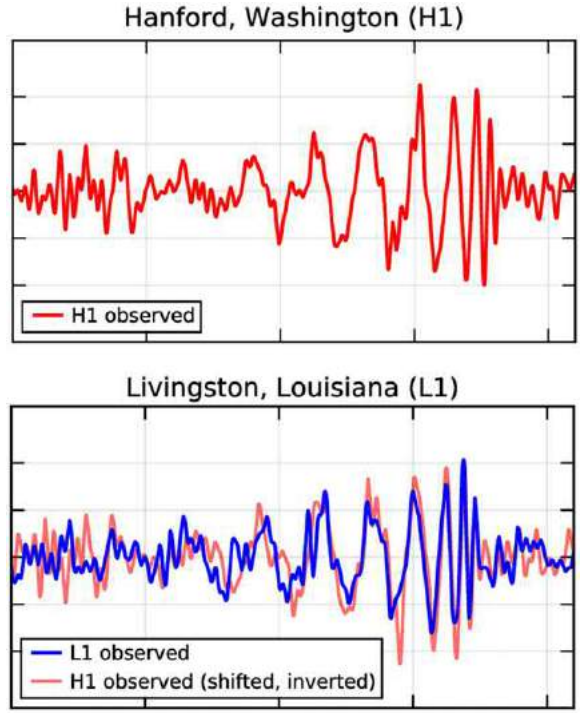


Fig. 10 The measured gravitational wave signal GW150914 as observed at the two Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) interferometric detectors. The data has been bandpass filtered (35–350 Hz), and the gravitational wave signal is clearly observable by eye. Top: the signal as observed from the LIGO Hanford detector. Bottom: the signal as observed from the LIGO Livingston detector (blue). In addition, the Hanford signal (red) is superimposed after it has been displaced by 7 ms and inverted (due to the relative orientation of the two detectors). The similarity of the two measured signals is clearly visible. Figures courtesy of the LIGO Open Science Center (losc.ligo.org).

(Abbott *et al.*, 2016a). Parameter estimation routines were used to determine that the gravitational wave signal was emitted from the merger of two black holes with masses of 36 $M_\odot$ and 29 $M_\odot$. The newly created black hole had a mass of 62 $M_\odot$, meaning that the total energy of gravitational wave emitted was equivalent to 3 $M_\odot c^2$. The system was 1.3 billions light-years away from us when it merged.

The measured gravitational wave signal, GW150914, from the two LIGO detectors is displayed in **Fig. 10** (Abbott *et al.*, 2016a). The peak amplitude of GW150914 is $h \sim 10^{-21}$ which corresponds to a displacement of the interferometers' arms of $\Delta L \sim 2 \times 10^{-18}$ m. The exquisite sensitivity of these interferometers can be seen from these numbers. In addition to GW150914, during Advanced LIGO's first observing run two other gravitational wave events were observed (also stellar mass binary black hole mergers) (Abbott *et al.*, 2016b,c).

Conclusions

The observation of gravitational waves using Michelson interferometers testifies to the utility of these devices, and to the scientists that have made them work so well. More than a hundred years ago Michelson succeeded in carrying off experiments of amazing difficulty as he measured the speed of light and disproved the existence of the aether. Gravitational wave detection is an experiment worthy of Michelson. In addition, the detection of gravitational waves a century after their prediction by Albert Einstein, and his observation that they will never be observed, testifies to the tremendous progress that has been made in technology, notably here in optics. A new window into the universe has been created, gravitational wave astronomy.

LIGO, Virgo, GEO, and KAGRA are creating a new type of telescope to peer into the heavens. With every new means of looking at the sky there has come unexpected discoveries. This has started with the unexpected observation of gravitational waves produced by binary black hole systems with tens of solar masses. Physicists do know that there will be other signals that they can predict: binary systems containing neutron stars, for example. It is suspected that short gamma ray bursts come from the coalescence of binary neutron stars, or neutron star – black hole binary systems. A core-collapse supernova will produce a burst of gravitational waves that will hopefully rise above the noise. Pulsars, or neutron stars spinning about their axes at rates sometimes exceeding hundreds of revolutions per second, will produce continuous sinusoidal signals that can be seen by integrating for sufficient lengths of time. Gravitational waves produced by the Big Bang will produce a background stochastic noise that can possibly be extracted by correlating the outputs from two or more detectors. These are exciting physics results that will come through tremendous experimental effort. The exciting initial observations of gravitational waves have been made, but it is just the beginning of a new astronomy.

Acknowledgements

NC is supported by National Science Foundation grant PHY-1505373. This article has been assigned LIGO Document number P1700043.

References

Aasi, J., Abadie, J., Abbott, B., *et al.*, 2015. Classical and Quantum Gravity 32, 074001. Available at: http://stacks.iop.org/0264-9381/32/i=7/a=074001.s.

Aasi, J., Abadie, J., Abbott, B., *et al.*, 2013. Enhanced sensitivity of the LIGO gravitational wave detector by using squeezed states of light. Nature Photonics 7, 613–619.

Abbott, B.P., Abbott, R., Abbott, T.D., *et al.*, LIGO Scientific Collaboration and Virgo Collaboration, 2016a. Physical Review Letters 116 (6), 061102. Available at: http://link.aps.org/doi/10.1103/PhysRevLett.116.061102.

Abbott, B.P., Abbott, R., Abbott, T.D., *et al.*, LIGO Scientific Collaboration and Virgo Collaboration, 2016b. Physical Review Letters 116 (24), 241103. Available at: http://link.aps.org/doi/10.1103/PhysRevLett.116.241103.

Abbott, B.P., Abbott, R., Abbott, T.D., *et al.*, LIGO Scientific Collaboration and Virgo Collaboration, 2016c. Physical Review X 6 (4), 041015. Available at: https://link.aps.org/doi/10.1103/PhysRevX.6.041015.

Abramovici, A., Althouse, W.E., Drever, R.W.P., *et al.*, 1992. Science, 256. pp. 325–333. Available at: http://science.sciencemag.org/content/256/5055/325; ISSN: 0036-8075.

Acernese, F., Agathos, M., Agatsuma, K., *et al.*, 2015. Classical and Quantum Gravity 32, 024001. Available at: http://stacks.iop.org/0264-9381/32/i=2/a=024001.

Affeldt, C., Danzmann, K., Dooley, K.L., *et al.*, 2014. Classical and Quantum Gravity 31, 224002. Available at: http://stacks.iop.org/0264-9381/31/i=22/a=224002.

Armano, M., Audley, H., Auger, G., *et al.*, 2016. Physical Review Letters 116 (23), 231101. Available at: http://link.aps.org/doi/10.1103/PhysRevLett.116.231101.

Billing, H., Maischberger, K., Rudiger, A., *et al.*, 1979. Journal of Physics E: Scientific Instruments 12, 1043. Available at: http://stacks.iop.org/0022-3735/12/i=11/a=010.

Bondi, H., 1957. Nature 179, 1072. Available at: http://dx.doi.org/10.1038/1791072a0.

Bradaschia, C., Fabbro, R.D., Virgilio, A.D., *et al.*, 1990. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 289, 518–525. Available at: http://www.sciencedirect.com/science/article/pii/016890029091525G; ISSN: 0168-9002.

Cho, A., 2016.Science. doi:10.1126/science.aaf4057.

De Witt, C.M. (Ed.), 1957 Proceedings: Conference on the Role of Gravitation in Physics, Chapel Hill, NC.

Drever, R.W.P., Ford, G.M., Hough, J., *et al.*, 1983. A gravity-wave detector using optical cavity sensing. In: Schmutzer, E. (Ed.), Proceedings of the Ninth International Conference on General Relativity and Gravitation, Jena, July 1980, pp. 265–267.

Drever, R.W.P., Hough, J., Munley, A.J., *et al.*, 1981. Optical Cavity Laser Interferometers for Gravitational Wave Detection. Berlin; Heidelberg: Springer, pp. 33–40. Available at: http://dx.doi.org/10.1007/978-3-540-38804-3_4; ISBN: 978-3-540-38804-3.

Evans, M., Gras, S., Fritschel, P., *et al.*, 2015. Physical Review Letters 114 (16), 161102. Available at: http://link.aps.org/doi/10.1103/PhysRevLett.114.161102.

Forward, R.L., 1978. Physical Review D 17 (2), 379–390. Available at: http://link.aps.org/doi/10.1103/PhysRevD.17.379.

Meers, B.J., 1988. Physical Review D 38 (8), 2317–2326. Available at: https://link.aps.org/doi/10.1103/PhysRevD.38.2317.
Pirani, F.A.E., 2009. Republication of: On the physical significance of the Riemann tensor. General Relativity and Gravitation 41, 1215–1232.
Pustovoit, V., Gertsenshtein, M., 1963.Soviet Physics JETP-USSR 16, 433–435.
Saulson, P., 1994. Fundamentals of Interferometric Gravitational Wave Detectors. Available at: https://books.google.fr/books?id=F87sCgAAQBAJ; ISBN: 9789814501903.
Shoemaker, D., Schilling, R., Schnupp, L., *et al.*, 1988. Physical Review D 38 (2), 423–432. Available at: https://link.aps.org/doi/10.1103/PhysRevD.38.423.
Somiya, K., 2012. Classical and Quantum Gravity 29, 124007. Available at: http://stacks.iop.org/0264-9381/29/i=12/a=124007.
Spero, R., 1986. The Caltech laser-interferometric gravitational wave detector. In: Ruffini, R., (Ed.), Fourth Marcel Grossmann Meeting on General Relativity, pp. 615–620.
Taylor, J., Weisberg, J., 1989. Further experimental tests of relativistic gravity using the binary pulsar PSR 1913 + 16. Astrophysical Journal 345, 434–450.
Unnikrishnan, C.S., 2013. International Journal of Modern Physics D 22, 1341010. Available at: http://www.worldscientific.com/doi/abs/10.1142/S0218271813410101.
Ward, H., Hough, J., Newton, G.P., *et al.*, 1985. IEEE Transactions on Instrumentation and Measurement IM-34, pp. 261–265. ISSN: 0018-9456.
Weiss, R., 1972. Electromagnetically coupled broadband gravitational antenna. Technical Reports MIT Quarterly Report of the Research Laboratory for Electronics. Available at: https://dcc.ligo.org/LIGO-P720002/public/main.

Cavity QED Effects in Molecular Systems

Stéphane Kéna-Cohen, Polytechnique Montréal, Department of Engineering Physics, Montreal, QC, Canada

Introduction

Organic thin films have found widespread use in display technology. The ease with which these materials can be processed and the versatility afforded by molecular design have been two of the main catalysts for the spread of organic light-emitting diode (OLED) technology. These characteristics have also made organic molecules attractive platforms for studying light–matter interaction in various regimes. The nature of optical excitations in molecular materials depends greatly on the type and structure of the molecular system. The underlying building block can range from small molecule, to polymer, to macromolecules such as proteins. In an isolated molecule, the optical absorption energy differs from the energy difference between the lowest unoccupied molecular orbital and highest occupied molecular orbital by the electron–hole pair (or exciton) binding energy. Furthermore, when strong intermolecular interactions are present, such as in certain polymers or crystalline aggregates, energy levels are broadened into bands and wavefunctions can become delocalized over many molecules. In this both of these limits, cavity QED effects have been exploited to modify the absorption and emission properties of organic molecules. For example, they can be used to increase the fluorescence efficiency of organic molecules by several orders of magnitude, to tailor the angular emission of molecules in organic thin films and to create hybrid light-matter modes called exciton-polaritons where the cavity photon and molecular excited state become intrinsically linked.

Modification of Molecular Spontaneous Emission

The possibility to modify molecular fluorescence by changing the optical environment was beautifully demonstrated in a series of pioneering experiments by Drexhage *et al*. where the radiative decay rate of fluorophore was measured as a function of its distance from a metal surface. In the quantum theory, the change in rate can be understood to occur from a modification by the metal of the local density of optical states (LDOS) at the emitter position. In the electric dipole approximation, Fermi's Golden rule predicts a radiative rate:

$$\gamma = \frac{\pi\omega_0 p^2}{3\hbar\varepsilon_0}\rho(\mathbf{r_0}, \omega_0)$$

where ω_0 is the emitter transition frequency, $\boldsymbol{p} = p\hat{n}$ is its transition dipole moment and $\rho(\mathbf{r_0}, \omega_0)$ the density of optical states evaluated at the emitter position $\mathbf{r_0}$. The LDOS consists of a sum of the normalized field intensities, performed over all of the normal modes supported by the optical environment. The form of the LDOS will favor emission into normal modes with large field intensities at the emitter position and be inhibited in those with vanishing intensities.

Classically, one can understand the change in radiative rate as being due to the action of the radiation field scattered by the environment back onto the molecular dipole. To quantitatively explain Drexhage's experiments, Chance, Prock and Silbey introduced the powerful Dyadic Green's function approach. The Dyadic Green's function, $\mathbf{G}(\mathbf{r}, \mathbf{r}')$, a tensor, can be used to express the electric field, $\mathbf{E}(\mathbf{r})$, resulting from a given current density, $\mathbf{J}(\mathbf{r})$, using:

$$\mathbf{E}(\mathbf{r}) = i\omega\mu\mu_0 \int_V \mathbf{G}(\mathbf{r}, \mathbf{r}')\mathbf{J}(\mathbf{r}')\mathbf{d}^3\mathbf{r}'$$

It is obtained by solving the inhomogeneous wave equation within the studied structure with the appropriate boundary conditions. Then, for a molecule that is modeled as a localized oscillating current density (a point dipole), the electric field generated by this dipole is simply given by:

$$\mathbf{E}(r) = \mu\mu_0\omega^2\mathbf{G}(\mathbf{r}, \mathbf{r}_0)\cdot\mathbf{p}$$

Given that the rate of energy dissipation due to the oscillating charge density is:

$$W = -\frac{1}{2}\int_V Re\{\mathbf{J}^*\cdot\mathbf{E}\}\mathbf{d}^3\mathbf{r}$$

we find that the energy dissipation rate of the dipole is:

$$\dot{W} = \frac{\omega}{2}\mathrm{Im}\{\mathbf{p}^*\cdot\mathbf{E}(\mathbf{r}_0)\} \propto \hat{n}\cdot\mathrm{Im}\{\mathbf{G}(\mathbf{r}_0, \mathbf{r}_0)\}\cdot\hat{n}$$

where $\hat{n}$ is a unit vector along the dipole. The radiative decay rate is thus be related to the imaginary part of the Dyadic Green's function evaluated at the dipole position. Physically, this term contains two contributions: one due to the typical dipole radiation and a second due to the scattered radiation field. In particular, this term can be directly related to the LDOS:

$$\rho(\mathbf{r}_0) = \frac{6\omega_0}{\pi c^2}\hat{n}\cdot \mathrm{Im}\{\mathbf{G}(\mathbf{r}_0,\mathbf{r}_0)\}\cdot\hat{n}$$

Note that changes in the optical environment also lead to frequency shifts of the emitted radiation, similar to the Lamb shift, but these tend to be very small.

Molecules in Plasmonic Resonators

Plasmonic resonators are particularly interesting for enhancing the fluorescence or phosphorescence rates of molecules given the large field intensities that can be obtained within a small mode volume. Although mode volume is implicitly included in the definition of the LDOS, its role can be explicitly highlighted using Purcell's formula. In plasmonic resonators, several unique features must be considered that are usually absent in dielectric structures. Firstly, the excitation of modes with field components in the metal leads to losses due to Joule heating. These non-radiative losses, which occur with a rate γ_{nr}, can be significant and should not be confused with the intrinsic non-radiative (e.g., vibrational) loss rate of the molecule, γ_{nr}^{mol}. Rather, part of the radiation emitted by the dipole will be quenched by the metal. As a result, emitters coupled to plasmonic cavities will always suffer from (often significantly) less than unity quantum efficiencies. Secondly, the open nature of these systems can lead to significant enhancement of the excitation fields, which also lead to enhanced fluorescence. Finally, as for Fabry–Pérot optical cavities, plasmonic resonators intrinsically change the emitter radiation pattern and this change can significantly modify the measured fluorescence enhancement, which will depend on the experimental configuration. The interplay between fluorescence enhancement and quenching was highlighted in an experiment by Anger *et al*. where single molecule fluorescence was studied as a function of distance from a spherical gold nanoparticle in a geometry where this distance could be actively tuned. The results are shown in Fig. 1, where it can be seen that the fluorescence rate, which is a function of both the increased excitation field and the modified quantum efficiency reaches a maximum enhancement of ~ 7 at a distance 5 nm from the metal surface, beyond which it drops rapidly due to increased quenching.

The enhancements in fluorescence rate can be dramatic for molecules with a low intrinsic quantum yield. In the presence of a resonator, the modified quantum efficiency is $\eta = \gamma_r/\left(\gamma_r + \gamma_{nr} + \gamma_{nr}^{mol}\right) \simeq \gamma_r/\gamma$, where $\gamma_r \equiv \gamma - \gamma_{nr}$. The approximation holds whenever the Purcell factor is large enough to dominate any internal loss. The Purcell factor is defined as $F_p = \gamma/\gamma_0$, where γ_0 is the radiative rate without the presence of the antenna. By using a dye with $\eta = 2.5\%$ within the gap of a bowtie antenna, the Moerner group were able to demonstrate a thousand-fold fluorescence rate enhancement. In particular the structure reduced the fluorescence lifetime to below 10 ps, as compared to the molecule's intrinsic lifetime of 275 ps. In this work, η was increased to a maximum of 25% and limited by the fraction of emission leading to Joule heating.

One of the most interesting proposals for achieving large Purcell factors, while maintaining high quantum yield is the use of optical patch antennas proposed by Esteban *et al*. These can maintain quantum efficiencies of nearly 50%, combined with Purcell

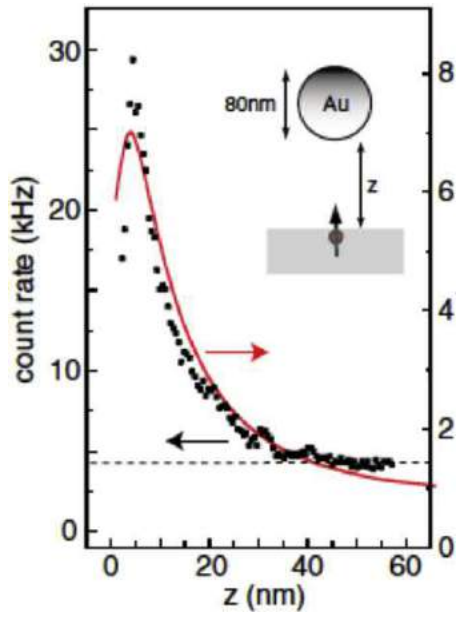


Fig. 1 Fluorescence count rate as a function of distance between an 80 nm-diameter gold nanoparticle and a single molecule of nile blue in a 2-nm-thick PMMA layer. The increase in local density of optical states (LDOS) near the nanoparticle leads to an increased count rate as it approaches the molecule. Very close to the nanoparticle, however, radiation occurs mostly in non-radiative modes, which leads to a quenching of the molecular emission. Reproduced from Anger, P., Bharadwaj, P., Novotny, L., *et al.*, 2006. Enhancement and quenching of single-molecule fluorescence. Physical Review Letters 96, 113002.

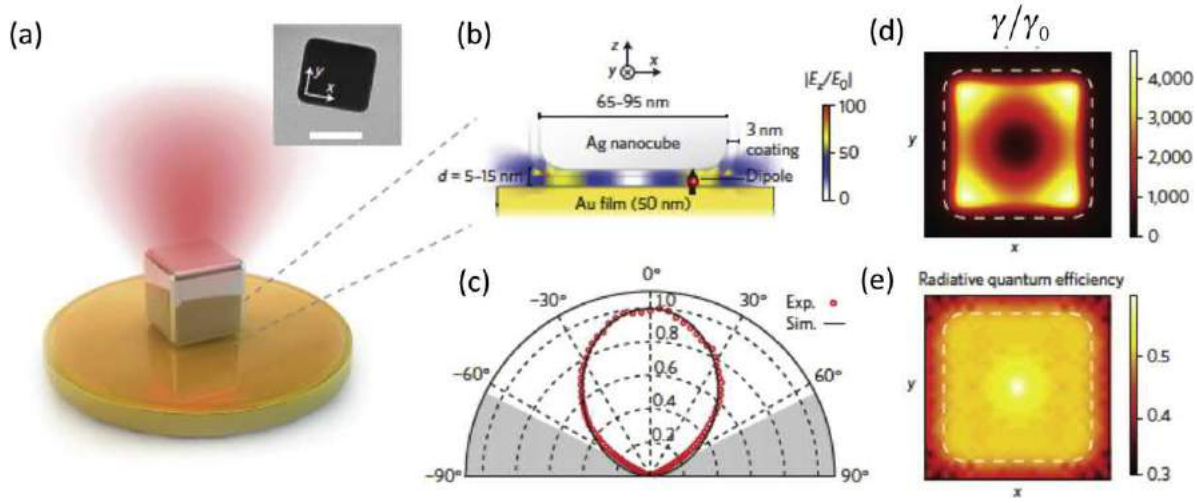


Fig. 2 (a) Schematic of the nanocube patch antenna from Akselrod *et al.* (the inset shows a transmission electron microscope image of a nanocube). (b) The electric field profile in the *z*-direction is shown for the fundamental plasmonic gap mode. (c) The experimental and simulated radiation profiles are shown. These show a maximum at normal incidence from the substrate, which is useful for out-coupling. (d) The calculated Purcell factor as a function of position for a vertically oriented dipole, which shows rate enhancements as large as 4000. The Purcell factor is closely related to field profile, which is peaked in the corners of the cube. (e) The radiative quantum efficiency for a vertically oriented dipole as a function of position. The quantum efficiency is remarkably high, staying above 50% for nearly all dipole positions. Reproduced from Akselrod, G.M., Argyropoulos, C., Hoang, T.B., *et al.*, 2014. Probing the mechanisms of large Purcell enhancement in plasmonic nanoantennas. Nature Photonics 8, 835.

factors of approaching a thousand. The structures consist of a small metaling disk (the patch) placed tens of nanometers above a much larger disk (the ground plane). The behavior of these structures can be understood by considering the modal structure of the infinite metal-insulator-metal waveguide. Such a waveguide supports a strongly confined mode in the gap, which does not exhibit a cutoff as a function of gap thickness. As a result, relatively large Purcell effects are possible, but this mode does not radiate. The wavelength scale patch serves to create discrete resonances as a function of patch width. These serve to further increases the Purcell factor, while allowing for efficient radiation out of the structure. A variant of this structure, consisting of a metallic nanocube (the patch) above an infinite metal plane was used by Akselrod *et al.* to demonstrate $F_p > 2000$ for vertically oriented dipoles, while maintaining a high quantum yield and directionality. The calculated field profile, Purcell factors and radiative quantum efficiencies of these antennas are shown in **Fig. 2**.

Dramatic changes in radiative decay rates can have important consequences for the photophysical processes within a molecule. Enderlein first considered the possible role of the Purcell effect on the photobleaching of organic molecules. Namely if photobleaching occurs from the excited state with a given rate, efficient radiation will reduce the time spent in the excited state and thus reduce the probability of a photobleaching. This can then dramatically increase the number of photons emitted before photobleaching. For many molecules, the photophysics are slightly different from those initially proposed by Enderlein, but the outcome is the same. In practice, photobleaching often occurs from the triplet state following intersystem crossing from the excited singlet state. For large Purcell factors, the triplet yield after optical excitation can be dramatically decreased due to the much faster radiative pathway, which can now compete effectively with intersystem crossing.

Cavity Effects in OLEDs

The typical OLED architecture consists of several organic layers deposited on top of an indium tin oxide (ITO) anode, which sits atop a glass substrate. Organic layer thicknesses tend to range between 100 and 150 nm and the structures are typically capped with an aluminum cathode. It is not obvious that cavity effects need to be carefully accounted for in such an innocuous structure. There is no obvious cavity and the emissive layers are typically located 50 nm away from the metal cathode. Saito *et al.* used the approach of Chance *et al.* to calculate the decay rate of an emitter in proximity to the cathode and found that significant reductions in lifetime could be observed for distances up to 150 nm away. This reduction is principally due to the excitation of surface plasmon-polaritons at the metal-organic interface by the dipole's near field components. Bulovic *et al.* highlighted the importance of dipole orientation on the angular emission profile and polarization behavior of OLEDs. In particular, they found that the presence of the metal cathode is sufficient to lead to weak microcavity effects and as a result, small changes in organic layer thicknesses are sufficient to shift the observed emission wavelength by ~15 nm. They also confirmed experimentally the polarization splitting in the emission, which occurs at high angle and the predominance of TM-polarized radiation.

Hobson *et al.* presented a thorough analysis of light extraction in an OLED geometry, by using the method of Chance, Prock and Silbey. They calculated the power dissipated via the different possible pathways: radiation into air, into the glass substrate,

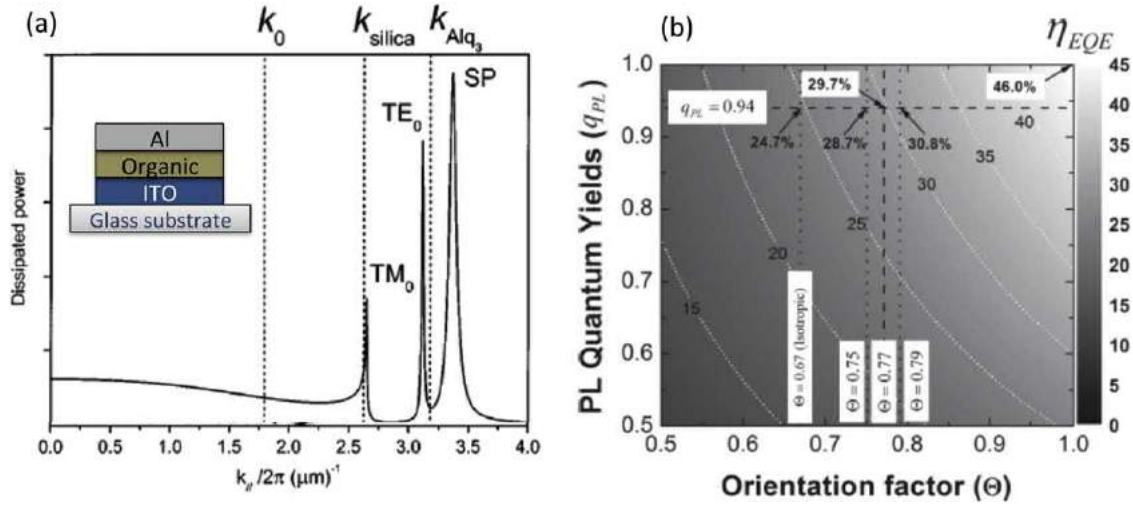


Fig. 3 (a) Power dissipation spectrum as a function of in-plane momentum $k_{//}$ for a 100-nm-thick model small-molecule organic light-emitting diode (OLED) with randomly oriented dipoles. The emitters are located 50 nm away from the cathode. The critical wavevectors for air, silica and the organic Alq_3 are indicated above the figure. A schematic OLED structure is shown in the inset. The fraction of power dissipated into a given mode is given integrating under each peak and the power waveguided in the substrate corresponds to that between k_0 and k_{silica}. Reproduced from Hobson, P.A., Wasey, J.A.E., Sage, I., *et al.*, 2002. The role of surface plasmons in organic light-emitting diodes. IEEE Journal of Selected Topics in Quantum Electronics 8 (2), 378. (b) Contour plot of the calculated external quantum efficiency (EQE) for a realistic OLED as a function of the dipole orientation factor and intrinsic quantum yield of the emitter. The dashed lines correspond measured values for $Ir(ppy)_2(acac)$, which has an intrinsic quantum yield of 0.94 and a horizontal to vertical dipole ratio of 0.77:0.23 ($\Theta = 0.77$). ITO, indium tin oxide. Reproduced from Kim, S.-Y., Jeong, W.-I., Mayr, C., *et al.*, 2013. Organic light-emitting diodes with 30% external quantum efficiency based on a horizontally oriented emitter. Advanced Functional Materials 23, 3896.

into waveguided modes or into surface plasmon modes as shown in Fig. 3(a) for the case of a model 100-nm-thick small-molecule OLED. Importantly, they correctly found that near-field coupling to surface plasmons are a considerable source of loss even for emitters located more than 50 nm away from the cathode. For polymer-based systems, coupling to the surface plasmon mode is typically weaker than for isotropically oriented small molecules, because the transition dipole moments are predominantly in-plane and cannot radiate efficiently in the TM-polarized surface plasmon mode.

A large number of studies have since focused on strategies to out-couple waveguided light from the substrate and active layers, such as using microlens arrays, corrugated structures, scattering films or low-index grids. Recently, however, the importance of molecular orientation has been reemphasized. In particular, the Kim and Brütting groups have studied the importance of molecular orientation on out-coupling efficiency. Fig. 3(b) shows the maximum theoretical external quantum efficiency (EQE) of a state-of-the-art green phosphorescent OLED as a function of dipole orientation factor and PL quantum efficiency of the emitter. One can see that the maximum out-coupling efficiency, and concomitant EQE for an ideal isotropic emitter tends to be around ~25%. However, this value can be increased to 46% for horizontally oriented dipoles. As a result, several small molecule guest–host combinations have since been identified that can lead to external quantum efficiencies surpassing 30% due to a natural tendency for their transition dipole moments to align in-plane.

Finally, the ability to spectrally control the emission in microcavity OLEDs where the ITO has been replaced or complemented by a thin metal anode has further advantages. Such microcavities have been used to increase the emitted light in the forward direction and more importantly, to tune the emission wavelength or increase spectral purity. For example, Sony's "Super Top Emission" architecture introduced this strategy in combination with a single white OLED design to produce full color displays.

Strong Light–Matter Coupling

The results of the previous sections, which were based on Fermi's golden rule, were obtained by treating light-matter interaction – in those cases, within the electric dipole approximation – using perturbation theory. This approach accurately predicts the absorption and emission dynamics of molecules at short times compared to the Rabi frequency of the combined light-matter system. In practice, this approach tends to give the correct results because of strong dissipation or short coherence times that damp any long-time oscillatory behavior (Rabi oscillations). In optical cavities where only a few discrete modes interact with the molecules, this approach can fail. Interaction of the molecule with the vacuum electromagnetic field occurs at a rate given by the so-called vacuum Rabi frequency, Ω. In a uniformly filled cavity with ideal mirrors:

$$\hbar\Omega = 2\sqrt{\frac{Ne^2 f}{4m\varepsilon_0\varepsilon_b}}$$

where N is the molecular number density, f is the oscillator strength, and ε_b is the background dielectric constant. There is no explicit dependence on mode volume because of the use of molecular density. The corresponding Rabi energy $\hbar\Omega$ can be as high as ~ 1 eV, which easily exceeds the typical photon and exciton linewidths (or equivalently dissipation rates) in organic microcavities. As a result, it is relatively simple to reach the strong light–matter coupling regime. The explicit factor of 2 emphasizes the fact that this definition corresponds to the energy splitting of new light–matter eigenmodes of the system: the lower and upper exciton-polaritons (LP and UP). For each in-plane wavevector, LP and UPs are formed, which are coherent superpositions of the molecular excitons (that with the correct wavevector) and the corresponding cavity photon mode. Their dispersion relation, which is shown in **Fig. 4(a)**, is inherited mostly from the typical Fabry–Perot dispersion and shows typical anticrossing behavior around the bare exciton resonance. For planar structures, this dispersion can be readily measured as a function of angle given the correspondence between angle of incidence and in-plane wavevector. This contrasts strongly with bulk exciton-polaritons, which can also exist in molecular crystals at low-temperature. Lidzey *et al*. were the first to observe the formation of polaritons in an organic microcavity at room-temperature using a porphyrin dye in a host matrix.

Because of the large number of molecules as compared to the number of photon modes supported in such structures, there exists several eigenstates of the molecule-photon Hamiltonian that are not coupled to the photon mode: the so-called dark states. These play an important role in dictating the dynamics under nonresonant excitation. Polaritonic modes can also be modified by the presence of structural and energetic disorder, effects that were considered analytically by Agranovich *et al*. and numerically by Michetti and La Rocca. The overarching conclusion is that in the photonic parts of the polariton branches, disorder does not cause a strong deviation from the ideal polariton mode, but very close to $k=0$ and in the flat (exciton-like) regions of the polariton branches, disorder causes a strong spatial localization of the modes.

As for inorganic polaritons, a strong motivation for studying organic microcavities in the strong-coupling regime has been to realize low-threshold sources of coherent light: polariton lasers. Indeed, if a large population of polaritons can be accumulated near the energy minimum at $k=0$, Bose stimulation will favor further population of this state. When scattering into this state exceeds the loss rate, a polariton condensate will form. At this point, the state becomes macroscopically occupied and spontaneously acquires a phase, which corresponds to a breaking of U(1) symmetry. Because photons emitted through the cavity mirrors preserve both the phase and momentum of the underlying polaritons, this results in both spatially and temporally coherent emission. **Fig. 4(b)** shows an angle-resolved contour plot of the luminescence emitted from a strongly-coupled organic microcavity pumped below and above the condensation threshold. Above threshold, the luminescence is emitted from the bottom of the LP and a small blueshift can be observed due to polariton nonlinearities.

The detailed mechanisms for relaxation from a high-energy molecular excited state created by a laser pump to the minimum of the LP are still under investigation. Measurements show that this relaxation occurs principally through the dark states. Although the decay of polaritons occurs on a rapid timescale given by the polariton lifetime (typically 0.01–1 ps), time-resolved photoluminescence measurements show a decay on the timescale of the reservoir lifetime. This strongly suggests that these are the initially excited states from which polaritons must then be populated. Various mechanisms have been proposed to explain relaxation from

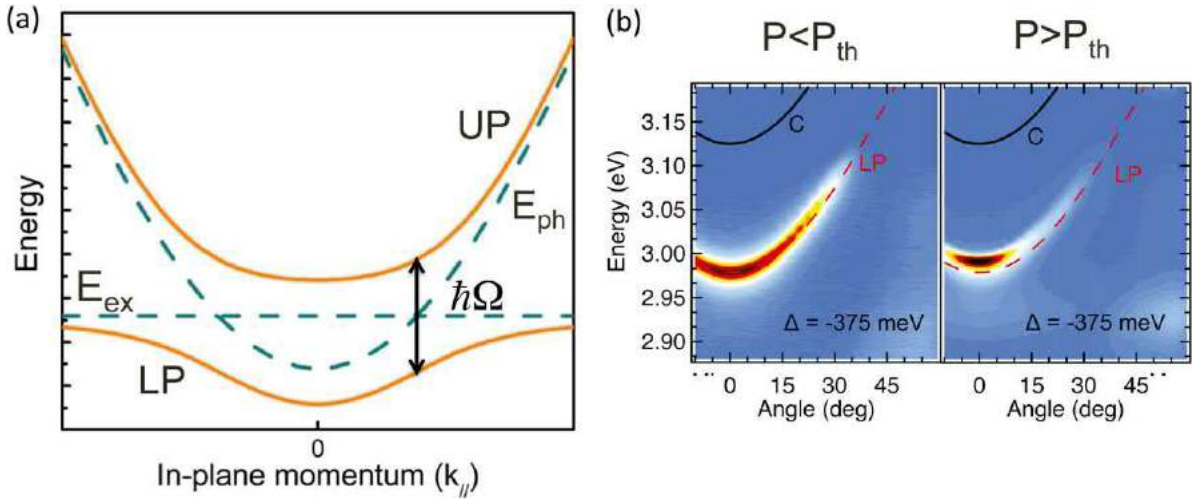


Fig. 4 (a) The dispersion relation of upper and lower polaritons (UP and LP) is shown as solid lines. The dispersion relations of the bare exciton (E_{ex}) and cavity photon (E_{ph}) are shown as dashed lines. The polariton branches show typical anticrossing behavior around the exciton energy. The minimum splitting occurs at the degeneracy point where both branches are split by the Rabi energy. (b) Contour plot of the angle-resolved photoluminescence from a strongly-coupled microcavity pumped below and above the condensation threshold. Below threshold the emission intensity decreases monotonically when going up from the minimum of the LP branch. Above threshold, all of the emission comes from the bottom of the LP and a small blueshift can be observed. The color scales are normalized. The measured irradiance above threshold is approximately 10^4 times larger than that below threshold. Modified from Daskalakis, K.S., Maier, S.A., Murray, R., *et al.*, 2014. Nonlinear interactions in an organic polariton condensate. Nature Materials 13, 271.

the dark states to polaritonic states including the emission of localized molecular vibrations, pure dephasing mechanisms due to low-energy vibrations and radiative emission through uncoupled vibrational modes of the ground electronic state.

The realization of cavities with low-enough dissipation rates have allowed for polariton condensates to be demonstrated in microcavities containing organic single-crystals (Kéna-Cohen *et al.*), disordered organic and polymer films (Daskalakis *et al.* and Plümhof *et al.*) and macromolecules (Dietrich *et al.*). One particularly interesting feature of microcavities in the strong coupling regime is the nonlinearity due which arises due to a depletion of the ground sate. This nonlinearity gives us rise to effective repulsive interactions between polaritons and between polaritons and dark excitons. The first has been used to demonstrate room-temperature superfluidity of polaritons, while the second gives rise to modulation instabilities and can be used to create artificial potential barriers using a pump laser. Until now, the nonlinearities observed in organic microcavities remain much weaker than those in inorganics and thus require much larger densities of polaritons to show dramatic effects.

Another fascinating possibility is that of using organic microcavities in the strong coupling regime to modify molecular properties or chemical reaction rates. For example, the Ebbesen group has shown a significantly reduced rate of photo-isomerization for a photochromic dye in a strongly coupled microcavity compared to the out-of-cavity case. Indeed, theoretical work by Galego *et al.* has shown how the modification of potential energy surfaces in the strongly coupled regime can indeed suppress photochemical reactions due to a stabilization of the molecular structure. More recently, strong coupling of optically active vibrational modes within planar optical cavities has also been demonstrated. This approach is particularly promising for the future control of chemical reaction pathways as it could allow for bond-specific targeting.

Conclusions

This article reviewed selected applications of cavity QED to solid-state molecular systems. We saw how in the weak coupling regime, optical resonators allow for the control of the spontaneous emission rate and radiation pattern. These effects have important applications for the realization of fluorescent markers, OLEDs and single photon sources. In the strong coupling regime, hybrid light–matter eigenstates-the polaritons-form with their own unique properties. These possess strong nonlinearities and have applications for the realization of low-threshold organic lasers and the control molecular properties and chemical reactions. We should also note that optical cavities are also routinely used in other applications which have not been covered here such molecular spectroscopy, conventional organic lasers and photon Bose–Einstein condensation with liquid dye solutions. However, in these applications, QED effects play a lesser role.

See also: Organic Lasers

Further Reading

Agranovich, V.M., Litinskaia, M., Lidzey, D.G., 2003. Cavity polaritons in microcavities containing disordered organic semiconductors. Physics Reviews B 67, 085311.
Bharadwaj, P., Deutsch, B., Novotny, L., 2009. Optical antennas. Advances in Optics and Photonics 1, 438–483.
Brutting, W., Frischeisen, J., Schmidt, T.D., Scholz, B.J., Mayr, C., 2012. Device efficiency of organic light-emitting diodes: Progress by improved light outcoupling. Physica Status Solidi: A 210, 44–65.
Bulović, V., Khalfin, V.B., Gu, G., *et al.*, 1998. Weak microcavity effects in organic light-emitting devices. Physics Review B 58, 3730–3740.
Carusotto, I., Ciuti, C., 2013. Quantum fluids of light. Reviews of Modern Physics 85, 299–366.
Chance, R.R., Prock, A., Silbey, R., 1978. Molecular fluorescence and energy transfer near interfaces. In: Prigogine, I., Rice, S.A. (Eds.), Advances in Chemical Physics, vol. 37. Hoboken, NJ: John Wiley & Sons, Inc., pp. 1–65.
Daskalakis, K.S., Maier, S.A., Kéna-Cohen, S., 2016. Polariton condensation in organic semiconductors. In: Bozhevolnyi, S.I., Martin-Moreno, L., Garcia-Vidal, F. (Eds.), Quantum Plasmonics. Berlin Heidelberg: Springer, pp. 151–163. Volume 185 of the Springer Series in Solid-State Sciences.
Enderlein, J., 2002. Theoretical study of single molecule fluorescence in a metallic nanocavity. Applied Physics Letters 80, 315–317.
Esteban, R., Teperik, T.V., Greffet, J.J., 2010. Optical patch antennas for single photon emission using surface plasmon resonances. Physical Review Letters 104, 026802.
Garcia-Vidal, G.J.F., Feist, J., 2015. Cavity-induced modifications of molecular structure in the strong coupling regime. Physics Reviews X 5, 041022.
Hobson, P.A., Wedge, S., Wasey, J.A.E., Sage, I., Barnes, W.L., 2002. Surface plasmon mediated emission from organic light-emitting diodes. Advance Materials 14, 1393–1396.
Hutchison, J.A., Schwartz, T., Genet, C., Devaux, E., Ebbesen, T.W., 2012. Modifying chemical landscapes by coupling to vacuum fields. Angewandte Chemie International Edition 51, 1592–1596.
Kéna-Cohen, S., Forrest, S.R., 2012. Exciton-polaritons in organic semiconductor microcavities. In: Sanvitto, D., Timofeev, V. (Eds.), Exciton Polaritons in Microcavities. Berlin Heidelberg: Springer, pp. 349–375. Volume 172 of the Springer Series in Solid-State Sciences.
Novotny, L., Hecht, B., 2012. Principles of Nano Optics. Cambridge: Cambridge University Press.
Saito, S., Tsutsui, T., Era, M., *et al.*, 1993. Progress in organic multilayer electroluminescent devices. Proceedings of SPIE 1910, 212–221.

Cavity QED

H Walther, University of Munich and Max-Planck-Institute for Quantum Optics, Garching, Germany

Modification of the Spontaneous Transition Rate in Confined Space

In order to understand the modification of the spontaneous emission rate in an external cavity-like structure, it must be remembered that in quantum electrodynamics this rate is proportional to the density of modes of the electromagnetic field, i.e., the vacuum field fluctuations above the atomic transition frequency ω_0. As a consequence the spontaneous emission rate is increased if the atom is surrounded by a cavity tuned to the transition frequency. Conversely, the decay rate decreases when the cavity is mistuned. This fact was already recognized in the pioneering days of nuclear magnetic resonance by Purcell.

The spontaneous decay rate of the atom in the cavity, γ_c, will then be enhanced in relation to that in free space, γ_f, by a factor given by the ratio of the corresponding mode densities

$$\frac{\gamma_c}{\gamma_f} = \frac{\rho_c(\omega_0)}{\rho_f(\omega_0)} = \frac{2\pi Q}{V_c \omega_0^3} = \frac{Q\lambda_0^3}{4\pi^2 V_c}$$

where V_c is the volume of the cavity and Q is the quality factor of the cavity which expresses the 'sharpness' of the mode. For low-order cavities $V_c \approx \lambda_0^3$ this means that the spontaneous emission rate is increased by roughly a factor of Q. As mentioned already, when the cavity is detuned, the decay rate will decrease. In this case the atom cannot emit a photon, since the cavity is not able to accept it, and the energy will stay with the atom, i.e., its decay is inhibited.

To change the decay rate of an atom, in principle no resonator has to be present; any conducting surface near the radiator affects the mode density and, therefore, the spontaneous radiation rate. Parallel conducting planes can alter the emission rate somewhat but can only reduce the rate by a factor of 2 owing to the existence of modes, independent of the plate separation with an electric field perpendicular to the planes.

In order to demonstrate experimentally the modification of the spontaneous decay rate, it is not necessary to go to single-atom densities. The experiments where the spontaneous emission is inhibited can also be performed with large atom numbers. However, in the opposite case, when the increase of the spontaneous rate is observed, a large number of excited atoms may disturb the experiment by induced transitions. The first experimental work on inhibited spontaneous emission was done by Drexhage, Kuhn and Schäfer in 1974. The fluorescence of a thin dye film near a mirror was investigated. A reduction of the fluorescence decay by up to 25% was observed. Later experiments with microcavities filled with dye solutions were performed by de Martini *et al.* These experiments demonstrated that thresholdless lasing is possible in such systems.

Inhibited spontaneous emission was observed by Gabrielse and Dehmelt. In these neat experiments with a single electron stored in a Penning trap they observed that cyclotron orbits show lifetimes which are up to 10 times longer than that calculated for free space. The electrodes of the trap form a cavity which decouples the cyclotron motion from the vacuum field leading to the longer lifetime.

A new stage in experiments on cavity QED was reached when it was recognized that highly excited alkaline atoms are very suitable for these experiments. The main quantum numbers are in the range $n = 20$-60. Those so-called Rydberg atoms couple very strongly to the radiation field as will be discussed later. The transitions to neighboring states are in the microwave region. The cavities can therefore be built as low-order cavities with dimensions on the order of the wavelength of the transitions being in the mm or cm regions. The spontaneous rate of these transitions in free space is small owing to the low value of their transition frequency, therefore an enhancement is possible. Experiments with Rydberg atoms on the inhibition of spontaneous emission have been performed by Kleppner and coworkers and by Haroche and coworkers. In the latter experiment a 3.4 μm transition of the Cs atom was suppressed.

The first observation of enhanced atomic spontaneous emission in a resonant cavity was published by Haroche *et al.* This experiment was performed with Rydberg atoms of Na excited in the 23s state in a niobium superconducting cavity resonant at 340 GHz. Cavity tuning-dependent shortening of the lifetime was observed. The cooling of the cavity had the advantage of totally suppressing the black-body field. The latter effect is completely absent if optical transitions are observed, however, in this case it is more difficult to obtain low-order cavities. The first experiments on optical transitions were performed by Feld and collaborators. They succeeded in demonstrating the enhancement of spontaneous transitions even in higher-order optical cavities.

In modern semiconductor devices both electronic and optical properties can be tailored with a high degree of precision. Therefore, electron–hole systems producing recombination radiation analogous to radiating atoms can be localized in cavity-like structures, e.g., in quantum wells. Thus optical microcavities of half or full wavelength size are obtained. Both suppression and enhancement of spontaneous emission in semiconductor microcavities were demonstrated in experiments by Yamamoto and collaborators. Yablonovitch *et al.* proposed the use of photonic band gap structures in semiconductors. In those systems spontaneous emission is forbidden in certain frequency regions owing to the special material properties.

Energy Shift in Confined Space

In the previous section we focused on changes of radiation rates of atoms near conducting walls or in cavity-like structures. Next we will discuss the more subtle phenomenon of energy shifts. While radiation rates are modified by the influence of the vacuum field on the real emission of photons the energy shifts are caused by the modification of a virtual photon interaction.

Resonant and nonresonant phenomena have to be distinguished. The resonant self-energy shift of a decaying atomic dipole in the vicinity of a conducting wall can be determined from the average polarization energy produced by the image dipole field. For distances comparable to the wavelength the near-field condition is satisfied, resulting in the z^{-3} dependence of the static dipole–dipole interaction characteristic for the van der Waals energy; under far-field conditions the distance dependence is given by z^{-1}. The polarization of the atom by the nonresonant parts of the broadband electromagnetic field causes energy shifts, of which the Lamb shift is the most prominent. In the sense of the nonrelativistic treatment by Bethe the major contribution of that shift can be described as being a result of the emission and reabsorption of virtual photons. It is plausible that just as the real emission of a photon is modified in confined space, so also is the virtual process. The latter 'real' radiation energy shift is thus a consequence of vacuum fluctuations only. It is identical with the energy shift predicted by Casimir and Polder and is analogous to the better-known result of Casimir on the force between two plane neutral conducting plates.

The question of modification of atomic energies in confined space has recently found considerable interest and many calculations of the phenomenon have been performed. Direct application to the energy shift of Rydberg atoms, which are of special interest for experimental studies, was performed by Barton in 1987. He calculated the direct electrostatic interaction with a conducting wall and the radiation induced (retarded) effects. The result is that in the case of two parallel plates the electrostatic effect is dominant when the distance L between the conducting plates is small, $L < n^3 a_0/\alpha$ (n is the principal quantum number of the Rydberg atom, a_0 the Bohr radius, and α the fine-structure constant), and the radiative effect plays the major role when large distances are used, $L > n^3 a_0 \alpha$.

Experiments to measure the van der Waals force of conducting planes have been performed by Hinds and coworkers. They could clearly demonstrate the z^{-3} dependence of the van der Waals force. The retarded effects were not detectable at large distances but at small distances from the wall a deviation from the z^{-3} dependence was found which was attributed to the retarded QED potential.

The energy shift of rubidium Rydberg atoms in confined space has been measured by Marrocco *et al.* in an experiment with extreme high-resolution using the Ramsey experiment two-field method. The atoms are excited in s-Rydberg states ($n \approx 30$) by two-photon transitions using the light of an ultrastable dye laser, the linewidth of which was less than 10 Hz. The laser intensity was enhanced in a folded cavity locked to the laser frequency. Both interaction zones necessary for the Ramsey method were enclosed in this cavity. Between the two interaction regions the atoms pass through a pair of conducting plates, the distance between which can be changed. The shift of the Ramsey interference was measured as a function of the plate distance. Using the Ramsey method has the advantage that the shift can be determined without a direct probing of the atoms in the space between the plates. A level shift on the order of about 150 Hz could be measured, in very good agreement with theory.

Maser Operation

If the rate of atoms crossing a cavity exceeds the cavity damping rate ω/Q the photon released by each atom is stored long enough to interact with the next atom. Here ω stands for the cavity frequency and Q for the quality factor of the cavity. The atom–field coupling becomes stronger and stronger as the field builds up and evolves to a steady state. Using Rydberg atoms with a large field–atom coupling constant leads to a new kind of maser which operates with exceedingly small numbers of atoms and photons. The photons corresponding to transitions between neighboring Rydberg atoms are in the microwave region at about 20–100 GHz. Atomic fluxes as small as a few atoms per second have generated maser action, as could be demonstrated by Walther *et al.* of the University of Munich in 1985 using a superconducting cavity. For such a low flux there is never more than a single atom in the resonator – in fact, most of the time the cavity is empty of atoms. It should also be mentioned that in the case of such a setup a single resonant photon is sufficient to saturate the maser transition.

The scheme for this one-atom maser or micromaser is shown in **Fig. 1**. The setup represents the simplest system in radiation–atom interaction: a single atom is interacting with a single mode of the radiation field. This device was used to verify the complex dynamics of a single atom in a quantized field predicted by the Jaynes–Cummings model. All of the features are explicitly a consequence of the quantum nature of the electromagnetic field: the statistical and discrete nature of the photon field leads to new characteristic dynamics such as collapse and revivals in the exchange of a photon between the atom and the cavity mode. The frequency of this exchange is usually called the Rabi flopping frequency. The field in the cavity is measured through the number of atoms in the lower state of the maser transition. Due to the strong coupling between the atom and maser field both are entangled. The strong coupling can also be used to entangle subsequent atoms.

The steady-state field of the micromaser shows sub-Poisson statistics. This is in contrast to regular masers and lasers where coherent fields (Poisson statistics) are observed. The reason for nonclassical radiation being produced is that a fixed interaction time of the atoms is chosen, leading to careful control of the atom–field interaction dynamics.

Under steady-state conditions, the photon statistics $P(n)$ of the field of the micromaser are essentially determined by the pump parameter $\Theta = N_{ex}^{1/2} \Omega t_{int}$, denoting the angular rotation of the pseudospin vector of the interacting atom. Here, N_{ex} is the average number of atoms that enter the cavity during the decay time of the cavity τ_{cav} Ω the vacuum Rabi flopping frequency, and t_{int} is the

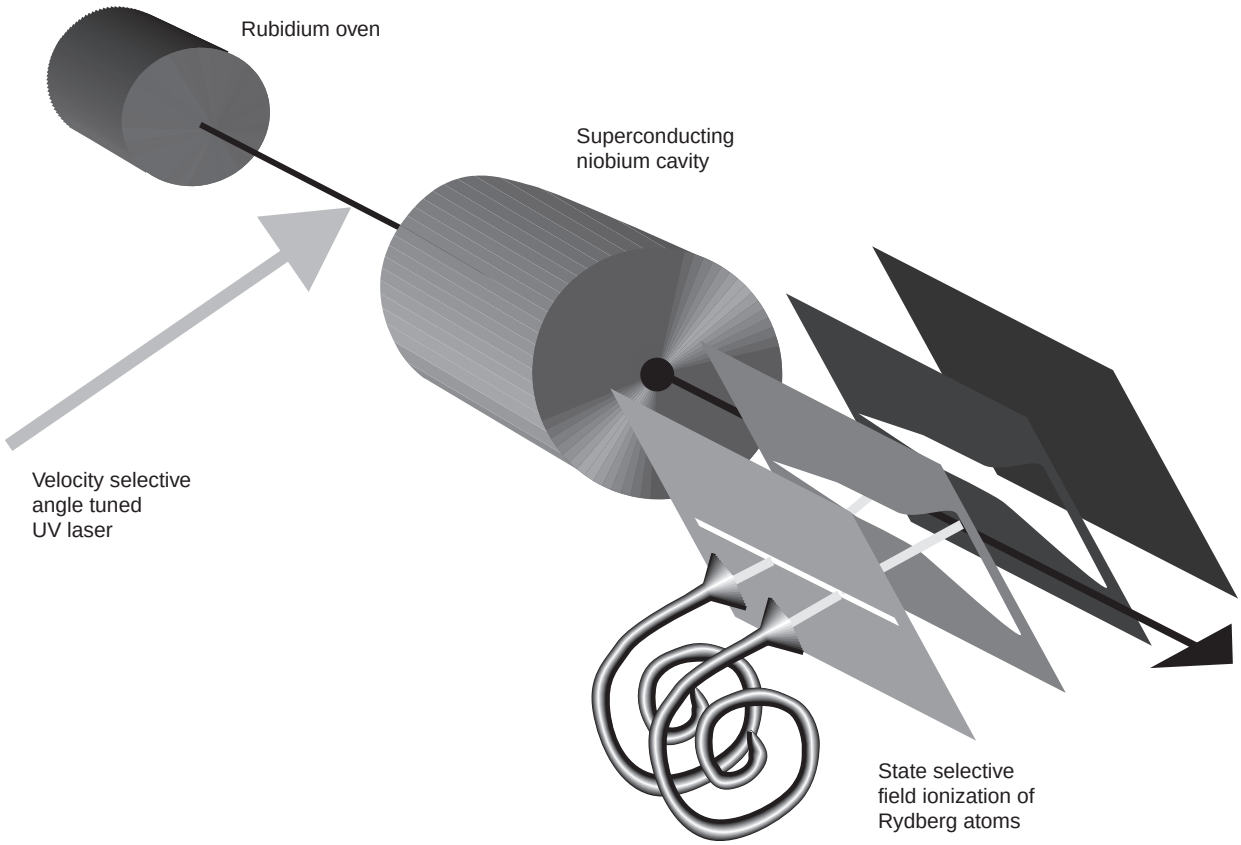


Fig. 1 Micromaser setup of rubidium Rydberg atoms in the 63p state. The velocity of the atoms is controlled by exciting a velocity subgroup of atoms in the atomic beam. The atoms in the upper and lower maser levels are selectively detected by field ionization. The number of photons deposited in the cavity is determined through the number of atoms in the lower state (61d).

atom–cavity interaction time. The normalized photon number of the maser $\langle v\rangle = \langle n\rangle / N_{ex}$ shows the following generic behavior (see Fig. 2). It suddenly increases at the maser threshold value $\Theta = 1$ and reaches a maximum for $\Theta = 2$ (denoted by 1 in Fig. 2). The behavior at threshold corresponds to the characteristics of a continuous phase transition. As Θ increases further, the normalized averaged photon number $\langle v\rangle / N_{ex}$ decreases to a minimum slightly below $\Theta = 2\pi$ and then abruptly increases to a second maximum (3a in Fig. 2). This general type of behavior recurs roughly at integer multiples of 2π but becomes less pronounced with increasing Θ. The reason for the periodic maxima of the average photon number is that for integer multiples of $\Theta = 2\pi$ the pump atoms perform an integer number of full Rabi flopping cycles, and start to flip over at a slightly larger value of Θ, thus leading to enhanced photon emission. The periodic maxima for $\Theta = 2\pi, 4\pi$, and so on may be interpreted as first-order phase transitions.

The photon statistics of the maser radiation is usually characterized by the Q-parameter introduced by Mandel:

$$Q_{\text{field}} = \left\{ \left(\langle n^2 \rangle - \langle n \rangle^2 \right) / \langle n \rangle \right\} - 1$$

It can be seen that $Q_{\text{field}} = 0$ corresponds to a Poissonian photon distribution. Q_{field} for the micromaser is plotted as a dotted line in Fig. 2. The value drops below zero in the region 2a, 2b, etc. This shows the highly sub-Poissonian character of the one-atom-maser field being present over a large range of parameters.

The reason for the sub-Poissonian atomic statistics is the following. A changing flux of atoms changes the Rabi frequency via the stored photon number in the cavity. Adjusting the interaction time allows the phase of the Rabi nutation cycle to be chosen such that the probability of the atoms leaving the cavity in the upper maser level increases when the flux, and hence the photon number in the cavity, is raised. This leads to sub-Poissonian atomic statistics since the number of atoms in the lower state decreases with increasing flux and photon number in the cavity. This feedback mechanism can be neatly demonstrated when the anticorrelation of atoms leaving the cavity in the lower state is investigated. Measurements of this anticorrelation phenomenon could be made.

The fact that anticorrelation is observed shows that the atoms in the lower state are more equally spaced than expected for a Poissonian distribution of the atoms in the beam. This means, for example, that when two atoms enter the cavity close to each other, the second one performs a transition to the lower state with reduced probability.

The interaction with the cavity field thus leads to an atomic beam with atoms in the lower maser level showing number fluctuations which are up to 40% below those of a Poissonian distribution found in usual atomic beams. This is interesting

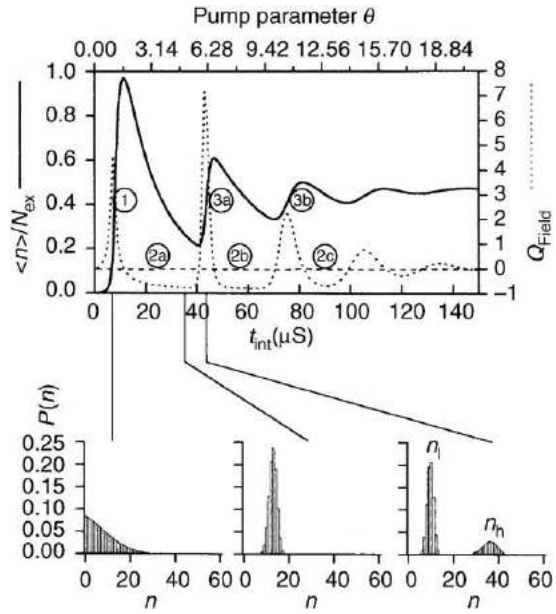


Fig. 2 One-atom maser or micromaser characteristics. The upper part of the figure shows the average photon number versus interaction time (solid curve) and the photon number fluctuations represented by the *Q*-factor (dotted line). Both curves are determined by the photon exchange dynamics between atom and field. The pump parameter gives the angular rotation of the pseudospin vector of the interacting atom (e.g., 2π corresponds to a full rotation, i.e., the atom is again in the upper level). In the lower part of the figure the steady-state photon number distribution $P(n)$ is shown for three values of t_{int}. The distribution on the left side corresponds to the maser threshold, that on the right gives an example of the double-peaked distribution associated with the quantum jump behavior. In this situation, the atom is back in the excited state and can again emit, leading to a higher steady-state photon number n_h. With increasing t_{int} this part will grow and n_l will decrease and disappear. A new jump occurs in the region 3b.

because atoms in the lower level have emitted a photon to compensate for cavity losses inevitably present. Although this process is induced by dissipation giving rise to fluctuations, the atoms still obey sub-Poissonian statistics.

Generation of Number States (Fock States) of the Radiation Field

When the micromaser is operated at low temperatures, a very interesting feature is observed in the inversion of the population of the two maser levels called trapping states. They are a steady-state feature of the maser field; they occur in the micromaser as a direct consequence of field quantization in a cavity. At low cavity temperatures the number of black-body photons, n_{th}, in the cavity mode is reduced and trapping states begin to appear; at higher temperatures they are washed out by the thermal photons. The trapping states show up when the atom–field coupling Ω and the interaction time t_{int} are chosen such that in a cavity field with n_q photons the atoms undergo an integer number k of Rabi cycles. This is summarized by the condition

$$\Omega t_{int}\sqrt{n_q+1} = k\pi \tag{1}$$

for the trapping state, denoted by ($n_q k$ When **Eq. (1)** is satisfied, the cavity photon number is left unchanged after interaction of an atom and hence the photon number is 'trapped'. This will occur regardless of the atomic pump rate N_{ex}. The trapping state is therefore characterized by the upper-bound photon number n_q and the number of integer multiples of full Rabi cycles k. In this situation the field is stabilized. Whenever a photon disappears, e.g., due to dissipation, the next incoming atom experiences a changed Rabi nutation frequency and emits a photon with high probability. At the trapping condition a quantum nondemolition situation is present. Through the dynamics of the Rabi nutation the field is measured, without any net change of the field.

The build-up of the cavity field can be seen in **Fig. 3**, where the emerging atom inversion is plotted against the interaction time and pump rate. At low atomic pump rates (low N_{ex}) the maser field cannot build up and the maser exhibits Rabi oscillations due to the interaction with the vacuum field. At the positions of the trapping states, the field increases until it reaches the trapping-state condition. This is manifested as a reduced emission probability and hence as a dip in the atomic inversion. Once in a trapping state, the maser will remain there regardless of the pump rate.

Under ideal conditions the micromaser field in a trapping state is a Fock state, but when the micromaser is operated in a continuous wave mode, the field state is very fragile and highly sensitive to external influences and experimental parameters. However, in contrast to continuous-wave operation, in pulsed operation trapping states are more stable and more practical and can be used over a much broader parameter range than in continuous-wave operation.

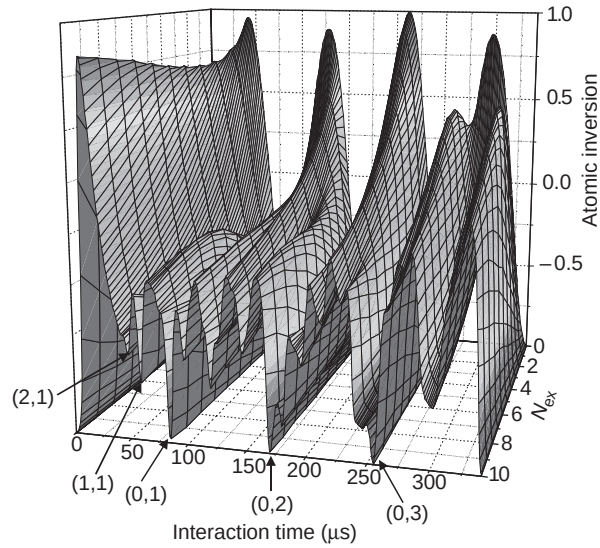


Fig. 3 Behavior of the micromaser at low temperature showing the trapping states, manifested as valleys along the N_{ex} axis. For the designation of the trapping states see text.

To demonstrate the principle, Fig. 4 shows a simulation of a sequence of 20 pulses of the pumping atoms in which an average of seven excited atoms per pulse are present. Under the trapping condition only a single emission event occurs, producing a single lower state atom which leaves a single photon in the cavity. Since the atom–cavity system is in the trapping condition, the emission probability is reduced to zero and the photon number is stabilized. Consequently, excited state atoms following the emitting atom stay in the upper maser level. The variation of the time when an emission event occurs during the atom pulses in Fig. 4 is due to the Poissonian spacing of upper-state atoms entering the cavity and the stochasticity of the quantum process.

The lower part of Fig. 4 shows the photon number distribution resulting from this process. It was demonstrated experimentally by Brattke *et al*. that a single photon number state could be generated with a success rate of 85%. By improving the experimental parameters one can expect to prepare single-photon Fock states in 98% of the pulses by this method.

Other Cavity Experiments

Besides the micromaser, there are also other experiments using Rydberg atoms in cavities and the strong interaction of these atoms with cavities which lead to interesting applications. One example is an experiment by S. Haroche, J. M. Raimond *et al.*, who succeeded in realizing a Schrödinger cat state in a cavity. Studies on the decoherence of this state could be conducted. The experiments are quite important in exploring the boundaries between the quantum and classical worlds. Another example is the quantum nondemolition detection of a single photon in a cavity by S. Haroche employing the dispersion level shift of probing atoms.

There is an interesting equivalence between an atom interacting with a single-mode field and a quantum particle in a harmonic potential, as was first pointed out by D. F. Walls and H. Risken; this connection results from the fact that the radiation field is quantized on the basis of the harmonic oscillator. The Jaynes–Cummings dynamics can therefore also be observed with trapped ions, as recently demonstrated in a series of beautiful experiments by D. Wineland *et al*. They also produced a Schrödinger cat state by preparing a single trapped ion in a superposition of two spatially separated wavepackets which are formed by coupling different vibrational quantum states in the excitation process.

Besides the experiments in the microwave region, a single-atom laser emitting in the visible range has also been realized by M. Feld *et al.* Furthermore, cavity quantum electrodynamic effects have been studied in the optical spectral region by J. L. Kimble *et al.*

Today's technology in microfabrication of semiconductor diode structures allows the realization of low-order cavity structures for diode lasers. In these systems the spontaneous emission is controlled in the same way as in the micromaser. Since spontaneous decay is a source of strong losses, control of this phenomenon leads to highly efficient laser systems. Quantum control of spontaneous decay thus has important consequences for technical applications. This topic will be briefly described in the next section.

Microlasers

The simplest approach to fabricating an optical microcavity is to shrink the spacing between the mirrors of a Fabry–Perot resonator to λ/n (where n stands for the refractive index) while reducing the lateral dimensions to a range of the same order of magnitude. This structure provides a single dominant longitudinal field mode that radiates into a narrow range of angles around the cavity axis.

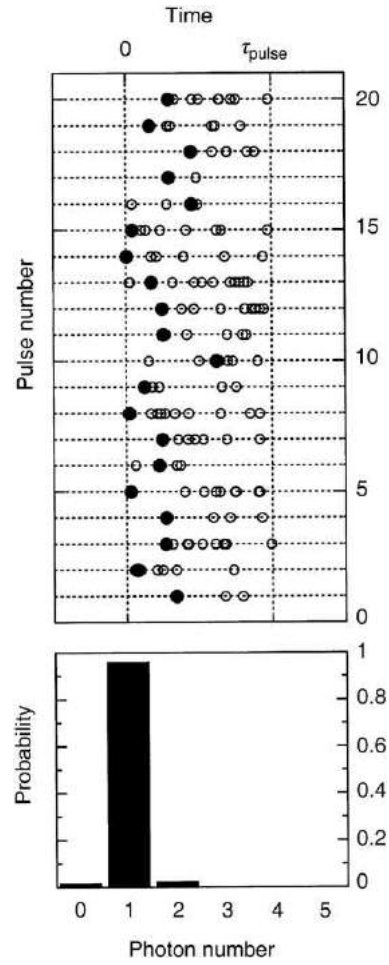


Fig. 4 A simulation of a subset of 20 subsequent atom bunches generated by pulsed laser excitation with the associated probability distribution for photons in the cavity or lower-state atom production (filled circles represent lower-state atoms and open circles represent excited-state atoms). The start and end of each pulse is indicated by the vertical dotted lines marked 0 and τ_{pulse}, respectively. For details, see Battke S, Varcoe BTH and Walther H (2001) Generation of photon number states on demand via cavity, quantum electrodynamics. *Physical Review Letters* 86: 3534–3537.

The first optical microcavity experiments used dye molecules between high-reflectivity dielectric mirrors in the Fabry–Perot configuration. Because spontaneous emission is a major source of energy loss, speed limitations, and noise in lasers, the capability to control spontaneous emission is expected to improve laser performance. If the fraction of spontaneous emission coupled into the lasing mode is made close to one, the 'thresholdless' laser is obtained, in which the light output increases almost linearly with the pump power instead of exhibiting a sharp turn-on at the pump threshold.

Semiconductor microcavities provide high-Q Fabry–Perot cavities for both basic studies and potential applications. Molecular beam epitaxy or organometallic chemical vapor deposition techniques are used to deposit high-reflectivity mirrors consisting of alternating quarter-wavelength layers of lattice-matched semiconductors. For example, 15–20 pairs of quarter-wave layers of $Al_{0.2}Ga_{0.8}As$ and AlAs result in a reflectivity greater than 99% and Q values greater than 500. The optically active layer in such a microcavity is typically a GaAs quantum well located at the midplane of the cavity, where the field strength is maximum.

Since the microcavities have an extremely low threshold, their efficiency will also be high. Low-cost, high-density light source arrays and photonic circuits are made possible by the small size and low power consumption of such resonators. It will be possible to produce entire wafers containing millions of microlasers with a multipole arrangement instead of having to cleave each individual semiconductor laser as at present. This improvement will lead to higher yields and lower cost per element. Another advantage of surface-emitting microcavity sources is the efficient optical coupling of their stable symmetric mode patterns into optical fibers or waveguides. These lasers are thus certain to spark a revolution in optical communication in the future.

Conclusion

This article reviewed experiments in cavity quantum electrodynamics. The experiments shed new light on our understanding of the radiation interaction of atoms. The achievable strong coupling between atoms and radiation leads to the possibility to generate

entanglement between the generated radiation field and the atoms, presenting a basis for interesting applications in connection with, e.g., quantum computing and quantum information processing. Furthermore Fock states of the radiation field can be generated. The control of spontaneous decay finally leads to interesting new laser systems offering high efficiency.

Further Reading

Berman, P.R., 1994. Cavity quantum electrodynamics. In: Bederson, B., Bates, D. (Eds.), Advances in Atomic, Molecular, and Optical Physics. Boston, MA: Academic Press. Supplement 2.

Englert, B.-G., Löffler, M., Benson, O., *et al.*, 1998. Entangled atoms in micromaser physics. Fortschritte der Physik 46, 897–926.

Haroche, S., 1992. Cavity quantum electrodynamics. In: Dalibard, J., Raimond, J.M., Zinn-Justin, J. (Eds.), Fundamental Systems in Quantum Optics. Amsterdam: North-Holland, pp. 767–935.

Haroche, S., Kleppner, D., 1989. Cavity quantum electrodynamics. Physics Today 42, 24–30.

Kimble, H.J., Carnal, O., Georgiades, N., *et al.*, 1995. Quantum optics with strong coupling. In: Wineland, D.J., Wieman, C.E., Smith, S.J. (Eds.), Atomic Physics, vol. 14. New York: American Institute of Phycis, pp. 314–335.

Marrocco, M., Weidinger, M., Sang, R.T., Walther, H., 1998. Quantum electrodynamic shifts of Rydberg energy levels between parallel metal plates. Physical Review Letters 81, 5784–5787.

Meschede, D., 1992. Radiating atoms in confined space: from spontaneous emission to micromasers. Physics Reports 211, 201–250.

Meystre, P., 1992. Cavity quantum optics and the quantum measurement process. In: Wolf, E. (Ed.), Progress in Optics, vol. 30. Amsterdam: North-Holland.

Walther, H., 1992. Experiments on cavity quantum electrodynamics. Physics Reports 219, 263–281.

Cavity QED in Semiconductors

M Kira, W Hoyer, and SW Koch, Philipps-University, Marburg, Germany
G Khitrova and HM Gibbs, University of Arizona, Tucson, AZ, USA

Introduction

According to classical electromagnetism, propagating light fields obey Maxwell's equations under which electric and magnetic fields exchange their strength in an oscillatory manner. If a propagating field is followed at a single reference point, it is convenient to represent the electric field as a complex number $|E(t)|\exp(-\mathrm{i},\omega,t)$ where $|E(t)$ is the magnitude of the field and the phase of the field $\exp(-\mathrm{i}\omega t)$ oscillates in time t with the optical frequency ω. **Fig. 1** illustrates how the classical field is determined by one single point in the complex plane; consequently, the phase and amplitude of the field are known exactly. According to quantum mechanics, this simple picture has to be altered because the Heisenberg uncertainty relation states that complementary quantities like phase and amplitude cannot be determined precisely at the same time. As a result, the real and imaginary parts of the field can be determined only with accuracy $\Delta\mathrm{Re}[E]$ and $\Delta\mathrm{Im}[E]$, respectively, and the Heisenberg uncertainty principle determines the best possible accuracy $\Delta\mathrm{Re}[E]\ \Delta\mathrm{Im}[E]\geq 1$. To incorporate the fundamental inaccuracy, the electric field has to be defined by using complex-valued distributions or equivalently wavefunctions as shown in **Fig. 1**. This quantization procedure leads to the Schrödinger equation for light analogous to that for particles. The field called quantum optics investigates the quantum electrodynamics (QED) features of light.

In general, light does not propagate freely in space, but it interacts strongly with the surrounding matter. For example, light can be absorbed, and its energy can be converted into excitation of the matter. As a result, the light may be slowed down, reflected, scattered, or diffracted. In order to explain the implications of the light–matter interaction in detail, one obviously has to apply a quantum mechanical description also for the matter. This approach determines the so-called eigenstates of the matter, so that the matter may occupy only certain discrete states with discrete energies. The Rydberg series of a hydrogen atom is a typical example. Similar discrete energy levels are also found for the quantized light; these levels are often referred to as photons which heuristically describe the particle properties of the light. The lowest-order light–matter interaction consists of processes where one photon is absorbed (emitted) while the matter is simultaneously excited (de-excited) from one eigenstate to another. This generic type of interaction leads to a microscopic description of the optical effects mentioned above. The magnitude of these effects depends on the strength of the interaction. By implementing different cavity configurations, the reflected light can be forced to travel across the same matter many times. As a result, the light–matter interaction is enhanced by a factor proportional to the multiple passes of the light. Thus, a cavity can be efficiently used to amplify optical phenomena. **Fig. 2** shows cavity configurations commonly used for semiconductors. Semiconductor cavity QED investigates quantum optical effects in semiconductor systems by enhancing them with cavity mirrors. In general, quantum optical features produce classically unexpected effects which typically stem from: (a) the discrete nature of eigenstates of light and matter; (b) the superposition principle of quantum mechanics; and (c) the Heisenberg uncertainty principle. In this article, we review both theoretical and experimental advances made so far to predict, observe, and control quantum optical effects in semiconductors with respect to effects (a)–(c).

During the last few decades, atomic quantum optics has already developed toward QED investigations, and semiconductor optics is developing rapidly into the same regime. Advanced atomic QED theories have successfully explained and guided intriguing experimental developments like laser cooling, atom condensation, and photon teleportation. However, the atomic

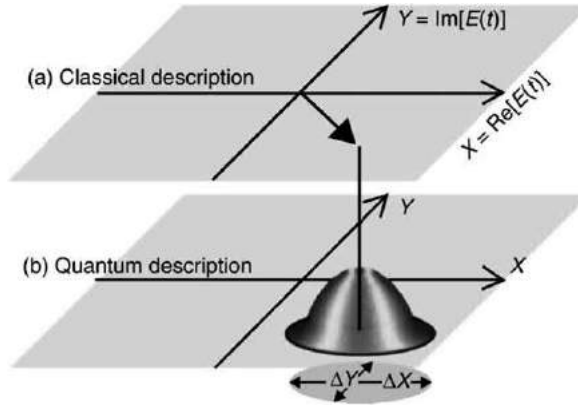


Fig. 1 (a) The classically described light field $E(t)$ is a single point in the complex XY-plane; the corresponding (X,Y) vector has a definite phase and angle. (b) Quantum mechanically described light is a distribution with fluctuations ΔX and ΔY described by the shaded circle.

 doi:10.1016/B978-0-12-809283-5.00915-0

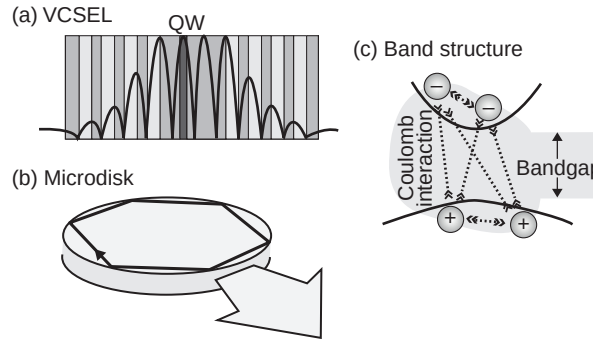


Fig. 2 (a) A typical vertical-cavity surface-emitting laser (VCSEL) structure; a stack of quarter-wavelength layers with different refractive index provides a mirror such that the light intensity (solid line) is strongly confined inside the cavity. A quantum well (QW) is positioned at the field maximum. (b) In a microdisk laser, light encounters ideal total reflection when the edge of the disk is reached. The so-called whispering gallery mode emits (leaks) light from the edge. (c) A schematic illustration of semiconductor band structure, bandgap, and Coulomb interaction for an electro–hole system.

approach can mostly be applied to describe dilute and only weakly interacting atomic gases since relatively simple models of few-level systems are used to describe the material. The elementary electronic excitations in semiconductors consist of electrons and holes (missing electrons) lifted into the conduction and valence bands, respectively. The corresponding eigenstates form continuous energy bands with the band-gap energy separation indicated in Fig. 2. Since electrons and holes have opposite charges, they experience Coulomb attraction whereas bare electrons or holes repel each other. Under favorable conditions, the attractive Coulomb interaction prevails and atom-like bound electron–hole pairs (excitons) may be formed. However, since the Coulomb force has an infinite range and an electron–hole system is typically dense, excitons cannot be treated as weakly interacting quasiparticles. As a result, the atomic QED approaches cannot be used to describe quantum optics of semiconductors in general. Thus, this article concentrates mostly on semiconductor QED theory which fully includes the interaction of charge carriers, i.e., electrons and holes, as well as quantum features of light.

The development work on quantum devices and processes is vital for advancements in several key technologies such as computers and telecommunications. The continuous decrease in component dimensions leads to a microscopic structure size (less than 1 μm); at the same time, increase in the device performance is accompanied by ultrafast operation time (faster than 1 ps). Due to these development trends, the properties of components and processes are becoming more and more quantum mechanical. Although the description of the quantum processes is complicated, the microscopic behavior offers entirely new operational functionality such as massively parallel quantum computers. These possibilities are based on the controlled manipulation of the quantum mechanical state of the light and matter. Due to rapid recent developments, optically coupled semiconductor devices have a great potential to become technologically successful and commercially viable quantum components.

Quantized States: Observation via Strong Light–Matter Interaction

In order to enhance the light–matter interaction, we choose a system where the semiconductor is placed inside a cavity in the position where the field intensity is maximum. For an empty cavity, the field intensity has a strong dependence upon frequency with a strong resonance at ω_R when the cavity length is an integer number of light wavelengths in the material. The width of the resonance is directly related to the quality of the cavity as determined by the number of back-and-forth reflections of light inside the cavity. If one chooses a vertical-cavity surface-emitting laser (VCSEL) structure, the most efficient coupling is obtained by placing a thin planar semiconductor structure at the field maximum as shown in Fig. 2. If the semiconductor is planar and thin, the structure is called a quantum well because electrons and holes are confined in one direction in analogy to the standard particle-in-a-box problem of fundamental quantum mechanics. As a result, the energy levels of a quantum well are discrete in the confinement direction. If the structure is narrow enough, the system dynamics is confined to the lowest quantum-well level such that the carriers are quasi-two-dimensional due to the free in-plane motion. The optical response of such a system to a weak classical probe beam has been successfully analyzed with the so-called semiconductor Bloch equations. For the quantum-well system alone, the absorption spectrum may have a sharp resonance; this is often referred to as an excitonic resonance since it is located below the fundamental bandgap energy at a position corresponding to the exciton binding energy. Obviously, the light–matter coupling becomes large when the exciton and cavity resonances coincide. However, since these resonances are coupled, the optical response is altered; we observe a splitting into two absorption peaks instead of the original degenerate resonances. This phenomenon is commonly referred to as normal-mode coupling; in general, it is a quite common feature in quantum mechanics that degenerate states split due to an additional interaction. Fig. 3 shows the first experimental observation of normal-mode coupling in semiconductors.

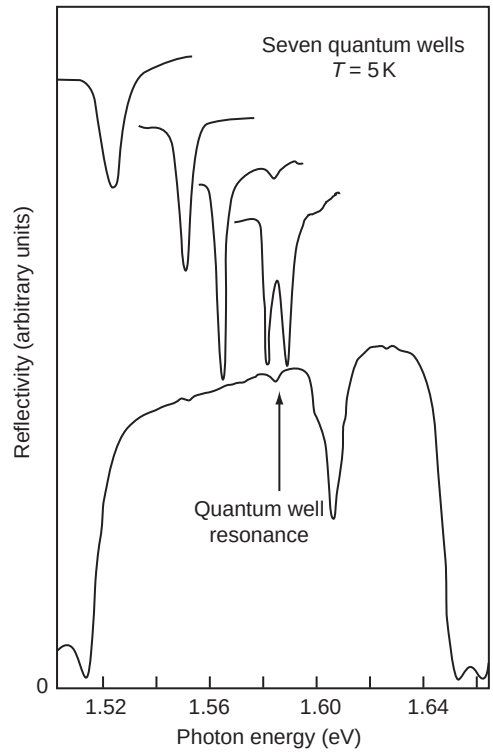


Fig. 3 Probe reflectivity of a seven-quantum-well microcavity structure. The various curves correspond to the energy difference between exciton and cavity resonances. When these resonances become degenerate, the reflectivity shows two resonances corresponding to the normal-mode splitting. From Weisbuch C, Nishioka M, Ishikawa A and Arakawa Y (1992) Observation of the coupled exciton-photon mode splitting in a semiconductor quantum microcavity. *Physical Review Letters* 69: 3314. Copyright (2004) by the American Physical Society.

Since a semiconductor is a strongly interacting many-body system, several effects alter the specific character of an excitonic resonance when electrons and holes are excited. The attractive interaction between electrons and holes becomes weaker for increased carrier density since the surrounding charges screen the bare Coulomb interaction. A particular exciton may also experience scattering from nearby electrons, holes, or other excitons via the Coulomb interaction. Furthermore, electrons and holes are Fermions, which requires that any given carrier state can only be occupied once. This poses fundamental limits on how many carriers can coexist in a specific volume; above a certain limit, additional occupation is prevented, and so-called Pauli blocking is observed. Since an exciton consists fundamentally of Fermionic constituents, eventually Pauli blocking effects become important even for excitons. Due to a combination of these many-body effects, the excitonic resonance is weakened by an increasing carrier density. **Fig. 4** shows a comparison between theory and experiment of the excitonic absorption spectrum for different carrier densities. For elevated densities, the exciton resonance is broadened and its height reduced. When the same investigations are repeated in a microcavity, we observe that the normal-mode peaks decrease in height but their separation is roughly unchanged for moderate densities. In general, the normal-mode spitting is proportional to the oscillator strength of the excitonic absorption, i.e., the area under the resonance. Thus, the observed constant splitting suggests an unchanged oscillator strength for moderate densities. The invariant oscillator strength was unexpected and can only be explained by using a theory which includes the Coulomb interaction of carriers microscopically. When the carrier density is increased even further, the exciton resonance is completely bleached, and the microcavity transmission has only a single peak, at the cavity mode energy; this is commonly referred to as the weak-coupling regime, in contrast to the nonperturbative strong coupling with two transmission peaks. The strong coupling regime provides several intriguing phenomena; for example, parametric amplification of emission has been demonstrated by applying simultaneously multiple light beams to the sample.

The ultimate cavity QED limit of normal-mode coupling follows when the light–matter interaction is so strong that even a single photon leads to splitting (genuine strong coupling). We briefly outline some aspects of the quantum statistical limit known from atomic physics. This situation can be analyzed with the so-called Jaynes–Cummings model where only two states of the atom are included together with a single light mode. In this case, the interaction between a single photon and an atom leads to a normal-mode splitting g_0. If two photons interact with one atom, the splitting increases to $2^{1/2}g_0$; more generally coupling between n photons and an atom leads to a splitting of $n^{1/2}g_0$. In the QED limit, the matter response is very nonlinear because one can detect the quantized nature of light directly from the energy splitting. This limit has already been reached in atomic cavity QED, whereas this regime is hard to approach with semiconductors because they behave more like multi-atom systems. If one has N atoms in a cavity, the first photon excites one atom, but there are N different ways of doing this, so the splitting is $N^{1/2}g_0$. If a second photon

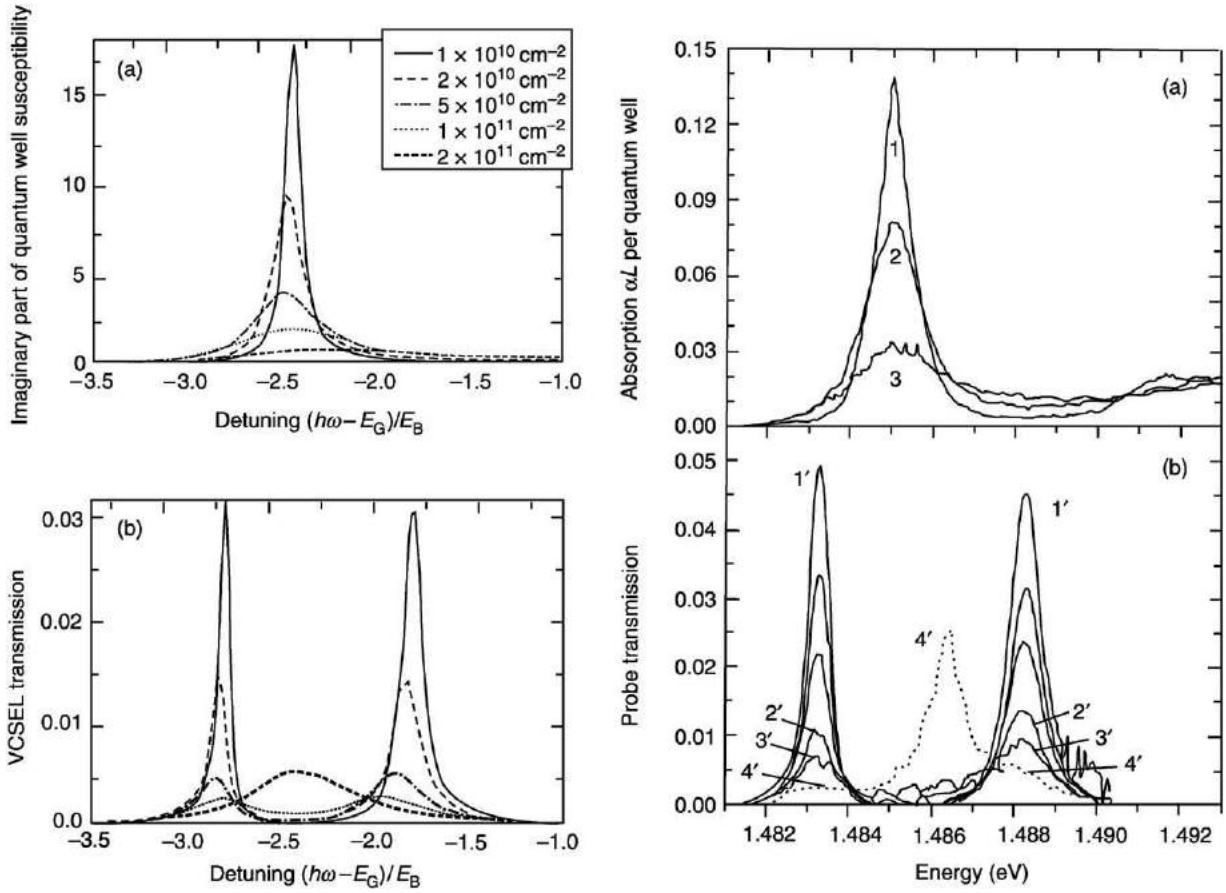


Fig. 4 Left column: (a) microscopically computed bare quantum-well absorption spectrum, i.e., the imaginary part of the susceptibility, as a function of carrier density. (b) Calculated transmission of the quantum-well microcavity for corresponding densities. The right column shows experimentally determined (a) absorption and (b) microcavity transmission for the same conditions as in the calculations. From Jahnke F, Kira M, Koch SW, *et al.* (1996) Excitonic nonlinearities of semiconductor microcavities in the nonperturbative regime. *Physical Review Letters* 77: 5257. Copyright (2004) by the American Physical Society.

arrives, it interacts with the remaining unexcited atoms rather than the excited one, so that the splitting remains $N^{1/2}g_0$ as long as the number of photons is much smaller than N. Thus, in an N-atom system, or equivalently in semiconductors, the quantized light effects are much harder to observe than for a single atom. Currently, the normal-mode splitting in a semiconductor microcavity is explainable by the classical features of light, so that only the quantized nature of the matter is observed. In order to approach the ultimate cavity QED limit, one obviously has to reduce N and increase the light–matter coupling. In the future, this objective might be possible in quantum-dot/nanocavity systems where the semiconductor is confined in all spatial directions.

When an excited semiconductor is not under any influence of external classical light fields, it can still emit light via spontaneous emission resulting from the recombination of electron–hole pairs. The resulting light emission is called photoluminescence. Since the emission process takes place in a strongly interacting many-body system, one has to systematically include the Coulomb interaction and Fermionic features. The emission properties can be consistently described by the so-called semiconductor luminescence equations. When the microcavity photoluminescence spectrum is investigated, one observes a similar normal-mode splitting as for the transmission studies. However, this normal-mode coupling is not in the ultimate cavity QED limit even though the quantum fluctuations of the light field trigger the spontaneous emission. Nevertheless, the luminescence shows an interesting new transition as a function of the excitation level. **Fig. 5** contains a comparison between theory and experiment of normal-mode peak heights and positions as a function of carrier density. For low densities, the high-energy peak is lower in height but it overtakes the low-energy peak for moderate carrier densities that are still below lasing threshold. This threshold-like overtaking was attributed to Boser action, involving exciton formation, final-state stimulation, and Bose condensation. The inset shows an estimate of just how Bosonic the excitons are, i.e., how much they actually behave as independent atoms; the value unity corresponds to the fully Bosonic situation. The nonlinear luminescence transition takes place at a density regime where the underlying Fermionic electron and hole contributions become important (commutator is roughly 0.5). A more detailed analysis with the semiconductor luminescence equations shows that Fermionic carrier nonlinearities in a detuned cavity are responsible for the experimental observations. This example of a misinterpretation based on a Bosonic analysis stresses how important it is to include the Coulomb interaction and Fermion character of carriers when analyzing the properties of semiconductors.

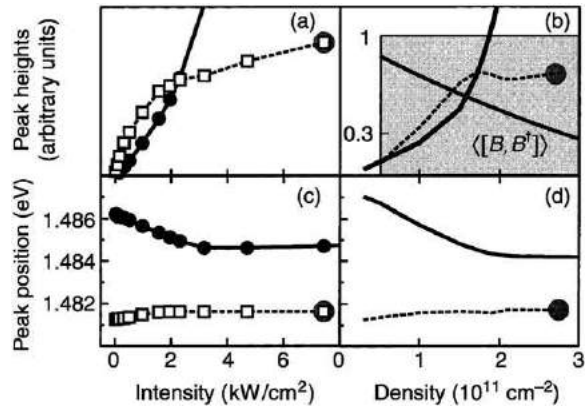


Fig. 5 Measured microcavity luminescence intensities (a) and peak energies (c) versus excitation intensity for the high-energy peak (solid line) and the low-energy peak (dashed line). The results of the microscopic theory are shown in (b) and (d). From Kira M, Jahnke F, Koch SW, *et al.* (1997) Quantum theory of nonlinear semiconductor microcavity luminescence explaining "boser" experiments. *Physical Review Letters* 79: 5170. Copyright (2004) by the American Physical Society.

Superposition Principle: Observation of Quantum Optical Correlations

The above normal-mode-coupling investigations have shown interesting nonlinear effects, but they also revealed that the ultimate QED limit – demonstrating discrete states of light directly – has not yet been reached. Alternatively, a QED investigation may concentrate on other fundamental features of quantum mechanics. One possibility is to study the implications of the superposition principle which states that the joint properties of a light–matter system can always be expressed as a superposition that combines light and matter wavefunctions. The most dramatic consequence can be observed in so-called entangled wavefunctions which cannot be factorized into light and matter parts. In a factorized wavefunction, the light and matter parts are independent whereas entangled wavefunctions conditionally connect light and matter wavefunctions. To elaborate the subtle details of entanglement, we first analyze a simple example. Assume that light can be in two different polarization states, $|\sigma^+\rangle$ or $|\sigma^-\rangle$, and the matter part is either excited $|\text{up}\rangle$ or de-excited $|\text{down}\rangle$. A wavefunction, $[|\text{up}\rangle + |\text{down}\rangle]\,[|\sigma^+\rangle + |\sigma^-\rangle]$, is clearly a superposition of the fundamental states, and at the same time the light and matter parts are completely factorized. However, the wavefunction $[|\text{up}\rangle\,|\sigma^+\rangle + |\text{down}\rangle\,|\sigma^-\rangle]$ is entangled, since the light and matter parts cannot be separated. In the entangled state, measurement on the light state will conditionally determine a definite state of the matter, whereas the factorized state has no such conditionality. The entanglement has far-reaching consequences which cannot be understood with a classical analysis. For example, the principles of teleportation and quantum computing follow directly from the controlled manipulation of different parts of the entangled wavefunction. For atomic systems, wavefunction entanglement has been demonstrated and utilized in several experiments. For semiconductors, the direct manipulation and detection of the wavefunction seems difficult due to the overwhelming complexity of the many-body wavefunctions. Once again, the entanglement and the wavefunction are simplest for low-dimensional structures; direct entanglement effects have been demonstrated recently in a single quantum dot and between a pair of dots. Also for more complicated semiconductor systems, such as a quantum well in a microcavity, the entanglement can be observed as correlations between light and matter. In this case, the existence of QED correlations basically means that light and matter properties depend conditionally on each other. In general, the direct entanglement and QED correlation investigations have a large development potential for semiconductors.

When a semiconductor is excited with an external light pulse, entanglement-related correlations couple the classical and luminescence emission dynamics. The resulting dynamic interplay between the semiconductor Bloch and luminescence equations is mainly mediated by the correlation between photons and electron–hole densities. In the following, we analyze a microcavity configuration where a strong pump pulse generates large QED correlations. The effect of the correlations is then measured by the response of a weak probe beam. **Fig. 6** shows a comparison of theory and experiment of probe reflection in a configuration where the pump and probe do not have any spectral overlap. The measured probe reflection displays long-living oscillations as a function of time delay, i.e., phase difference, between the pump and the probe. Only by including the QED correlation in the theory can the oscillatory probe reflection be explained. When the QED contributions are omitted from the theory, the phase difference does not have any effect on the probe reflection. Thus, this experiment–theory example represents a direct observation of the cavity QED effects in semiconductors. **Fig. 6** also shows more pronounced oscillations when two phase-locked pump pulses provide the excitation; again the full QED theory explains the enhanced oscillation features.

The QED features can also alter the normal-mode coupling characteristic of a weak probe beam. **Fig. 7** shows a comparison of theory and experiment for a situation where the pump spectrum is located between the two normal-mode coupling resonances. Both the theory and experiment show a new third resonance which follows the energetic position of the pump. The third peak is a direct consequence of the QED correlations. Since the third peak is generated from the QED field–carrier correlations generated by

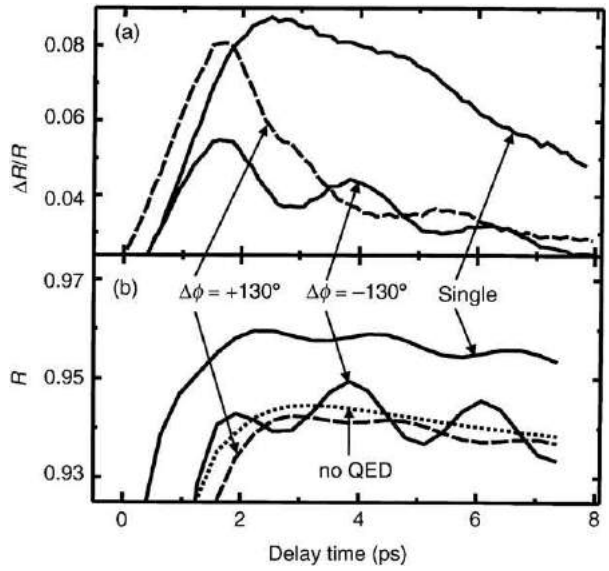


Fig. 6 (a) Measured differential reflectivity and (b) computed reflection probability of the probe pulse as function of probe delay with respect to the excitation pulse. For two-pulse excitations, $\pm 130^{\circ}$ relative phase shifts are used. The dotted line is computed without the QED correlations. From Lee Y-S, Norris TB, Kira M, *et al.* (1999) Quantum correlations and intraband coherences in semiconductor cavity QED. *Physical Review Letters* 83: 5338. Copyright (2004) by the American Physical Society.

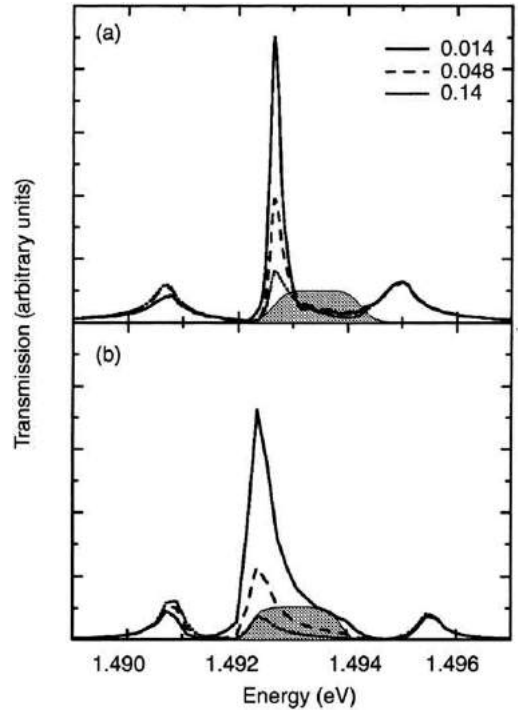


Fig. 7 Dependence of probe transmission on probe intensity (nJ cm^{-2}) at a constant pump intensity of 330 nJ cm^{-2} for the microcavity. (a) Experiment, and (b) theory. The shaded regions indicate the pump spectrum. From Ell C, Brick P, Hübner M, *et al.* (2000) Quantum correlations in the nonperturbative regime of semiconductor microcavities. *Physical Review Letters* 85: 5392. Copyright (2004) by the American Physical Society.

the pump pulse, its magnitude follows the intensity of the pump. As the probe intensity is decreased the relative effect of the QED correlations is increased so that the third peak grows in the probe transmission.

In these QED investigations, the cavity plays an important role since the mirror feedback leads to a significant amplification of the QED effects through the enhanced light–matter coupling. From a practical point of view, normal-mode coupling also provides well-separated spectral features which can be used as classical-emission reference points.

Heisenberg Uncertainty Principle: Squeezing of Light Emission

Quantized light effects can also be investigated by measuring the light fluctuations of the field directly. To maximize the quantum effects, we analyze the field fluctuations in the emission directions where the classical field vanishes. This situation can be realized in planar quantum-well structures which are nearly free of disorder. In such systems, a light pulse propagating perpendicular to the structure leads to transmission and reflection of classical light only along the excitation axis. If the detection is performed at an angle, the so-called secondary emission is purely quantum mechanical. However, the classical and quantum emissions are coupled in the same way as for the QED correlation study. The resulting fluctuations of secondary emission obey the Heisenberg uncertainty principle; in the following, the special quantum features of these fluctuations are investigated.

In the full analysis, we determine the variance ΔX and ΔY of the emission as shown in Fig. 8 (see also Fig. 1). The Heisenberg uncertainty principle requires that the quadrature fluctuations obey $\Delta X\,\Delta Y \geq 1$. For a specific quadrature, the minimum uncertainty limit is usually defined to be $\Delta X = 1$ or $\Delta Y = 1$. If ΔX and ΔY are different, the observed light field is squeezed; and if the variance in one quadrature is less than one, the field is squeezed below the minimum uncertainty limit. In both cases, the field has a strong quantum nature; in particular, squeezing below unity suggests that measurements can be more accurate than the standard quantum limit for that quadrature.

To illustrate the behavior of the quantum fluctuations, we excite the quantum well resonantly with a relatively strong pulse. The emission is detected at an angle of 45° away from the excitation axis. Since the excitation pulse is relatively strong, the carrier density starts to oscillate during the pulse because Fermionic carriers can be excited only once; thus, further excitation actually leads to de-excitation. In other words, when these states become almost fully excited, the pulse can no longer excite the system, and we observe periodic de-excitation and excitation analogous to the Rabi-flopping of a strongly excited two-level system. Fig. 8 shows the exciting pulse and corresponding quadrature fluctuations during the excitation process. As long as the pulse is present, the squeezing is at the 4% level and oscillates with the Rabi-flopping frequency. The quadrature fluctuations clearly show squeezing below the minimum uncertainty limit. Hence, the field has obvious quantum properties.

Similar squeezing and quantum characteristics statistics have been predicted and observed in the photon statistics of resonance fluorescence from two-level atoms subjected to an intense coherent light beam. The quantum properties of the scattered light in both quantum-well and atomic emission are enhanced when the driving field forces the system to oscillate between the excited and de-excited state. Just as described above, when the system becomes de-excited, it can no longer emit an additional photon; this inhibition manifests itself as sub-Poissonian photon statistics and squeezing in the mode quadratures. When the excitation pulse re-excites the system, such restrictions are no longer imposed. Thus, the field properties show quantum features oscillating with the Rabi-flopping frequency. Squeezing effects can also be observed with excitation by an electric current; for example, amplitude squeezing of diode laser emission has been demonstrated by controlling the electron statistics of the current flow.

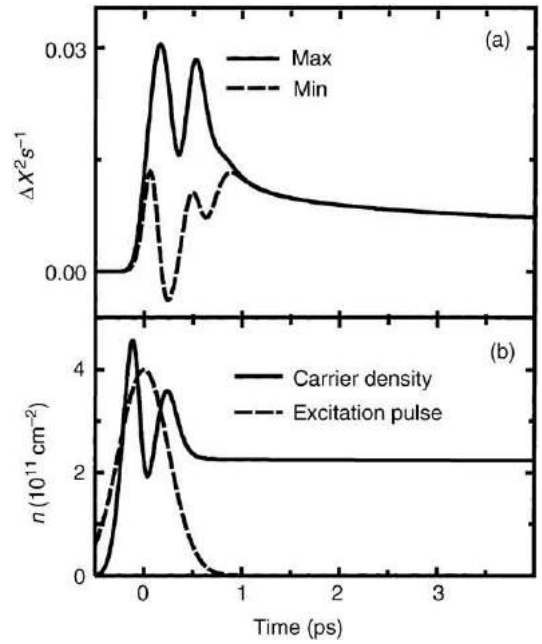


Fig. 8 (a) Time-resolved maximum (solid line) and minimum (dashed line) deviations of the mode quadratures from the vacuum value of unity. (b) Corresponding time-resolved carrier density (solid line) and excitation pulse (dashed line). From Kira M, Jahnke F and Koch SW (1999) Quantum theory of secondary emission in optically excited semiconductor quantum wells. *Physical Review Letters* 82: 3544. Copyright (2004) by the American Physical Society.

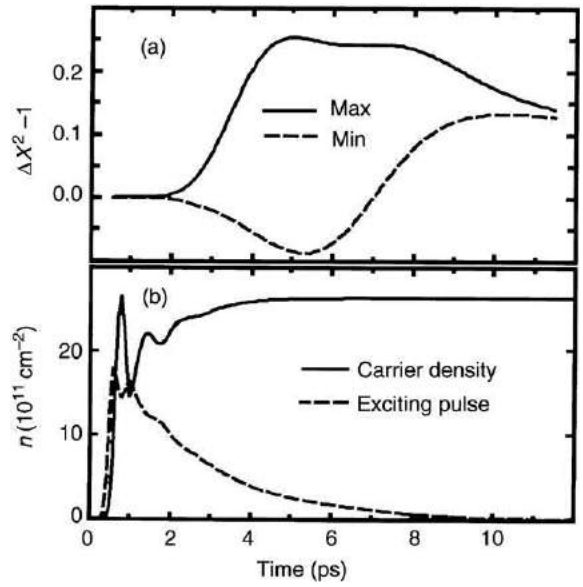


Fig. 9 Squeezed light from a quantum well inside a microcavity: (a) the time evolution of maximum and minimum of the quadrature fluctuations; (b) the corresponding density and light field intensity at the quantum well.

The squeezing investigations demonstrate that some QED effects are expected to be seen without a cavity. **Fig. 9** shows the quadrature fluctuations when the cavity is added to the squeezing analysis. We observe qualitatively similar squeezing, but now the level of squeezing is enhanced up to 30% compared to the minimum uncertainty. This is once again a demonstration how QED effects can be amplified by using an increased light–matter coupling provided by the cavity.

Summary

This article overviews some of the most intriguing features of cavity quantum electrodynamics in semiconductors. Even though the related research has started only recently, several important quantum phenomena have already been predicted and measured: (a) the discrete quantum mechanical states of light and/or matter can be measured with strong coupling; (b) the consequences of the superposition principle have been detected as light–matter entanglement; and (c) the Heisenberg uncertainty principle has been tested via the squeezing of light. In all cases, the cavity enhances light–matter coupling leading to more pronounced quantum effects. Compared to atomic systems, the semiconductor has strong Coulomb correlation effects due to the relatively high density, which makes the analysis challenging. However, the semiconductor can also provide new mechanisms like the photon/carrier-density correlations which trigger new unexpected quantum phenomena.

The field of semiconductor quantum optics is very active and is developing rapidly. It benefits both from advances in computer capabilities and in semiconductor crystal growth. The increasing computer capacity allows more profound, accurate, and realistic modeling of semiconductor structures. At the same time, advances in crystal growth technology provide us with improved samples which are almost disorderless. In particular, the growth of quantum wells with narrower exciton linewidths and quantum dots with larger dipole moments and reduced dephasing rates may be achieved, which is certain to make quantum features increasingly apparent and unavoidable.

All these research efforts eventually focus on producing devices utilizing quantum mechanical principles. One of the main objectives endeavors to develop quantum logic components for building blocks of quantum computers. Similar expanding possibilities can be expected, e.g., for accuracy of detection, device efficiency, and component design in general. Considering all of this, we are almost guaranteed to see new astonishing advancements; in this bright future, semiconductor cavity QED research will most likely be a pre-eminent element.

See also: Cavity QED

Further Reading

Allen, L., Eberly, J.H., 1987. Optical Resonance and Two-Level Atoms, 1st edn. New York: Dover.
Banyai, L., Koch, S.W., 1993. Semiconductor Quantum Dots, 1st edn. Singapore: World Scientific.
Berman, P.R. (Ed.), 1994. Several articles in *Cavity Quantum Electrodynamics*., 1st edn., Boston, MA: Academic Press.

Cohen-Tannoudji, C., Dupont-Roc, J., Grynberg, G., 1989. Photons and Atoms, 3rd edn. New York: Wiley.
Haug, H., Koch, S.W., 2004. Quantum Theory of the Optical and Electronic Properties of Semiconductors, 4th edn. Singapore: World Scientific.
Khitrova, G., Gibbs, H.M., Jahnke, F., Kira, M., Koch, S.W., 1999. Nonlinear optics of normal-mode-coupling semiconductor microcavities. Reviews of Modern Physics 71, 1591–1639.
Kira, M., Jahnke, F., Hoyer, W., Koch, S.W., 1999. Quantum theory of spontaneous emission and coherent effects in semiconductor microstructures. Progress in Quantum Electronics 23, 189–279.
Scully, M.O., Zubairy, M.S., 1999. Quantum Optics, 1st edn. Cambridge, UK: Cambridge University Press.
Walls, D.F., Milburn, G.J., 1994. Quantum Optics, 1st edn. New York: Springer-Verlag.

Silicon Qubits

Thaddeus D Ladd, HRL Laboratories, LLC, Malibu, CA, United States
Malcolm S Carroll, Sandia National Laboratory, Albuquerque, NM, United States

Motivation for Silicon Qubits

The first motivation for basing qubits on silicon is the obvious foundation of classical microelectronics. Although silicon quantum computers would operate in a fundamentally different way from classical computers – for example, at cryogenic temperatures – still the level of development in material quality, crystal growth, and fabrication methodologies for silicon is unrivaled by any other material in the world. Leveraging even a small fraction of the worldwide investment in silicon for qubit development could potentially put silicon-based qubits far ahead of other solid-state alternatives.

The second, less obvious reason for choosing silicon is the remarkably clean magnetic environment witnessed by spins in highly purified and isotopically enriched silicon material. Fortuitously, 95.3% of the naturally occurring isotopes of Si nuclei (^{28}Si and ^{30}Si) are spin-0. These nuclei therefore have a "closed shell" of nuclear moments, providing no external magnetic field whatsoever. Add to this the possibility of intrinsic silicon with part-per-billion chemical quality and the system is remarkably close to "vacuum" with respect to magnetic noise properties.

The most dramatic experimental demonstrations of this "semiconductor vacuum" are from inductively and optically probed magnetic resonance experiments in bulk silicon crystals which have been isotopically enriched to contain 99.9995% ^{28}Si. These experiments probe dilute phosphorus impurities (at a density of 10^{15} cm^{-3}, i.e. one out of every 50 million atoms), which serve as donors in the material. For conduction electrons, these are closely analogous to a suspended single-charged ion, although it is worth recalling that these "ions" are suspended in a highly quiet material featuring negligible motion due to lattice vibrations without any need for laser cooling, laser trapping, or electrodes, in contrast to trapped ion or neutral atom qubit experimental setups.

In cryogenic, electron spin resonance (ESR) experiments, electron spins precess in an applied magnetic field, kicked off by a microwave pulse. The spinning electrons dephase first and foremost due to quasi-static inhomogeneities in the local magnetic field, an effect readily reversed by spin-echo techniques, which periodically invert the relative phases accrued by static rotation speed differences. At high levels of enrichment (e.g., 99.9995% ^{28}Si) and once inhomogeneity is removed as a source of dephasing, he most important term causing dephasing is the dipole-dipole couplings between the dilute phosphorus atoms themselves. By applying a magnetic field gradient, these dipole-dipole effects can also be reduced. In Tyryshkin *et al.* (2012), it is shown that electron spins may be estimated to precess in phase for nearly a minute in this material.

Phosphorus impurities are optically addressable as well. They have a hydrogen-like energy structure in the THz energy range, which is inconvenient for both electronics and optics, but still a noteworthy structure for qubit studies (Litvinenko *et al.*, 2015). They also feature a set of sharp optical transitions corresponding to the transition from an atomic-hydrogen-like neutral donor to a molecular-hydrogen-like bound excitonic state. These transitions are atomically sharp in isotopically enhanced silicon, enabling the ability to polarize and measure the hyperfine couplings to the spin of the ^{31}P nucleus. Again using radio-frequency pulses to refocus dephasing effects from inhomogeneous magnetic fields, these nuclei can be observed to precess without loss of phase coherence for hours, including at room temperature (Saeedi *et al.*, 2013).

These bulk experiments are a testament to the capability to preserve spin coherence in silicon, and are not replicated in any other material. But while a "vacuum" may be an apt description for the magnetic environment seen by conduction electrons, unfortunately the electronic structure of silicon is not completely "vacuum"-like. The electron structure of crystalline silicon produces an indirect bandgap. The energy minima for electrons in the conduction band are not at crystal momentum $k=0$, but rather along the six crystalline axes of the cubic structure, providing an anisotropy quite unlike a vacuum. These sixfold degenerate minima of the conduction band are referred to as valleys, and their degeneracy poses a problem for qubits since they can represent an uncontrolled degree of freedom for electrons that prevents clean control of qubit states. For the phosphorus impurity, the neutral donor ground state is an equal superposition of valley states which is energetically separated from other states by an energy of 11.7 meV in unstrained bulk (Zwanenburg *et al.*, 2013). While this effectively breaks the valley degeneracy for this system, a further consequence is that the atomic-like structure of the zero-phonon donor-bound exciton transitions comes with an extreme optical inefficiency: decay of the donor-bound exciton state is dominated first by Auger recombination, in which excitonic recombination ionizes the impurity, and second by broad phonon-assisted transitions.

Valley physics is more severe for electrons bound by large, shallow potentials in planar silicon devices. The valley degeneracy is partially broken by tensile strain in heterostructure stacks or by vertical electrostatic confinement in close proximity to an oxide interface, but while these effects raise four of the six valley states by 10s to 100s of meV, the remaining two valley states are much closer to degenerate. The splitting of these last two valleys is determined by atomic-scale details of the structure and the magnitude of the vertical electric field; assuring a sufficiently large value of this splitting is a key design constraint in silicon qubits.

Within a given valley or valley superposition, conduction electrons in silicon are reasonably well described by an effective mass model, with an in-plane effective mass of $0.19m_0$. This number is substantially higher than in GaAs, an earlier material studied for

spin qubits, requiring much higher demands on the lithographic resolution of structures for trapping or controlling electrons relative to GaAs. These limits are, however, within the limits of e-beam lithography and even modern implementations of photolithography. A substantial number of approaches to lithographically defined qubits have been introduced in the last 20 years. A technical review of many of these proposals and the progress on implementing them was compiled by Zwanenburg *et al.* (2013).

In this article, we highlight what the authors consider three of the most important categories of silicon qubits for silicon-based quantum computing, most of which have made substantial progress in demonstrating coherent operation since the Zwanenburg *et al.* review was published. These three principle categories of qubits are single phosphorus impurities, metal-oxide-semiconductor (MOS)-based dots, and dots based on heterostructures of strained silicon quantum wells with SiGe barriers. In what follows we will elaborate on these three essential categories, followed by categorization of the ways of controlling these qubits.

Three Principal Material Types of Silicon Qubits

The three qubit categories are summarized in **Fig. 1**. All three types ultimately rely on the spins of trapped conduction electrons, although most variations of phosphorus-based qubits heavily employ the ^{31}P nuclear spin as well. The first principle difference is the choice for how to confine electrons in the z-direction, defined as the substrate growth direction, normal to the silicon wafer. In the case of phosphorus impurities, it is the Coulomb attraction of the phosphorus impurity, which appears positively charged when replacing a silicon atom. This potential strongly attracts the electron, resulting in a binding energy of 45.6 meV (Zwanenburg *et al.*, 2013). In contrast, metal-oxide-semiconductor (MOS) quantum dots trap electrons in an electric potential defined by the hard "wall" of an oxide and the electric field created by the bending of the silicon bandgap due to the oxide interface. This potential is the same as that of the channel of a silicon MOS Field Effect Transistor (MOSFET), an extremely successful device forming the basis of modern microelectronics. The ground state of this approximately triangular potential sits at energies 10s of meV above the conduction band. SiGe quantum dots are similar to the MOS case, except in this system the barrier is pushed away from the surface of the semiconductor material, being defined instead by the band-offset between the strained silicon quantum well and a SiGe barrier, typically engineered to be larger than 150 meV (Zwanenburg *et al.*, 2013).

The trade-offs of the three approaches to vertical confinement have largely to do with disorder. A key advantage of the phosphorus system is that every phosphorus impurity provides effectively the same potential in a silicon lattice, assuming the impurity is sufficiently removed from surfaces or dielectric interfaces. This is in contrast to the quantum dot systems, where the exact energy of the vertical confinement may be impacted by random fluctuations in the amorphous oxide in the MOS case or by the heterostructure interface in the SiGe case.

A related distinguishing feature is the method for horizontal electron confinement. Here, the phosphorus system confines electrons horizontally in principally the same way as it does vertically: by its $1/r$ Coulomb potential. In contrast, the MOS and SiGe quantum dot systems rely on electrostatic gates to provide a horizontal potential profile for electrons. There are a number of strategies for designing these gates. Example scanning electron microscope (SEM), cross-sectional transmission electron microscope (TEM), and scanning tunneling microscope (STM) images of fabricated gates for all three vertical confinement types are shown in **Fig. 2**.

Most recent silicon quantum dot qubits are enhancement-mode devices, which are designed for an empty MOS or quantum well channel. A global field gate or individual dot gates pull the confined electrons from ohmic contacts. This contrasts with depletion mode structures, more typical in the AlGaAs/GaAs system, in which the system is doped such that a gas of electrons would reside in the channel region in equilibrium, but then negatively biased gate electrodes push electrons away from the dot regions, leaving behind a controlled number of charges. Depletion mode behaves poorly in silicon due to the disorder created by the dopants, which is more poorly screened in silicon than in GaAs. Besides avoiding disorder, however, enhancement-mode

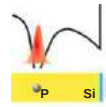
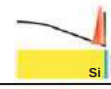
Type	Sketch	Main advantage	Main challenge
Donor		Lowest disorder and largest valley splitting	Multi-qubit fabrication and operation
MOS		Most similar to commercial transistors	Disordered potential
SiGe		Clean epitaxial barrier (low disorder)	Valley degeneracy

Fig. 1 Table indicating the three principle types of silicon qubits; the sketch column gives a rough image of the band structure as a black line, the electron wave function in red, and the material stack below. The confinement illustrated is referred to as "vertical" in the text (i.e. it the confinement in the growth direction), however the growth stack is here drawn horizontally, so that the energy of the band-structure above may be drawn vertically.

devices have the advantage that they could in principle confine electrons using only a single gate per dot. Most architectures mix enhancement and depletion operation using multiple gates per qubit. These multiple gates are useful for crafting each electron's confining potential, in part to overcome the effects of disorder in the vertically confining structure and in part to improve flexibility of electron control.

For all three silicon qubit categories, at least one form of control – under strong consideration since the seminal proposals of Loss and DiVincenzo (1998) for quantum dots and Kane (1998) for phosphorus donors – is the kinetic exchange interaction. This interaction relies on the ability to tune the energy level of a dot or donor based on a gate directly above it (called the A-gate in the Kane proposal and in this article, but often referred to as a plunger gate), and to control the interaction between nearby spins either via A-gates or with an additional gate controlling a tunnel barrier between the dots (called the J-gate in the Kane proposal and in this article, but often referred to as an exchange gate).

The kinetic exchange interaction is a consequence of quantum tunneling in the Coulomb blockade regime and Pauli exclusion (Zwanenburg *et al.*, 2013). The essential physics is that the spin-singlet state of a pair of electrons (with total angular momentum $S=0$ and an antisymmetric spin state) is capable of tunneling into the ground state for two electrons on a single dot or donor site, since the resulting state is fully antisymmetric. Spin-triplet states (with total angular momentum $S=1$ and a symmetric spin state) would, in contrast, result in a disallowed symmetric doubly-occupied ground state, and are therefore forbidden from tunneling unless the dots are detuned in energy by a sufficient amount to allow tunneling into excited orbital or valley states. The kinetic exchange interaction lowers the energy of the singlet state relative to the triplet state due to the singlet's allowed tunneling process. Electrical control of tunneling via gate voltages modulate this interaction, either by increasing the tunnel coupling directly (via J-gates) or by bringing the chemical potentials of the quantum dots or donors closer to a resonance condition (via A-gates).

Kinetic exchange provides one of the key proposed mechanisms for coupling silicon qubits, but there are others which will be addressed later in this article. Nonetheless, we will frame our discussion in terms of the basic A-gate/J-gate control of quantum dot chemical potentials and exchange energies for the sake of a meaningful comparison of the three basic categories. **Fig. 2(a)** provides the simplest visual realization of the A/J or plunger/exchange gate concept, with the visible "paddle" gates intending to serve the role of A gates and the straight gates between intending to serve the role of J gates. While harder to see in those images of dot-based devices employing overlapping gates, similar A/J functionality, in addition to horizontal confinement, is intended here as well.

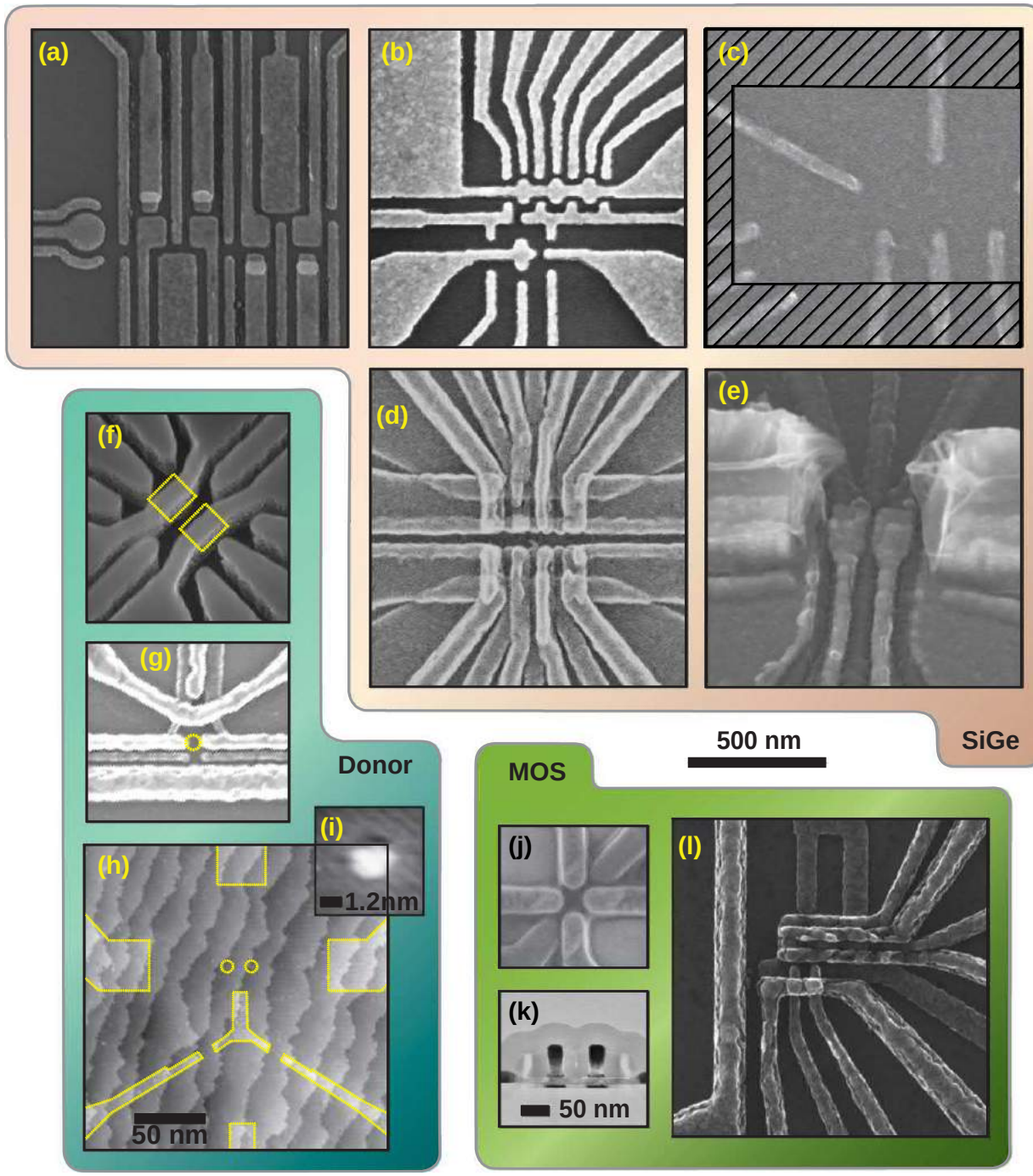
Due to their naturally occurring confining potentials, phosphorus impurities would appear to have an advantage relative to dots in that the requirement of multiple gates for horizontal confinement is relaxed. However, the phosphorus system can introduce substantially more strict fabrication requirements compared to quantum dots, especially in multi-qubit systems. The phosphorus potential results in extremely small electron wavefunctions. The Bohr radius of an electron on a donor is approximately 2.5 nm. This means that in order to implement a reasonably strong exchange interaction entirely in the bulk, theory estimates that A-gates would need to be separated by approximately 10–20 nanometer pitch (Zwanenburg *et al.*, 2013). An additional challenge is that the nondegenerate phosphorus valley state creates a lattice-scale wavefunction oscillation, which results in a kinetic exchange interaction that varies from crystal-lattice site to crystal-lattice site (Zwanenburg *et al.*, 2013). Some early donor qubit designs, as a consequence, proposed extremely challenging donor and electrode placement, both vertically and horizontally (Hollenberg *et al.*, 2006; Testolin *et al.*, 2007).

Alternative long-range coupling schemes between donors, as we describe below when we elaborate on coupling mechanisms, might be employed to circumvent many of the strict requirements of exchange-based donor qubit architectures. Such approaches could use standard silicon foundry processing, using ion implantation to place an average number of donors within the placement requirements of these proposed schemes. For exchange-based donor approaches, however, a new fabrication technique with atom-scale precision is under parallel development. Using a scanning-tunneling-microscope (STM) to do lithography on a hydrogen-terminated Si surface, combined with exposure of the patterned surfaces to phosphine, a quasi-three-dimensional molecular-beam-epitaxy capability with spatial resolution at a single lattice site has been demonstrated (Fuechsle *et al.*, 2012). See **Fig. 2(h)** and **(i)** for qubit devices including both donor qubit islands and electrode gates fabricated this way (Watson *et al.*, 2015, 2017). The benefit of STM-defined phosphorus gates is not limited to their small size and pitch: the atom-by-atom fabrication leads to conducting channels that have shown substantially lower $1/f$ voltage noise than other lithographically defined gate stacks (Shamim *et al.*, 2016), which is a critical feature given that such $1/f$ noise may ultimately limit the fidelity of silicon qubits employing exchange interactions. The extremely promising and rapidly evolving STM-lithography approach, however, is still immature relative to the much more firmly established silicon foundry processes with respect to yield, integration, and manufacturing.

The appeal of quantum dot approaches relative to the phosphorus system is the engineering advantage of defining dots entirely by traditional lithography. The MOS system is especially promising in this regard due to its close affinity to the ubiquitous complementary-MOS (CMOS) transistor. A further advantage of MOS relative to the SiGe system is that the large vertical electric field present in this system breaks the valley degeneracy from a two-fold degeneracy to valley splittings that can easily exceed 300 μeV, corresponding to temperatures of 3.5 K (Gamble *et al.*, 2016). This large energy means that thermally initialized electron states have negligible excited valley population at typical dilution refrigerator temperatures. This vertical field is highly reduced in the SiGe system, leading to a less certain energy difference between the two lowest valley states. While comfortable valley splittings in excess of 50 μeV have been observed and used for high quality initialization, control, and read-out in some devices (Eng *et al.*, 2015), in other devices the valley splitting is too small for measurable qubit behavior (Borselli *et al.*, 2011; Zajac *et al.*, 2015).

Despite the challenge of valley splitting, the primary reason to consider SiGe-based dots relative to MOS-based dots has to do with noise and disorder. MOS-based dots place the qubit wavefunction against amorphous thermal SiO_2. While this is certainly the most developed and studied semiconductor-oxide interface in history, even the highest quality oxides have intrinsic interfacial

disorder. There is inherent atomic mismatch between the two materials. MOS transistors also routinely observe $1/f$ character trap noise. Oxide disorder, therefore, might have adverse impact on the yield of MOS dots and the $1/f$ noise may limit the fidelity of MOS-dot operation. Si/SiGe dots hope to reduce the effect of interfacial disorder by pushing the qubit wavefunction away from any oxide or metal interfaces and into a region of cleaner epitaxial crystalline semiconductor. Indirect measures of disorder such as mobility and the density at which transport is first observed (i.e., the metal insulator transition) are both appreciably lower in the strained-Si channels of the SiGe-based structures. A flexible gate design in MOS may enable tuning configurations to compensate for this difference in electrostatic disorder and indeed high quality single MOS dots have been demonstrated (Yang *et al.*, 2013). It is unclear whether this can be extended to bigger multi-dot networks. With respect to separation of the electron channel from imperfect amorphous interfaces with SiGe, to date, this has not completely eliminated the presence of $1/f$ noise, presumably coming from the remote gate and dielectric stack of the device. Recent measures of charge noise in MOS quantum dots (Freeman *et al.*, 2016) and recent qubits that include a MOS interface (Harvey-Collard *et al.*, 2017) show noise amplitudes comparable to



SiGe devices (Jock *et al.*, 2017). As of the date of this article, the relative performance of MOS vs. SiGe quantum dots relative to both interfacial disorder and charge noise remains unclear.

Following the choice of how to trap electrons, qubits are further defined by how those electrons are controlled. Again, there are a few principle strategies for defining and controlling qubits which we now define, noting that they are not exclusive, as well-engineered qubits are likely to use a combination of control mechanisms.

Memory, Control, and Readout Mechanisms

Different types of silicon qubits make different choices for how to store quantum information, how to perform quantum logic, and how to read out quantum states. See **Table 1** for a summary of the qubit types which we detail below.

For memory, the first basic decision is whether to store information as charge or spin, the latter being by far the favored choice due to its much longer coherence time. However, when performing quantum logic, spin qubits are often partially converted to charge qubits. We therefore discuss both charge and spin qubits, and then proceed to indicate their control mechanisms.

Charge Qubits

The simplest form of silicon qubit is the charge qubit. For this, consider two dots (or a dot and a donor) in charge states (n,m) and $(n+1,m-1)$, where the two numbers indicate the number of conduction electrons or valence holes in each dot. The simplest such qubit would encode information on whether an electron is on the left dot (or donor) or the right, e.g. the charge states $(1,0)$ and $(0,1)$. A key motivation for this type of qubit is the simplicity of control, measurement and initialization. For control, direct modulation of gate voltages will reliably move charges between dots with speeds typically limited only by the control electronics employed. For measurement, single-electron-transistor or similar devices are capable of sensing the motion of single charges at low-temperature (Zwanenburg *et al.*, 2013; Gonzalez-Zalba *et al.*, 2015), enabling a direct read-out of this qubit's state. By sweeping voltage bias of confinement gates, broad ranges of charge numbers n and m can be detected by observing the tunneling of

Table 1 Summary of memory and control mechanisms for silicon qubits discussed in this article.

<table>
<tr><th colspan="2">Memory</th><th>Single-qubit control</th><th colspan="2">Multiqubit control</th></tr>
<tr><td colspan="2">Charge</td><td>Direct charge motion</td><td rowspan="2">Capacative couplings</td><td></td></tr>
<tr><td rowspan="4">Spin</td><td rowspan="2">Multi-spin</td><td>Exchange interactions</td><td rowspan="3">Exchange and contact hyperfine interactions</td></tr>
<tr><td>Single-triplet oscillations in gradients</td></tr>
<tr><td rowspan="2">Single-spin</td><td>ESR with AC magnetic fields</td><td rowspan="2">Magnetic dipole-dipole couplings</td></tr>
<tr><td>ESR with AC electric fields and gradients</td></tr>
</table>

Fig. 2 Sample micrographs of silicon qubits. Qubits in the pink box employ SiGe heterostructures beneath the gate structures shown for gate confinement. Qubits in the blue box employ single donors. Qubits in the green box employ MOS confinement. All images are from SEMs except (h) and (i) which are STM images and (k) which is a TEM. All images are roughly scaled to the 500 nm scalebar shown in the center of the figure, except where indicated otherwise. Devices in Fig. (a)-(c) use Ti/Au gates, while those in (d), (e), (g), and (l) employ aluminum. Devices in (f) and (g) indicate windows for phosphorus donor implantation as dashed yellow regions. (a) Quadruple dot, image courtesy Mark Eriksson, University of Wisconsin, United States; simplified and adapted from Ward *et al.* (2016). (b) Triple dot structure fabricated by Christian Volk, Center for Quantum Devices, University of Copenhagen, 2016. (c). Double dot, image courtesy Kenta Takeda, University of Tokyo; see Takeda *et al.* (2016). doi: 10.1126/sciadv.1600694. The cross hatched region indicates the region in which a micromagnet made from deposited cobalt provides a magnetic gradient across the double-dot. A similar structure, without micromagnet, was demonstrated by HRL in Maune *et al.* (2012). (d) Six dot structure using overlapping gates, image courtesy Jason Petta, Princeton University; see Zajac *et al.* (2015). A similar structure was employed at HRL in Reed *et al.* (2016). (e) Double dot structure using overlapping gate with visible cobalt micromagnet. Image courtesy the Vandersypen group, copyright TU Delft, the Netherlands. (f) Polysilicon quantum dot and readout gate structure from Sandia National Laboratory; the quantum dot couples to an underlying phosphorus donor; see Harvey-Collard *et al.* (2017). (g) Gates for SET readout-channel for a single phosphorus device, courtesy Andrea Morello, UNSW; see Muhonen *et al.* (2014). (h) A phosphorus device in which both the qubit and the gates, outlined by dashed yellow lines, are fabricated by STM lithography. Courtesy of T.F. Watson from the Simmons Group, Center of Excellence for Quantum Computation and Communication Technology; see Watson *et al.* (2015). (i) A zoom into the yellow circle of the device in (h). (j) Quadruple quantum dot device in a fully depleted silicon-on-insulator (FDSOI) nanowire field-effect transistor. The gates shown are polysilicon separated by silicon nitride; electrons accumulate in corner states between the underlying silicon nanowire and an oxide surrounding them. Image courtesy Fernando Gonzalez-Zalba, Hitachi Cambridge Laboratory; see Betz *et al.* (2016). (k) TEM of a CMOS nanowire in the source/drain direction; the central features are two gates made out of TiN/polysilicon on SiO_2 and Hf-based dielectric. They are surrounded by silicon nitride spacers. See Maurand *et al.* (2016). (l) Aluminum MOS triple-dot structure. See Veldhorst *et al.* (2014). Image courtesy Andrew Dzurak, UNSW.

single charges from nearby baths, enabling initialization into a single (n,m) state. The tunneling amplitude between a pair of dots may be voltage-controlled via the energy detuning or the size of the electrostatic tunnel barrier between the dots; indeed driving voltages at different rates may enable any superposition of qubit states via voltage control and appropriately controlled Landau-Zener-Stueckelberg oscillations (Ward *et al.*, 2016; Gonzalez-Zalba *et al.*, 2016).

Charge qubits are often used to study basic properties of quantum dots (Wang *et al.*, 2013), as well as provide proving grounds for the electrostatic interactions which may provide controlled-logic gates between qubits. For example, Ward *et al.* (2016) indicates a controlled-phase gate between a pair of single-electron charge qubits defined in a Si/SiGe heterostructure, using the device layout shown in **Fig. 2(a)**.

The disadvantage of a charge qubit is its sensitivity to charge noise. Charge noise may result from any noisy electric field, which may originate from fluctuators in the materials used in the device (e.g. $1/f$ noise from dielectrics) or from the control gates (e.g. Johnson noise from control circuitry). While the magnitude of charge noise will vary quite a bit between devices and experimental set-ups, typical dephasing times of charge qubits due to charge noise is in the range of 0.1–10 nanoseconds (Zwanenburg *et al.*, 2013). Given typical constraints of the electronics used to control qubits, these timescales are simply too fast for high fidelity operation.

Spin Qubits

A comparably simple qubit is the single-spin qubit. A single electron spin bound to a donor or to a quantum dot is naturally a two-level quantum system, corresponding to spin-up and spin-down. As a fundamental magnetic dipole, a bound spin-1/2 has no sensitivity to electric fields, and interacts only magnetically. In silicon and silicon heterostructures, conduction-band electrons have a g-factor very close to 2, indicative of the small bulk spin-orbit interaction in silicon. Spin relaxation times typically exceed seconds at reasonable magnetic fields (Zwanenburg *et al.*, 2013).

The key advantage of the single-spin implementation relative to charge-qubit implementations is access to the long coherence times as observed in bulk systems. Recently, a single phosphorus qubit with single-spin encoding, implemented in silicon isotopically enhanced to contain only 800 ppm ^{29}Si, showed a static dephasing time of 270 μs, extended to a multiple-spin-echo decoherence time exceeding 30 s (Muhonen *et al.*, 2014), despite its proximity to the aluminum gates forming the readout mechanism (see **Fig. 2(g)**). Similarly, a single spin qubit in an MOS dot, similar to that in **Fig. 2(l)**, also in 800 ppm ^{29}Si, showed comparable times with static dephasing of 120 μs extended to a multiple-spin-echo decoherence time of 28 ms (Veldhorst *et al.*, 2014). Single-spin control for this system inherits the high-fidelity control of spin-resonance, enabling single-qubit gates with 99.6% fidelity as measured by randomized benchmarking (Veldhorst *et al.*, 2014).

A key consideration for single-spin encodings is that they typically require a large magnetic field, and their control requires careful tracking of each spin's Larmor precession in order to achieve phase-synchronous control. Ultimately, some limit to qubit coherence will exist due to the stability of the applied magnetic field or of the local oscillator. It is possible to circumvent these issues while still maintaining the coherence of spin qubits by using multi-spin qubits, in which qubit states are encoded by the total angular momentum of a set of spins. For example, the two-spin singlet-triplet qubit encodes information in the total angular momentum $S=0$ (singlet) and $S=1$ (triplet) subspaces of the two spins. A three-spin encoding combines a singlet-triplet pair with a third spin, assuring a total angular momentum $S=1/2$; a four-spin encoding likewise assures a total angular momentum $S=0$. These encodings are used in a way which imposes a constant total projection of angular momentum for all qubit states. In this way the qubit spin states can be brought to degeneracy for use as memory, assuring that the states do not evolve any phases relative to one another due to any constant or globally fluctuating magnetic fields. For this reason these qubits are referred to as decoherence-free subspaces or subsystems (DFS) (DiVincenzo *et al.*, 2000).

Spin Control by ESR and EDSR: To control single-spins, an obvious option is to use magnetic resonance, in which a large static applied magnetic field is present and transverse microwave magnetic fields are applied resonant with the electron Larmor precession of about 28 GHz per tesla of applied field. A clear technical difficulty with this mode of operation is that the controlling microwaves will typically not be localized, affecting multiple spins at once. To circumvent this difficulty, individual spins must have their resonant frequency shifted relative to their neighbors. In phosphorus devices, such a shift is available from the longitudinal component of the gate-controlled hyperfine interaction with the ^{31}P nucleus (Kane, 1998; Laucht *et al.*, 2015), enabling resonance shifts of order MHz for reasonable gate voltage variations. In MOS and SiGe devices, g-factor inhomogeneity has been shown to result from spin-orbit effects in the quantum dot barrier, enabling dot-to-dot and voltage-tunable g-factor variations again on the order of MHz (Kawakami *et al.*, 2014; Veldhorst *et al.*, 2015).

Even more localized microwave control is possible using electric dipole spin resonance (EDSR), in which magnetic field gradients via micromagnets or spin-orbit effects convert AC electric fields to magnetic fields. If a magnetic field gradient is present, then voltage-induced motion of the electron in that gradient may mimic the effect of transverse microwave fields, except with the clear advantage of maintaining localization to the electron undergoing spatial modulation. In SiGe devices, fabricated cobalt micromagnets have provided this gradient (Takeda *et al.*, 2016; Kawakami *et al.*, 2014), as shown in the devices of **Fig. 2(c)** and **(e)**. Another possibility here is the use of hole spins, which naturally have a stronger spin-orbit interaction in confined nanostructures such as that in **Fig. 2(k)**, enabling EDSR without additional gradients (Maurand *et al.*, 2016).

Spin Control by Kinetic Exchange: The kinetic exchange mechanism enables a conversion between charge qubits (including, unfortunately, sensitivity to charge noise) and spin-qubits. It does this by using voltages to attempt to generate a superposition of (2,0) and (1,1) charge states (reminiscent of the charge qubit), the amplitudes and energies of which are different for

Pauli-allowed spin singlets and Pauli-excluded spin triplets in the (2,0) state. This spin-to-charge conversion is useful since it enables initialization, control, and measurement of spin-qubits without requiring high magnetic fields or magnetic field gradients.

The action of kinetic exchange is to reduce the energy of a singlet, which may be considered a single-qubit control axis for multiple spins encoded in a DFS. For a pair of spins in a singlet-triplet qubit, this voltage-induced interaction may therefore be considered a rotation about the z-axis of the multi-spin qubit's Bloch sphere. To control other axes in this case, another mechanism is needed. Examples of successful introduction of a second interaction for the additional control axes are magnetic gradients, arising from either deliberately introduced magnetic field gradients (Takeda *et al.*, 2016), single nuclear spin (Harvey-Collard *et al.*, 2017), or an ensemble of nuclear spins (Maune *et al.*, 2012). The latter case has been a fruitful qubit implementation in the nuclear-spin-rich system of GaAs (Nichol *et al.*, 2017).

Remarkably, if DFS encodings are used employing at least three spins per qubit, the exchange interaction can provide universal quantum logic, including multi-qubit gates, with no other control mechanisms. This was shown in DiVincenzo *et al.* (2000), which shows a pulse sequence locally equivalent to controlled-phase for two three-spin encoded DFS qubits. More recently, pulse sequences have been found which provide exchange-only multiqubit control without any specification made on the m-quantum number for either qubit, potentially allowing operation in much lower magnetic fields (Fong and Wandzura, 2011). This exchange-only encoding works best in highly uniform magnetic fields, and is particularly important for silicon where nuclear-induced gradients may be made very small via the use of isotopic enhancement. The basic operation of exchange-only triple-dots qubits in 800 ppm ^{29}Si, using a gate layout similar to that in **Fig. 2(d)**, has been demonstrated in Eng *et al.* (2015) and Reed *et al.* (2016). This encoded triple-dot spin qubit showed a ^{29}Si-limited static dephasing time of 3 μs, extended via exchange echo to over 600 μs at high magnetic field (Eng *et al.*, 2015), with operational fidelity limited by charge noise (Reed *et al.*, 2016).

Initialization and Readout of Spin Qubits: While the inductive techniques of magnetic resonance are excellent and highly mature for control, the readout-mechanisms of traditional magnetic resonance are many orders of magnitude too insensitive. One effective way to initialize and measure a single spin is via spin-selective tunneling to or from a bath. In particular, if a bath of electrons has its Fermi energy set via voltage to be between the two Zeeman sublevels of the dot spin, one spin will have available states for tunneling and the other will not. This basic principle depends on operating at a field and temperature such that the electron Zeeman energy $g\mu_B B$ vastly exceeds thermal energy $k_B T$, where the temperature here is that of the electrons in the bath (including any noise introduced from electronics, etc.). This is reasonable at fields of about a tesla, which also corresponds to convenient microwave ESR control frequencies of 3–30 GHz. This method of measurement has been employed for single-shot single-spin qubits in MOS (Veldhorst *et al.*, 2015), phosphorus (Muhonen *et al.*, 2014; Watson *et al.*, 2017), and SiGe (Kawakami *et al.*, 2014) systems.

Multi-spin encodings enable initialization and measurement using the Pauli-blockade mechanism, which enables the preparation and detection of singlets of spin pairs. This mechanism depends on distinguishing spin-dependent electron tunneling regimes, which requires electron temperatures much less than any excited state in the system. For SiGe qubits, low-lying valley states can easily prohibit this type of measurement, but nonetheless Pauli-blockade singlet-triplet measurement has been demonstrated in both SiGe and MOS dots (Zwanenburg *et al.*, 2013). Attempts to partially circumvent the effects of small valley splitting by operating in a filled shell (i.e., (3,1)–(4,0) occupation) have also been shown in coupled donor-dot and multi-dot systems (Harvey-Collard *et al.*, 2017).

Qubit Couplings and Hybridizations

Coupling of qubits is necessary for computation. Four predominant coupling mechanisms in silicon qubit device architectures are: (1) contact hyperfine between electron and nuclear spins, (2) capacitive or electric dipole coupling; (3) spin dipole coupling; and (4) kinetic exchange. Shuttling of electrons has been considered as a complementary function with coupling mechanisms for donors and quantum dots. This is especially of interest for short range mechanisms that can restrict layout flexibility such as exchange and contact hyperfine (Skinner *et al.*, 2003; Witzel *et al.*, 2015; Pica *et al.*, 2016). Shuttling of electrons through quantum dot networks has been experimentally demonstrated in GaAs (Baart *et al.*, 2016; Fujita *et al.*, 2017).

The selection of coupling mechanism has a strong influence on both the physical manifestation (e.g., layout and signal control) as well as performance (e.g., speed and dominant error sources). Coupling mechanisms, less obviously, also allow redefinition of the single qubit encoding through qubit hybridizations. Hybridizations often seek "the best of both worlds," circumventing a challenge of one of the principal encodings through merging properties of a second qubit's properties. The invention of the transmon for Josephson-junction-based qubits is an example of the extraordinary success that can come from hybridization (Houck *et al.*, 2009).

Contact Hyperfine Interaction

The Fermi contact hyperfine interaction is ubiquitously important for both quantum dot and donor qubits because it is the dominant mechanism through which nuclear spins interact with the electron spin (i.e., either directly through a strong coupling to a ^{31}P nucleus or through overlap with trace ^{29}Si or ^{73}Ge). The contact hyperfine term, therefore, plays a role in the Hamiltonian of any qubit leading to contributions to line width or resonant frequency. In donors for which the electron spin is the primary

encoded qubit, as discussed earlier, the contact hyperfine is used as a tuning mechanism through the Stark shift (Kane, 1998; Laucht *et al.*, 2015).

Alternatively, the nuclear spin of the phosphorus atom has been proposed as either the data or memory qubit in several computing schemes (for example see Kane (1998), Skinner *et al.* (2003), Hill *et al.* (2015), Witzel *et al.* (2015), Pica *et al.* (2016)), due to several important advantages: it features extremely long coherence times, it can be isolated from the electric environment when the donor is ionized, and very high fidelity NMR rotations are possible. Furthermore, it naturally occurs with 100% abundance as the spin-1/2 ^{31}P isotope and it is a qubit that, therefore, cannot be lost (i.e., there is always a spin ½ system present). A strong coupling between the electron and a nucleus is "built-in" to the donor system, with a substantial amplitude of 117.53 MHz (Zwanenburg *et al.*, 2013). This provides a convenient spin-spin coupling to this long-lived quantum memory.

The Kane proposal (Kane, 1998) and many variations of it published since, have indicated device architectures using electrons to mediate interactions between donor nuclear spins through the contact hyperfine interaction, although of course all proposals further need a mechanism for donor-donor coupling either through shuttling or other mechanisms that we discuss below. It is also worthwhile to note that the hyperfine contact term has already enabled repeated quantum-non-demolition measurement of the nuclear spin, offering extremely high initialization and read-out fidelity, while magnetic resonance techniques offer excellent control fidelity (Muhonen *et al.*, 2014). As an example process exploiting these high fidelity operations, Bell inequality violations in this electron-nuclear system were recently demonstrated (Dehollain *et al.*, 2016).

Spin Qubit Coupling with Exchange

In the highly influential Loss/DiVincenzo (Loss and DiVincenzo, 1998) and Kane proposals (Kane, 1998), the kinetic exchange interaction is used as the entangling mechanism between single-spin qubits. As the name suggests, this interaction "exchanges" spins, enabling a "full swap" if tuned to a π-pulse. Spin-information can be moved across a device by a series of such swaps. Of course, if pulsed for half the duration, spins may be entangled. The sqrt-SWAP interaction between a pair of spins provides maximal entanglement; a controlled-NOT (CX) or controlled-phase (CZ) gate could then be accomplished using two sqrt-SWAP gates, with the additional use of single-spin rotations implemented via the methods described in the previous sections. Alternatively, exchange in the presence of Zeeman inhomogeneity due either to micromagnetic gradient fields or g-factor inhomogeneity can provide a CZ gate. Two-qubit gates for single spin qubits were recently demonstrated this way in MOS dots (Veldhorst *et al.*, 2015) and SiGe dots (Watson *et al.*, 2017; Zajac *et al.*, 2017). Sensitivity to charge noise was observed, providing one of the key present challenges in improving control fidelity of this type of coupling in QDs.

Exchange is a short-range interaction of order of the size of the electron wavefunctions. A possible longer term challenge for quantum dots and an immediate challenge for donors is the lithographic layout of the dense packing of electrodes needed to control the tunnel couplings, establish electron occupation, and provide readout and initialization capabilities. More complex layouts, such as extensions from present 1D layouts to 2D layouts, introduces challenges for exchange based architectures (Veldhorst *et al.*, 2016; Vandersypen *et al.*, 2017). For direct coupling of donor electron spins, the lithographic challenge is exaggerated both because of the more tightly confined electron wavefunction in the bulk and because of crystal-orientation-dependent tunnel coupling strengths. An alternate approach to couple donors through ionizing or hybridizing the donor electron to a larger surface quantum dot state (Kane, 1998; Calderon *et al.*, 2009) has been proposed to ease the lithographic requirements; recent experimental validations in this direction demonstrate coherent donor-quantum-dot hybrids (Harvey-Collard, *et al.*, 2017).

Variations of multispin and charge qubit hybridization can enable a variety of possibilities. One example is an encoding that resembles the triple-dot exchange-only qubit, but employs three electrons in two dots. An excited valley-orbit state in one of the two-dots effectively provides the third state of exchange-only control, with its energy splitting acting as an effectively constant exchange interaction. One axis of control is provided by voltage-controlled exchange-like interactions, while the other is provided by this excited-state splitting. Despite relying on what is ultimately a charge-state splitting for one axis of control, the coherence in these qubits can be extended to time scales of microseconds at the cost of substantially slowing down the qubit rotation times (Thorgrimsson *et al.*, 2017). Another example combining exchange with single spin qubits is that single spin qubit encodings can be combined with the singlet-triplet-readout methods of exchange qubits to provide direct means to make the parity measurements associated with quantum error correction (Jones *et al.*, 2016; Veldhorst *et al.*, 2016).

Capacitive and Electric Dipole Coupling

Charge qubits, as well as spin-qubit encodings with a charge qubit component, offer a natural long-range coupling scheme. The electric field created by charge displacement in one qubit can be used to control the state by displacing the charge of another qubit. At short range, this is effectively a quantum cross-capacitance effect; at larger distances, it has the character of an electric dipole-dipole coupling. Capacitive coupling has enabled multiqubit logic in silicon-based charge qubits, such as the one shown in **Fig. 2(a)** (Ward *et al.*, 2016). Multispin qubits may be coupled this way, for example between singlet-triplet qubits that exploit the dipole difference between the (2,0) and (1,1) charge states to influence a neighboring qubit's charge states. The approach has been successfully demonstrated with nearest neighbor qubits reaching approximately 90% entanglement fidelity in a GaAs system (Nichol *et al.*, 2017). The distance between the qubits was approximately 500 nm in this case providing a modest relaxation of space constraints compared to exchange coupling.

Such coupling mechanisms may be especially pertinent for donor-based systems, since exchange couplings are challenging to engineer in this system. In one proposal, the electric dipole of a phosphorus impurity is "stretched" by the action of an A-gate, enabling electric control of a long-distance dipole-dipole coupling. Since this long-range coupling has a weak spatial dependence in comparison to exchange, these coupling mechanisms may allow phosphorus to be fabricated through a controlled ion implantation process, with the inevitable placement straggle compensated for by gate calibration (Tosi *et al.*, 2017). An electric dipole may also be created through the combination of a phosphorus impurity and an exchange-coupled quantum dot above it. Coherent quantum dot coupling to a single donor potential has been recently demonstrated, including characterization of the charge noise in that system (Urdampilleta *et al.*, 2015; Harvey-Collard *et al.*, 2017). Key to the success of these electric dipole coupling approaches is assuring that the mechanisms which provide the charge sensitivity for long-distance coupling may be rapidly modulated to prevent undue decoherence from charge noise.

Capacitive couplings may be enhanced or lengthened using extra hardware. Increasing the coupling range using floating gates has been proposed (Trifunovic *et al.*, 2012), although the small predicted effects of dissipation in these gates awaits experimental validation. Superconducting coupling elements seem especially promising, since the coupling of charge qubit components of qubits and the management of charge noise have been well addressed in the superconducting qubit community. These ideas are beginning to be adopted by silicon qubits, and superconductors may offer a powerful spin-charge hybridization offering benefits such as longer-range coupling. Conversion of exchange qubits to charge qubits allow charge-induced coupling to microwave fields in superconducting resonators, providing rapid spin-spin interactions across a chip; the recently demonstrated strong coupling of a single electron charge qubit in a SiGe device to a superconducting resonator is a critical step in this direction (Mi *et al.*, 2017).

Electric dipoles with optical oscillation frequencies could enable optical connections between silicon qubits, which would be especially convenient for those applications in optical quantum communication requiring memory and quantum logic. Unfortunately, silicon's indirect bandgap drastically reduces this system's efficiency for most existing concepts for spin-photon entanglement employing near-bandgap excitons, as demonstrated in III-V semiconductors. However, optically efficient emitters do exist in silicon and may provide future qubits with practical optical interfaces; a recent proposal employing chalcogen donors is provided in Morse *et al.* (2017).

Magnetic Dipole Coupling

One of the potential advantages of single spin encodings is the long decoherence times due to the relatively good decoupling from environmental factors like charge noise. One approach to fully realize the long coherence times of spins is to altogether avoid all coupling schemes that have a charge qubit component. The spin is a magnetic dipole and the magnetic dipole-dipole interaction therefore provides a mechanism of coupling qubits not subject to charge noise. A critical challenge in exploiting dipole-dipole couplings is that they are long-range and "always on," making scheduled control challenging. Also they are slow, typically requiring millisecond interaction times. Nonetheless, this mechanism is appealing to donors because the dipole-dipole coupling avoids the atomic precision fabrication requirement for exchange and a system may exploit the long memory times of ^{31}P nuclear spins. An early proposal for dipole coupled donor quantum computing argues that using a long range "always on" component is manageable combined with g-factor tuning (de Sousa *et al.*, 2004). The Stark shift of the donor contact hyperfine and g-factor has been modelled (Rahman *et al.*, 2009) and Stark shifted resonant frequency have been experimentally demonstrated both in ensembles and single donor cases (Wolfowicz *et al.*, 2014; Laucht *et al.*, 2015). A more recent architecture proposes a more direct modulation of the dipolar interaction through use of precisely timed ionization and deionization of donors to control the magnetic dipole-dipole coupling, holding information on the associated ^{31}P nuclear spins (Hill *et al.*, 2015). This can substantially reduce the complication of long range coupling. In another recent proposal, mechanical motion of a scanning stage of ^{31}P modulates the dipole-dipole coupling strengths (O'Gorman *et al.*, 2016). Although these architectures no longer require atomic-scale placement and gating, they still depend on highly uniform energy levels for a regular array of gated donors and the two qubit rotations are slower because of the weaker interaction strengths.

The Future of Silicon Qubit Systems

In this article, we have outlined three principal ways of fabricating single-electron silicon systems, and three principal ways of encoding qubits on those systems. These fabrication and encoding techniques can furthermore be hybridized and combined producing improved features. Obviously, a large amount of design space exists for constructing systems of silicon qubits, and it remains an active, worldwide area of research to optimize the many trade-offs between the many possibilities. At the time of writing, the largest published coherently coupled silicon qubit system contains only two entangled qubits, leaving a long road ahead to the size and scale required for fault-tolerant quantum computing. Many concepts have been proposed for ways to scale, although many fundamental demonstrations remain to be proven prior to assurance that any such scaling path will succeed. However, the incredible scaling success of classical silicon microprocessors based on CMOS provides continual and dramatic encouragement of the notion that once the basic problems of charge-noise limits in control fidelities and valley- or disorder-limits in device yield are solved, a viable path for scaling to a useful technology will be found.

Acknowledgement

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

See also: Cavity QED

References

Baart, T.A., *et al.*, 2016. Single-spin CCD. Nat. Nanotechnol. 11, 330.
Betz, A.C., *et al.*, 2016. Reconfigurable quadruple quantum dots in a silicon nanowire transistor. Appl. Phys. Lett. 108, 203108.
Borselli, M.G., *et al.*, 2011. Pauli spin blockade in undoped Si/SiGe two-electron double quantum dots. Appl. Phys. Lett. 99, 063109.
Calderon, M.J., Saraiva, A., Koiller, B., Das Sarma, S., 2009. Quantum control and manipulation of donor electrons in Si-based quantum computing. J. Appl. Phys. 105, 122410.
Dehollain, J.P., *et al.*, 2016. Bell's inequality violation with spins in silicon. Nat. Nanotechnol. 11, 242.
de Sousa, R., *et al.*, 2004. Silicon quantum computation on magnetic dipolar coupling. Phys. Rev. A 70, 052304.
DiVincenzo, D.P., *et al.*, 2000. Universal quantum computation with the exchange interaction. Nature 408, 339.
Eng, K., *et al.*, 2015. Isotopically enhanced triple-quantum-dot qubit. Sci. Adv. 1, e1500214.
Fong, B.H., Wandzura, S.M., 2011. Universal quantum computation and leakage reduction in the 3-qubit decoherence free subsystem. Quantum. Inf. Comput. 11, 1003–1018.
Freeman, B.M., Schoenfield, J.S., Jiang, H.W., 2016. Comparison of low frequency charge noise in identically patterned Si/SiO_2 and Si/SiGe quantum dots. Appl. Phys. Lett. 108, 253108.
Fuechsle, M., *et al.*, 2012. A single atom transistor. Nat. Nanotechnol. 7, 242.
Fujita, T., Baart, T.A., Reichl, C., Wegscheider, W., Vandersypen, L.M.K., 2017. Coherent shuttle of electron-spin states. npj Quantum Inf. 3, 22.
Gamble, J.K., *et al.*, 2016. Valley splitting of single-electron Si MOS quantum dots. Appl. Phys. Lett. 109, 253101.
Gonzalez-Zalba, M.F., Barraud, S., Ferguson, A.J., Betz, A.C., 2015. Probing the limits of gate-based charge sensing. Nat. Commun. 6, 6084.
Gonzalez-Zalba, M.F., *et al.*, 2016. Gate-sensing coherent charge oscillations in a silicon field-effect transistor. Nano Lett. 16, 1614.
Harvey-Collard, P., *et al.*, 2017. Coherent coupling between a quantum dot and a donor in silicon. Nat. Commun. 8, 1029.
Hill, C.D., *et al.*, 2015. A surface code quantum computer in silicon. Sci. Adv. 1, e1500707.
Hollenberg, L., *et al.*, 2006. Two-dimensional architectures for donor-based quantum computing. Phys. Rev. B 74, 045311.
Houck, A.A., Koch, J., Devoret, M.H., Girvin, S.M., Schoelkopf, R.J., 2009. Life after charge noise: Recent results with transmon qubits. Quantum Inf. Process 8, 105.
Jock, R.M., *et al.*, 2017. Probing low noise at the MOS interface with a spin-orbit qubit. arXiv:1707.04357.
Jones, C., *et al.*, 2016. A logical qubit in a linear array of semiconductor quantum dots. arXiv:1608.06335.
Kane, B.E., 1998. A silicon-based nuclear spin quantum computer. Nature 393, 133.
Kawakami, E., *et al.*, 2014. Electrical control of a long-lived spin qubit in a Si/SiGe quantum dot. Nat. Nanotechnol. 9, 666.
Laucht, A., *et al.*, 2015. Electrically controlling single-spin qubits in a continuous microwave field. Sci. Adv. 1, e1500022.
Litvinenko, K.L., *et al.*, 2015. Coherent creation and destruction of orbital wavepackets in Si:P with electrical and optical read-out. Nat. Commun. 6, 6549.
Loss, D., DiVincenzo, D.P., 1998. Quantum computation with quantum dots. Phys. Rev. A 57, 120.
Maune, B.M., *et al.*, 2012. Coherent singlet-triplet oscillations in a silicon-based double quantum dot. Nature 481, 344.
Maurand, R., *et al.*, 2016. A CMOS silicon spin qubit. Nat. Commun. 7, 13575.
Mi, X., *et al.*, 2017. Strong coupling of a single electron in silicon to a microwave photon. Science 355, 156.
Morse, K.J., 2017. A photonic platform for donor spin qubits in silicon. Sci. Adv. 3, e1700930.
Muhonen, J.T., *et al.*, 2014. Storing quantum information for 30 seconds in a nanoelectronic device. Nat. Nanotechnol. 9, 986.
Nichol, J.M., *et al.*, 2017. High-fidelity entangling gate for double-quantum-dot spin qubits. npj Quantum Inf. 3. 3
O'Gorman, J., *et al.*, 2016. A silicon-based surface code quantum computer. npj Quantum Inf. 2, 15019.
Pica, G., *et al.*, 2016. Surface code architecture for donors and dots in silicon with imprecise and nonuniform qubit couplings. Phys. Rev. B 93, 034306.
Rahman, R., *et al.*, 2009. Gate-induced g-factor control and dimensional transition for donors in multivalley semiconductors. Phys. Rev. B 80, 155301.
Reed, M.D., *et al.*, 2016. Reduced sensitivity to charge noise in semiconductor spin qubits via symmetric operation. Phys. Rev. Lett. 116, 110402.
Saeedi, K., *et al.*, 2013. Room-temperature quantum bit storage exceeding 39 minutes using ionized donors in silicon-28. Science 342, 830.
Shamim, S., *et al.*, 2016. Ultralow-noise atomic-scale structures for quantum circuitry in silicon. Nano Lett. 16, 5779.
Skinner, A.J., Davenport, M.E., Kane, B.E., 2003. Hydrogenic spin quantum computing in silicon: A digital approach. Phys. Rev. Lett. 90, 087901.
Takeda, K., *et al.*, 2016. A fault-tolerant addressable spin qubit in a natural silicon quantum dot. Sci. Adv. 2, e1600694.
Testolin, M.J., *et al.*, 2007. Robust controlled-NOT gate in the presence of large fabrication-induced variations of the exchange interaction strength. Phys. Rev. A 76, 012302.
Thorgrimsson, B., *et al.*, 2017. Extending the coherence of a quantum dot hybrid qubit. npj Quantum Inf. 3, 32.
Tosi, G., *et al.*, 2015. Silicon quantum processor with robust long-distance qubit couplings. Nat. Commun. 8, 450.
Trifunovic, L., *et al.*, 2012. Long-distance spin-spin coupling via floating gates. Phys. Rev. X 2, 011006.
Tyryshkin, A.M., *et al.*, 2012. Electron spin coherence exceeding seconds in high-purity silicon. Nat. Mater. 11, 143.
Urdampilleta, R.M., *et al.*, 2015. Charge dynamics and spin blockade in a dybrid double quantum dot in silicon. Phys. Rev. X 5, 031024.
Vandersypen, L.M.K., *et al.*, 2016. Interfacing spin qubits in quantum dots and donors – hot, dense and coherent. npj Quantum Inf. 3, 34.
Veldhorst, M., *et al.*, 2014. An addressable quantum dot qubit with fault-tolerant control-fidelity. Nat. Nanotechnol. 9, 981.
Veldhorst, M., *et al.*, 2015. A two-qubit logic gate in silicon. Nature 526, 410.
Veldhorst, M., Eenink, H.G.J., Yang, C.H., Dzurak, A.S., 2016. Silicon CMOS architecture for a spin-based quantum computer. arXiv:1609.09700.
Wang, K., Payette, C., Dovzhenko, Y., Deelman, P.W., Petta, J.R., 2013. Charge relaxation in a single electron Si/SiGe double quantum dot. Phys. Rev. Lett. 111, 046801.
Ward, D.R., *et al.*, 2016. State-conditional coherent charge qubit oscillations in a Si/SiGe quadruple quantum dot. npj Quantum Inf. 2, 16032.
Watson, T.F., *et al.*, 2015. High-fidelity rapid initialization and read-out of an electron spin via the single donor d-charge state. Phys. Rev. Lett. 115, 166806.
Watson, T.F., *et al.*, 2017. A programmable two-qubit quantum processor in silicon. arXiv:1708.04214.
Witzel, W.M., Montaño, I., Muller, R.P., Carroll, M.S., 2015. Multi-qubit gates protected by adiabaticity and dynamical decoupling applicable to donor qubits in silicon. Phys. Rev. B 92, 081407(R).
Wolfowicz, G., *et al.*, 2014. Conditional control of donor nuclear spins in silicon using Stark shifts. Phys. Rev. Lett. 113, 157601.

Yang, C.H., Rossi, A., Ruskov, R., *et al.*, 2013. Spin-valley lifetimes in a silicon quantum dot with tunable valley splitting. Nat. Commun. 4, 2069.
Zajac, D.M., Hazard, T.M., Mi, X., Wang, K., Petta, J.R., 2015. A reconfigurable gate architecture for Si/SiGe quantum dots. Appl. Phys. Lett. 106, 223507.
Zwanenburg, F.A., *et al.*, 2013. Silicon quantum electronics. Rev. Mod. Phys. 85, 961.
Zajac D.M., *et al.*, 2017. Quantum CNOT gate for spins in silicon. arXiv:1708.03530.

Entanglement and Quantum Information

PG Kwiat, University of Illinois at Urbana-Champaign, Urbana, IL, USA
DFV James, Los Alamos National Laboratory, Los Alamos, NM, USA

Glossary

Entanglement a property of quantum systems consisting of two or more distinct subsystems, often separate particles. When transformations on one system do not affect the other, the state is said to be separable; when this is not the case, and the quantum state of the overall system cannot be resolved into separate states of the individual pieces, the system is entangled.

Pure and mixed states when the probability amplitudes specifying a quantum state are deterministic, the state is said to be pure; when the amplitudes are random quantities, the state is mixed. The distinction is similar to that of coherent and partially coherent optical fields.

Quantum computer an information processing device that exploits quantum mechanical phenomena to greatly enhance computational power for certain problems. Some mathematical problems thought to be intractable on conventional computers can, in theory, be performed efficiently on a quantum computer.

Quantum crytography (also known as quantum key distribution) a quantum information protocol by which two parties acquire a shared series of random numbers (a cryptographic key) by exploiting the quantum nature of light. Absolute security can be proven by virtue of the indivisibility and uncopy-ability of individual quanta.

Quantum state tomography a means by which quantum states can be determined experimentally by performing a series of appropriate measurements on multiple identically prepared systems. From such measurements, the elements of the density matrix, which fully specifies the state, may be inferred.

Quantum superdense coding a quantum information protocol by which two bits of classical information may be communicated by a single quantum bit, initially part of an appropriate entangled quantum state.

Quantum teleportation a quantum information protocol by which the unknown quantum state of one particle can be transferred to another distant particle, using a pair of entangled particles, a projective measurement, and exchange of two bits of classical information.

Qubit (quantum bit) a two-level quantum mechanical system, which constitutes the building blocks of quantum information processing devices. Examples include a spin-1/2 particle, an atom with two well-isolated levels, or the polarization of a photon.

Quantum information is an emerging field of technology that encompasses the application of fundamental quantum mechanical phenomena, such as entanglement, to tasks in information processing and communications. The importance of optical technology in quantum information is not surprising, given the success of quantum optics in the investigation of these fundamental phenomena and the pre-eminent role of optics in modern communications. This article describes some of the important advances in photonic quantum information and discusses the prospects for the future.

Basic Notions: Qubits and Entanglement

Qubits

Quantum information is usually confined to two-level quantum systems, which are referred to as quantum bits, or qubits. The qubits form the register of a quantum computer or carry the information in a quantum communications channel. Physical examples of qubits under active study include the spin degrees of freedom of an electron or spin-1/2 nucleus, two well-isolated energy levels of an atom or trapped ion, and the polarization degrees of freedom of a photon; here we concentrate on the latter (Recently some attention has been devoted to quantum information based on continuous variables, such as those arising in connection with 'squeezed light'. However, here we will focus on discrete systems, i.e., qubits.). The two states of each qubit are generically denoted by $|0\rangle$ and $|1\rangle$ (although, because of our concentration on optical polarization, we shall use the horizontal and vertical states $|H\rangle$ and $|V\rangle$ interchangeably with $|0\rangle$ and $|1\rangle$ when describing experiments). An arbitrary state of a single qubit, $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, is specified by the complex probability amplitudes α and β associated with the two possible states, constrained by $|\alpha|^2 + |\beta|^2 = 1$. States for which α and β are deterministic quantities are referred to as pure states; when they are stochastic, the state is mixed, and must be described by a density matrix, e.g., $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$, where p_i is the probability of the pure state $|\psi_i\rangle$. Pure states are a useful but unattainable ideal, just like a perfectly coherent laser beam; in reality even the best-prepared quantum states have some level of mixedness, which can be quantified by the entropy: $S \equiv -Tr(\rho \ln \rho) = -\sum_i \lambda_i \ln \lambda_i$, where λ_i are the eigenvalues of the matrix ρ.

 doi:10.1016/B978-0-12-809283-5.00964-2

Multiqubit States and Entanglement

For two qubits, the possible pure states of the system are specified by four probability amplitudes: (Qubits are assumed to be distinguishable particles, and so the wavefunctions describing their states need not be symmetrized under exchange of particles.) $|\psi_2\rangle = \alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle$, with $|\alpha|^2 + |\beta|^2 + |\gamma|^2 + |\delta|^2 = 1$, and, e.g., $|00\rangle \equiv |0\rangle \otimes ,|0\rangle$ denoting the state in which the first and the second qubits are both in state $|0\rangle$. In general, 2^n amplitudes are needed to characterize a pure state of n qubits. Immediately we can see one advantage of quantum systems for information storage: if each distinct probability amplitude is regarded as a data register, the size of the memory grows exponentially with the number of qubits. Further, if one were to flip just one qubit, for example the second, the state would be transformed into $|\psi'_2\rangle = \beta|00\rangle + \alpha|01\rangle + \delta|10\rangle + \gamma|11\rangle$, where all of the amplitudes have been affected. This simple example demonstrates the notion of quantum parallelism, one of the most powerful properties of quantum information processors.

If the two-particle state can be written as a product of two single-particle states, i.e., if $\alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle = (A|0\rangle + B|1\rangle) \otimes ,(C|0\rangle + D|1\rangle)$, then these two particles can be considered as separate, unconnected entities: the state is said to be separable (When mixture is also present, the state is separable if it can be written as $\rho = \sum_i p_i \rho_i^A \otimes ,\rho_i^B$, i.e., as a sum of product states for systems A and B.). When the state cannot be written in this form, it is called an entangled state. Entanglement is one of the most fascinating fundamental properties of quantum systems: Erwin Schrödinger described it as '*the* characteristic trait of quantum mechanics … that enforces its entire departure from classical lines of thought.' Consider performing a measurement on the second of the two qubits, to establish which of its two states the particle was in, leaving the first qubit untouched. The result would collapse the state of the first qubit into one of two possibilities: either $|\phi_0\rangle = (\alpha|0\rangle + \gamma|1\rangle)/\sqrt{P_0}$ if the outcome of the measurement were 0 (which occurs with probability $P_0 = |\alpha|^2 + |\gamma|^2$) or $|\phi_1\rangle = (\beta|0\rangle + \delta|1\rangle)/\sqrt{P_1}$ if it were 1 (with probability $P_1 = |\beta|^2 + |\delta|^2$). Thus the state of the first qubit is instantaneously projected into a specific state by performing a measurement on the second qubit, which can be in a completely different location. This example demonstrates the nonlocality of quantum mechanics, which has been confirmed experimentally in some depth by a number of elegant quantum optics experiments over the last 30 years.

The fidelity $F = |\langle\Phi|\Psi\rangle|^2$ characterizes the overlap of two states (The generalized fidelity between mixed states ρ_A and ρ_B is $F = |Tr\sqrt{\sqrt{\rho A}\rho B\sqrt{\rho A}}|^2$.). The fidelity of $|\phi_0\rangle$ and $|\phi_1\rangle$ is a way to quantify the amount of entanglement in the initial two-qubit state. If the final two states are identical, the fidelity will be unity, and the initial state is separable; conversely if the two final states are orthogonal and occur with equal probability, then in a sense the initial state is maximally entangled, since the largest possible nonlocal influence would occur due to the measurement. Mathematically, the fidelity of the two final states is $|\langle\phi_0|\phi_1\rangle|^2 = 1 - C^2/4P_0P_1$, where $C = 2|\alpha\delta - \beta\gamma|$ is called the concurrence of the state. If $C = 0$, the state is separable; if $C = 1$, the state is maximally entangled (The concurrence can also be generalized to mixed states, although it has a much more complicated expression. The quantification of entanglement for mixed states with more than two subsystems is presently an active area of research.).

As discussed below, entanglement forms the heart of a number of quantum information protocols, such as dense coding, teleportation, and one type of quantum key distribution (also known as quantum cryptography), and large-scale entangled states of many qubits seem to be a requirement for the more ambitious goal of practical quantum computing.

Creating Entangled States Experimentally

Entangled states can currently be created in a controlled manner using technologies such as ion traps, cavity quantum electrodynamics, and optical spontaneous parametric down-conversion (SPDC). Down-conversion is a nonlinear optical process by which an incident 'pump' photon can be split (or down-converted) into a pair of longer-wavelength daughter photons (historically called 'signal' and 'idler') in a crystal possessing a $\chi^{(2)}$ nonlinearity, such as beta-barium borate (BBO). Mathematically, the process is described using the creation and annihilation operators of the field modes $(\hat{a}_\lambda^\dagger, \hat{a}_\lambda)$, thus:

$$|\Phi_{\text{out}}\rangle = \exp[-i\sum_{p,s,i} c_{p,s,i}(\hat{a}_p^\dagger \hat{a}_s \hat{a}_i + \hat{a}_i^\dagger \hat{a}_s^\dagger \hat{a}_p)]|\Phi_{\text{in}}\rangle$$

where the triple sum is over all field modes (subscripts p, s, and i refer to pump, signal, and idler modes, respectively) and $c_{p,s,i}$ is a coefficient linearly dependent on the second-order nonlinear susceptibility $\chi_{ijk}^{(2)}$ and also on the birefringent properties of the crystal. Further, $c_{p,s,i}$ will be negligibly small unless both total photon energy and momentum are conserved (i.e., $\omega_p = \omega_s + \omega_i$ and $\vec{\kappa}_p = \vec{\kappa}_s + \vec{\kappa}_i$, where $\vec{\kappa}$ is the intracrystal momentum).

One particularly efficient method for using this phenomenon to create entangled states is as follows. For a specific geometry (type-I phase matching), the daughter photons emerge from the crystal with identical polarizations (perpendicular to the parent polarization and the crystal optic axis) on opposite sides of a cone that is centered about the pump beam (**Fig. 1(a)**). Because each photon is in a definite state of polarization, the two photons are not in an entangled state. Two crystals, aligned with their axes of symmetry oriented at 90° to each other, as shown in **Fig. 1(b)**, can then be used to create an entangled state (Another widely used method employs type-II phase matching in a single crystal. The down-conversion photons are emitted with perpendicular polarizations along a pair of cones. Pairs emitted along particular directions are in the maximally polarization-entangled state $|\psi_-\rangle = \frac{1}{\sqrt{2}(|HV\rangle - |HV\rangle}$). Such a source was used in the first demonstrations of superdense coding and quantum teleportation.). With crossed crystals, two processes are possible: the parent photon can down-convert in the first crystal to yield two vertically polarized photons, or it can down-convert in the second crystal to yield two horizontally polarized photons. Because it is impossible to

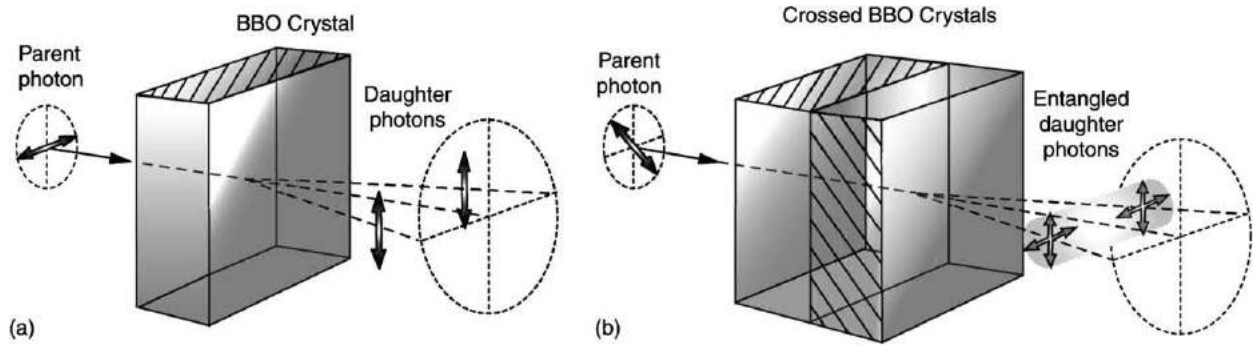


Fig. 1 An entangled-photon source. (a) For a given orientation of the nonlinear crystal, a horizontally polarized parent photon produces a pair of vertically polarized daughters. The daughters emerge on opposite sides of an imaginary cone. The cone's axis is parallel to the original direction taken by the parent photon. The two daughter photons are not in an entangled state. Reorienting the BBO crystal by 90° will produce a pair of horizontally polarized daughters if a vertically polarized pump beam is used. (b) Passing a photon polarized at 45° through two crossed BBO crystals can produce two photons in an entangled state. Because of the Heisenberg uncertainty principle, there is no way to tell in which crystal the parent photon 'gave birth,' and so a coherent superposition of two possible outcomes results: a pair of vertically polarized photons or a pair of horizontally polarized photons. Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.

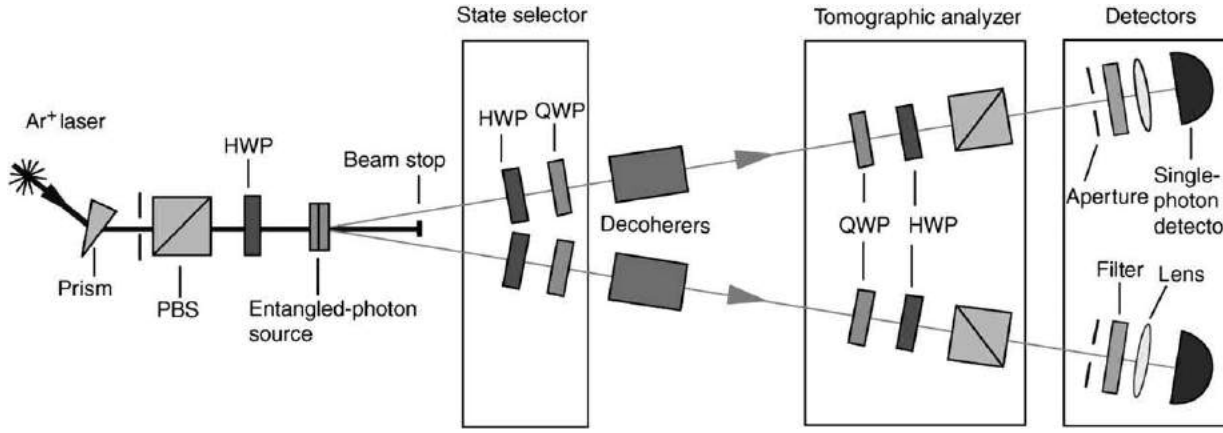


Fig. 2 Creating and measuring two-photon entangled states. The 'pump' photons are created, for example, in an argon ion laser and are linearly polarized with a polarizing beamsplitter (PBS). The half-wave plate (HWP) rotates the polarization state before the photon enters the pair of nonlinear crystals that constitute the entangled-photon source; the initial angle of pump polarization controls the entanglement of the pair produced. Each photon's polarization state can be altered at will by the subsequent HWP and quarter-wave plate (QWP). The decoherers following the state selection allow the production of mixed photon states. The optical elements (QWP, HWP, and PBS) in the tomographic analyzer allow the measurement of each photon in an arbitrary basis. Coincidence measurements of photons allow the quantum state to be determined. Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.

distinguish which of these processes has occurred, the state of the daughter photons is a coherent quantum-mechanical superposition of the states that would arise from each crystal alone; the output of the crossed crystals is photons in the maximally entangled state $|\Phi_+\rangle \equiv \frac{1}{\sqrt{2}}(|\mathrm{HH}\rangle + |\mathrm{VV}\rangle)$. (Note that in an arbitrary basis $|\theta\rangle \equiv \cos\theta|H\rangle + \sin\theta|V\rangle$ and $|\theta^{\perp}\rangle \equiv -\sin\theta|H\rangle + \cos\theta|V\rangle$, this maximally entangled state has the form $|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|\theta\theta\rangle + |\theta^{\perp}\theta^{\perp}\rangle)$, demonstrating that the nonlocal correlations are present regardless of the bases used to represent the state).

Fig. 2 shows how this basic source can be adapted to produce any pure quantum state of two photons by placing rotatable half- and quarter-wave plates (which can be used to transform the polarization state of a single photon) before the crystal pair and in the paths of the two daughter photons. To create mixed states, a long birefringent crystal can be used to delay one polarization component with respect to the other. If the delay is longer than the coherence time of the photons, the horizontal and vertical components are effectively decohered; that is, the phase relationship between the different states is destroyed.

The figure also shows schematically the apparatus required for measuring the quantum state. In classical optics, four Stokes parameters are required to specify the polarization of a single beam (i.e., an ensemble of uncorrelated photons); for a pair of photons, 16 projective measurements, each with different wave plate settings, is required. From these 16 measurements, all of the elements of the 4×4 density matrix describing the (in general, mixed) state of the photon pairs can be deduced. This is an example of quantum state tomography, a technique that has found application to a number of quantum optical systems.

Quantum Key Distribution

Two parties, historically known as Alice and Bob, want to have a secret conversation. A generic, classical encryption protocol would begin when Alice and Bob convert their messages to separate binary streams of 0's and 1's, which are then encrypted and decrypted with a set of secret 'keys' known only to the two. Each key is a random string of 0's and 1's that is as long as the binary string comprising each message. To encrypt, Alice (the sender) sequentially adds each bit of the key to each bit of her message, using addition modulo 2. She then sends the encrypted message over a public channel to Bob, who decrypts it by simply repeating the addition modulo 2 of the key to the message. This type of encryption, known as a one-time pad, is currently the only provably secure encryption protocol. But the one-time pad is effective only if Alice and Bob never reuse the key, and more obviously, if the key remains secret. A potential eavesdropper, Eve, cannot be allowed to glean any part of the bit stream that makes up the key. Therein lies a central problem of cryptography – how can secret keys be created and then securely distributed?

Quantum key distribution (QKD) exploits the fundamentally indivisible nature of photons to perform this task. There are a variety of QKD protocols; here we describe one that employs entangled photon pairs (see **Fig. 3**). Alice and Bob use a source such as the one described above to produce maximally entangled photons in the state $|\Phi_+\rangle \equiv \frac{1}{\sqrt{2}}(|\mathrm{HH}\rangle + |\mathrm{VV}\rangle) = \frac{1}{\sqrt{2}}(|45^\circ, 45^\circ\rangle + |-45^\circ, -45^\circ\rangle)$. One photon goes to Alice and the other to Bob. For each pair, Alice and Bob randomly and independently analyze their respective photons (using a polarizing beamsplitter) in the H/V or the $45^\circ/-45^\circ$ basis. They record a bit value of 0 for all H or 45° results, and a 1 for all V or -45° results. After a sufficient number of measurements (dictated by the length of the key), Alice and Bob have a public discussion, e.g., over the internet. For each detected photon, they announce which basis they used for the measurement, but not the actual measurement result. Whenever they made the same basis choice (50% of the time), the correlations of the entangled state ensure their measured bit values agree. By contrast, they discard the results when they used different bases, because their measurements are completely uncorrelated (see **Table 1**).

An eavesdropper (Eve) cannot tap the line, as she might with conventional communications, because of the indivisibility of individual photons and the fact that arbitrary quantum systems cannot be accurately cloned (The no-cloning proof is very simple. If we have a copying operation such that $|0\rangle|c\rangle \rightarrow |0\rangle|0\rangle$ and $|1\rangle|c\rangle \rightarrow |1\rangle|1\rangle$ ($|c\rangle$ is the initial state of the copier), then by linearity, a superposition becomes an entangled state, instead of two copies: $\frac{(|0\rangle+|1\rangle)}{\sqrt{2}}|c\rangle \rightarrow \frac{(|0\rangle|0\rangle+|1\rangle|1\rangle)}{\sqrt{2}} \neq \frac{(|0\rangle+|1\rangle)}{\sqrt{2}}\frac{(|0\rangle+|1\rangle)}{\sqrt{2}}$). If Eve steals Bob's photon (a 'denial-of-service' attack), the photon's information never becomes part of the key. Thus, although a wiretap would reduce the rate of the transmission, it would not jeopardize the security of the key. Eve can try to intercept the photon, measure it, and send another one to Bob. But any measurement Eve would make to determine the photon's polarization state would perturb the photon and collapse the entangled state. The photon she sends to Bob would therefore only be 'classically' correlated with Alice's photon. Consequently, Eve's intervention necessarily induces additional errors into Bob's key, which Alice and Bob can detect by

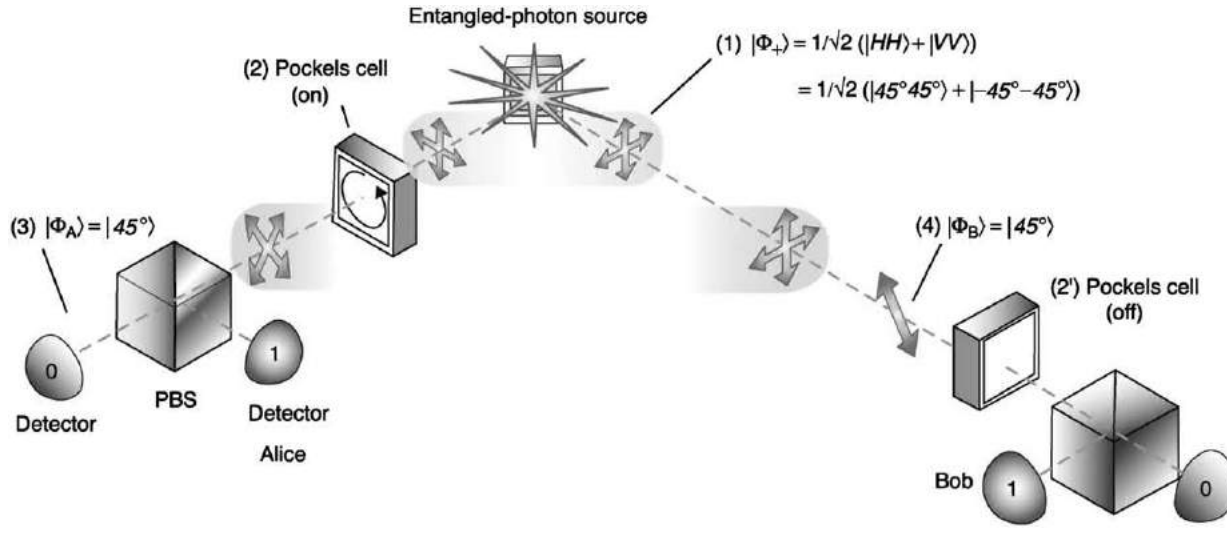


Fig. 3 Quantum cryptography using entangled photons. Entangled photons can be used to create a pair of identical cryptographic keys. One member of an entangled pair (1) is sent to Alice, and the other to Bob. Alice and Bob each randomly and independently analyze their respective photon in one of two linear polarization bases with a polarizing beamsplitter (PBS); the basis can be actively chosen using a Pockels cell before the PBS, as shown, to rotate the polarization (alternatively, the 'choice' could be made by directing the photon onto a nonpolarizing 50/50 beamsplitter; in the transmitted path, one analysis basis is used, in the reflected path, the other is used). In the example shown, Alice used the $45^\circ/-45^\circ$ basis (2), and measured 45° polarization (3), thus projecting Bob's photon into the identical state (4). Since he chose the H/V analysis basis (2′), he is equally likely to detect a 0 or a 1. By subsequent public discussions Alice and Bob determine the events for which they used the same analysis basis (and discard the other events). For these events Alice and Bob will have obtained identical measurement results, which they may interpret as raw key material. After classical error correction and privacy amplification techniques are applied, the remaining string is the sought-after shared secret key. Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.

Table 1 Polarization entanglement-based quantum cryptography protocol

The analysis basis	H/V	H/V	45°/ − 45°	H/V	45°/ − 45°	45°/ − 45°	H/V	45°/ − 45°
Alice's measurement result[a]	*H*	–	*− 45°*	*V*	*45°*	*45°*	*V*	*− 45°*
Alice's bit value	0	–	1	1	0	0	1	1
Bob's analysis basis	H/V	45°/ − 45°	45°/ − 45°	45°/ − 45°	H/V	45°/ − 45°	H/V	HV
Bob's measurement result	*H*	*− 45°*	*− 45°*	*45°*	*H*	–	*V*	*H*
Bob's bit value	0	1	1	0	0	–	1	0
Public discussion: Both photons detected and same basis used?	yes	no	yes	no	no	no	yes	no
Remaining secret key	0		1				1	

[a]In some cases Alice or Bob may not detect a photon due to loss or detector inefficiency. These events simply do not contribute to the key data.

Table 2 Method to convert the initial state $|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$ into any of the four Bell states

Alice's operation	*Polarization transformation*	*Resultant two-qubitstate*
$\hat{I}$, the identity	H→H; V→V	$\|\Phi_+\rangle = \frac{1}{\sqrt{2}}(\|00\rangle + \|11\rangle)$
$\hat{\sigma_x}$	H→V; V→H	$\|\Psi_+\rangle = \frac{1}{\sqrt{2}}(\|01\rangle + \|10\rangle)$
$i\hat{\sigma_y}$	H→V; V→ − H	$\|\Psi_-\rangle = \frac{1}{\sqrt{2}}(\|01\rangle - \|10\rangle)$
$\hat{\sigma_z}$	H→H; V→ − V	$\|\Phi_-\rangle = \frac{1}{\sqrt{2}}(\|00\rangle - \|11\rangle)$

publicly revealing a small subset of their actual key. Unfortunately, even with no eavesdropper, the encryption keys created by any real-world quantum cryptography system typically possess a few-percent errors. To make sure their key is secure, Alice and Bob ascribe all errors to Eve and then estimate the maximum amount of information available to the eavesdropper. They then use a classical privacy amplification protocol to reduce Eve's knowledge of the secret key to less than one bit by reducing the length of the key. It has been proven that if the initial error probability per bit is greater than $\sim 15\%$, no secret bits will remain after error detection and privacy amplification.

Researchers are working to make entanglement-based quantum cryptography more practical. Also, a number of longer distance quantum cryptography demonstrations using weak pulses have been carried out, over tens of kilometers in fibers and in free space. In fact, the first commercially available systems have recently been announced (in Europe and the United States).

Quantum Superdense Coding

It is possible for Alice to send Bob two bits of classical information using a single qubit in the quantum superdense coding protocol. Suppose that Alice and Bob share one qubit each of an entangled pair in the maximally entangled state $|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, where we have returned to the generic labeling scheme for the qubit states. Alice has two classical bits of information, which is equivalent to one of four choices. She can encode this information by applying one of four possible transformations on her qubit; a suitable set of transformations are the three Pauli matrices and the identity (i.e., do nothing). This set of operations can be performed experimentally on photon qubits quite easily, e.g., using waveplates, as indicated in **Table 2**. The four resultant two-qubit states are all maximally entangled and are orthonormal. They form a special basis for the two-qubit states, the Bell basis, which is of particular use in analyzing entanglement.

Alice now sends her qubit to Bob, e.g., through an optical fiber. Bob can then perform a Bell state analysis (Complete discrimination of all four Bell states is presently an unsolved technical problem: for photons one needs either a nonlinear interaction (which is typically very weak) or to exploit so-called 'hyper-entanglement' involving other entangled degrees of freedom of the photons.). on his qubit pair, i.e., a projective measurement of the two-qubit state in the Bell basis. The result immediately reveals the choice of operation Alice made, and in effect, two bits of classical information have been encoded on a single qubit.

Quantum Teleportation

Another application of entanglement is quantum state teleportation (**Fig. 4**), in which the infinite amount of information contained in an arbitrary qubit state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ is transferred by communication of two bits of classical information between Alice and Bob, if they also share two qubits in the maximally entangled state $|\Psi_-\rangle$. The three-photon initial state (i.e., the input photon plus the two entangled photons) is

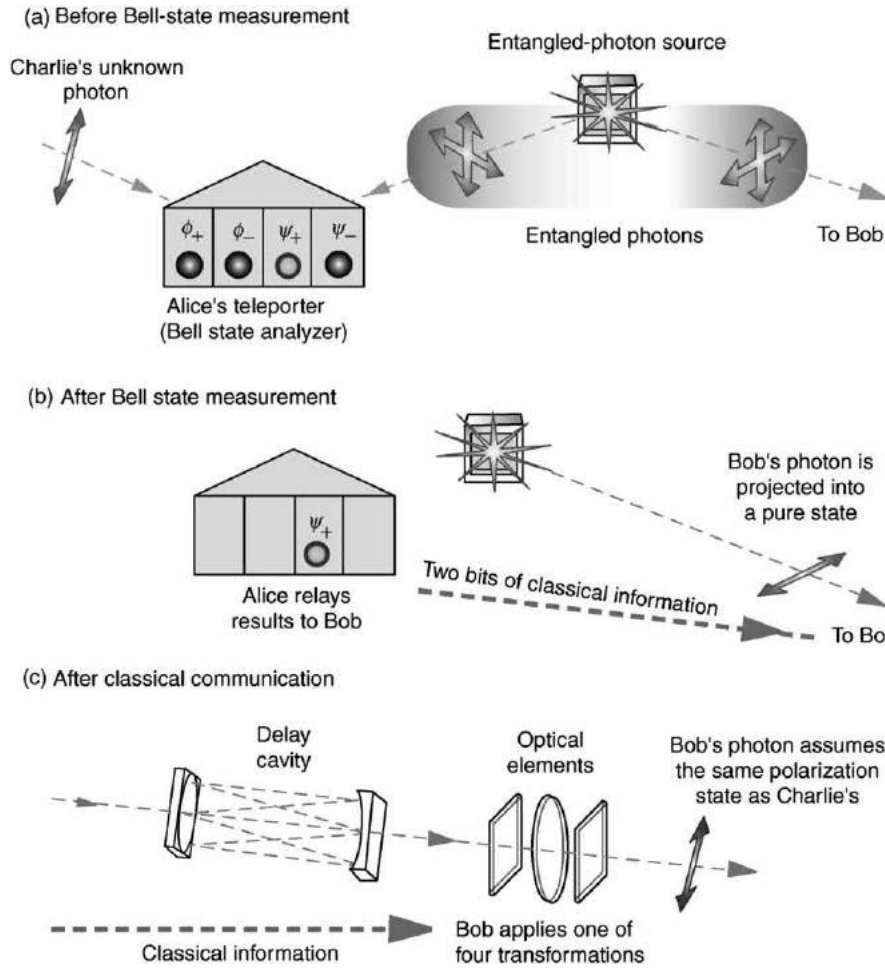


Fig. 4 Quantum state teleportation. (a) Teleportation requires an entangled photon source, a Bell state analyzer and a classical communications channel. One entangled photon goes to Bob and the other to Alice, who also receives a photon of unknown polarization state. (b) Alice performs a joint polarization measurement of the two photons and relays the result to Bob using two classical bits of information. The photon going to Bob is projected into a pure state as a result of Alice's measurement. (c) Upon receiving Alice's classical information, Bob performs a simple transformation on his photon (which he has been storing), such as a rotation of the polarization vector. The resulting state of his photon is then identical to the unknown state Alice wished to teleport. Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.

$$|\psi_0\rangle = (\alpha|H\rangle + \beta|V\rangle) \otimes , \frac{1}{\sqrt{2}(|HV\rangle - |VH\rangle}),$$

which can be rewritten with the first two photons (the input plus the first half of the entangled pair) represented by the Bell state basis:

$$\left|\psi_0\rangle = \frac{1}{2}\{|\Phi_+\rangle \otimes , (-\beta|H\rangle + \alpha|V\rangle\right) + |\Phi_-\rangle \otimes , (\beta|H\rangle + \alpha|V\rangle) + |\Psi_+\rangle \otimes , (-\alpha|H\rangle + \beta|V\rangle) + |\Psi_-\rangle \otimes , (-\alpha|H\rangle - \beta|V\rangle)\}.$$

Now suppose that a Bell state measurement is performed on the first two photons. The third photon is immediately projected into one of four possible states, which can be transformed back into the state of the original input photon by a simple operation, e.g., with a waveplate. For example, if the Bell state measurement produced the result $|\Psi_+\rangle$, the third photon is immediately 'collapsed' into the pure state $|\psi_3\rangle = -\alpha|H\rangle + \beta|V\rangle$. By applying a π-phase shift to the horizontal polarization (relative to the vertical), $|\psi_3\rangle$ can be transformed into the original input state: $|\psi_3'\rangle = \alpha|H\rangle + \beta|V\rangle$. Alice (who does the Bell state measurement) need only communicate to Bob (who wants to receive the teleported photon) which of the four Bell states she measured. Note that during the entire teleportation procedure, neither Alice nor Bob can obtain any idea of the values of the parameters α and β, which specify the state. Also, because Alice's measurement collapses the unknown state, there is only a single copy at the end of the protocol.

Quantum state teleportation was first demonstrated experimentally by Zeilinger and coworkers (University of Innsbruck, Austria). The group was able to determine two of the four Bell states unambiguously (the other two states gave the same experimental signature) and proved for those cases that the state of the input photon could indeed be transferred to Bob. Other

experiments have realized modified forms of quantum teleportation in different systems. For example, Kimble's group (California Institute of Technology) teleported the coherent state of an optical mode using squeezed (rather than polarization-entangled) light. Researchers have recently suggested how teleportation might form the basis of a distributed network of quantum communication channels, and how it might enable quantum computing in all-optical systems.

Other Application of Optical Entanglement

Quantum Computing

The most challenging and powerful application of quantum information is large-scale quantum computing. Two features that make quantum information processing potentially powerful are the exponentially large Hilbert space, which gives quantum registers very large capacities, and quantum parallelism, which means that data processing tasks can be performed very efficiently. One fundamental drawback, which severely constrains useful applications of quantum computers, is that the final measurement can only produce a number of classical bits equal to the number of qubits. Thus, quantum computers are limited to performing tasks in which a small amount of information is meant to be gleaned from processing a large amount of data; examples include searching an unstructured database for a specific entry (Grover's algorithm) or finding the periodicity of a function (the quantum Fourier transform). This second task is central to Shor's factor-finding algorithm, the most famous quantum computing algorithm to date.

A practicable quantum computer technology must have at least the following features, first identified by DiVincenzo:

- A set of well-characterized, distinguishable qubits to form a quantum data storage register.
- The ability to initialize the qubits of the register in a simple fiducial state.
- Decoherence times that are much longer than the time needed to perform logical operations.
- The ability to perform any single qubit operation on any qubit in the register, and the ability to perform two-qubit conditional logic gates (such as the CNOT gate: $\alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle \rightarrow \alpha|00\rangle + \beta|01\rangle + \gamma|11\rangle + \delta|10\rangle$). Together, these operations constitute a universal set of quantum gates, from which all other gates can be synthesized.
- The ability to measure each qubit.

All-optical schemes have been used to implement small quantum algorithms. However, most of the approaches are not scalable, due to the limitations of non-linear optics to perform a CNOT gate at the single-photon level. Recent proposals have suggested that very high efficiency single-photon detectors, along with sources of single photons 'on-demand,' allow scalable quantum computing with only linear optics; preliminary two-qubit gates have been experimentally demonstrated. A number of other promising candidate technologies that meet DiVincenzo's five requirements are being pursued vigorously. For example, the ability to create multiqubit entanglement and perform reliable measurements on trapped ions cooled and manipulated by lasers has recently been demonstrated. Solid state systems offer the possibility of scalability, and a number of schemes, such as quantum dots, isolated impurities with nuclear spins, and superconducting quantum interference devices (SQUIDs), are being investigated. It is possible that the final quantum computing technology may take the form of a hybrid between various present approaches.

Lithography

Lithography, in which a pattern is optically imaged onto some photoresistive material, is the primary method of manufacturing microscale or nanoscale electronic devices. An inherent limitation of this process is that details smaller than a wavelength of light cannot be written reliably. However, quantum state entanglement might circumvent this limitation. Under the right circumstances, the interference pattern formed by beams of entangled photons can have half the classical fringe spacing. Quantum lithography requires two beams of photons, a coherent superposition consisting of the state in which two photons are in beam A while none are in B, and the state in which no photon is in beam A while two photons are in B. Such number-entangled states can be made in the laboratory, and the predictions about fringe spacings have been verified. However, other obstacles must be overcome to surpass current classical-lithography techniques.

Two-Photon Imaging and Microscopy

At present, two-photon microscopy is widely used to produce high-resolution images, often of biological systems. However, the classical light sources (lasers) used for the imaging have random spreads in the temporal and spatial distributions of the photons, and the light intensity must be very high if two photons are to intersect within a small enough volume to cause a detectable excitation. Such high intensity can damage the system under investigation. Because the temporal and spatial correlations may be much stronger between members of an entangled photon pair, much weaker light sources could be used, which would be much less damaging to the systems being observed. The development of such systems is currently an active area of research.

Further Reading

Berman, G.P., Doolen, G.D., Mainieri, R., Tsifrinovitch, V., 1998. Introduction to Quantum Computers. Singapore: World Scientific.
Bouwmeester, D., Ekert, A.K., Zeilinger, A. (Eds.), 2000. The Physics of Quantum Information. Berlin: Springer-Verlag.
Braunstein, S.L., Lo, H.K. (Eds.), 2001. Scalable Quantum Computers: Paving the Way to Realization. Berlin: Wiley-VCH.
DiVincenzo, D.P., 2000. The physical implementation of quantum computation. Fortschrcitte der Physik 48, 771–783.
Gisin, N., Ribordy, G., Tittel, W., Zbinden, H., 2002. Quantum cryptography. Review of Modern Physics 74, 145–195.
Kok, P., Braunstein, S.L., Dowling, J.P., 2002. Optics and Photonics News 13 (9), 24–27.
Lo, H.K., Popescu, S., Spiller, T. (Eds.), 1998. Introduction to Quantum Computation and Information. Singapore: World Scientific.
Macchiavello, C., Palma, G.M., Zeilinger, A. (Eds.), 2000. Quantum Computation and Quantum Information Theory. Singapore: World Scientific.
Mandel, L., Wolf, E., 1995. Optical Coherence and Quantum Optics. Cambridge, UK: Cambridge University Press.
Nielsen, M.A., Chuang, I.L., 2000. Quantum Computation and Quantum Information. Cambridge, UK: Cambridge University Press.
Williams, C.P., Clearwater, S.H., 1998. Explorations in Quantum Computing. Santa Clara: Telos.
Yariv, A., 1989. Quantum Electronics. New York: Wiley.

Applications in Semiconductors

HM van Driel and JE Sipe, University of Toronto, Toronto, ON, Canada

Introduction

Interference phenomena are well-known in classical optics. In the early 19th century, Thomas Young showed conclusively that light has wave properties, with the first interference experiment ever performed. He passed quasi-monochromatic light from a single source through a pair of double slits using the configuration of **Fig. 1(a)** and observed an interference pattern consisting of a series of bright and dark fringes on a distant screen. The intensity distribution can be explained only if it is assumed that light has wave or phase properties. In modern terminology, if E_1 and E_2 are the complex electric fields of the two light beams arriving at the screen, by the superposition principle of field addition the intensity at a particular point on the screen can be written as

$$I \propto |E_1 + E_2|^2 = |E_1|^2 + |E_2|^2 + |E_1||E_2| \cos(\phi_1 - \phi_2) \tag{1}$$

where $\phi_1 - \phi_2$ is the phase difference of the beams arriving at the screen. While this simple experiment was used originally to demonstrate the wave properties of light, it has subsequently been used for many other purposes, such as measuring the wavelength of light. However, for our purposes, here the Young's double slit apparatus can also be viewed as a device that redistributes light, or controls its intensity at a particular location. For example, let's consider one of the slits to be a source, with the other slit taken to be a gate, with both capable of letting through the same amount of light. If the gate is closed, the distant screen is nearly uniformly illuminated. But if the gate is open, at a particular point on the distant screen there might be zero intensity or as much as four times the intensity emerging from one slit because of interference effects, with all the light merely being spatially redistributed. In this sense the double slit system can be viewed as a device to redistribute the incident light intensity. The key to all this is the superposition principle and the properties of optical phase.

The superposition principle and interference effects also lie at the heart of quantum mechanics. For example, let's now consider a system under the influence of a perturbation with an associated Hamiltonian that has phase properties. In general the system can

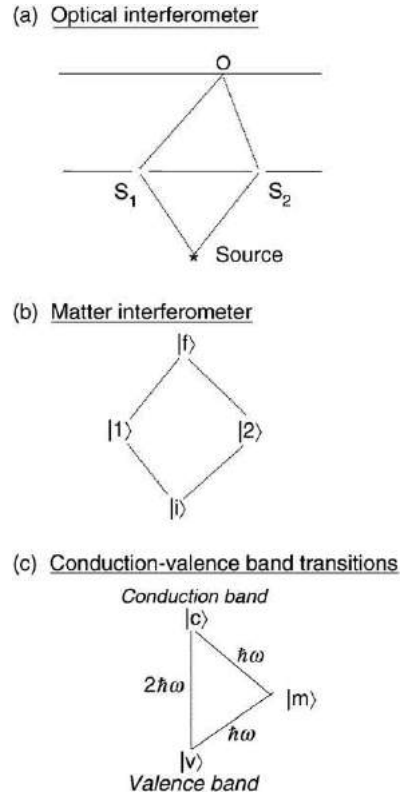


Fig. 1 (a) Interference effects in the Young's double slit experiment; (b) illustration of the general concept of coherence control via multiple quantum mechanical pathways; (c) interference of single- and two-photon transitions connecting the same valence and conduction band states in a semiconductor.

 doi:10.1016/B978-0-12-809283-5.00904-6

evolve from one quantum state $|i\rangle$ to another $|f\rangle$ via multiple pathways involving intermediate states $|m\rangle$. Because of the phased perturbation, interference between those pathways can influence the system's final state. If a_m is the (complex) amplitude associated with a transition from the initial to the final state, via intermediate virtual state $|m\rangle$, then for all possible intermediate states the overall transition probability, W, can be written as

$$W = \left| \sum_m a_m \right|^2 \tag{2}$$

In the case of only two pathways, as illustrated in **Fig. 1(b)**, W becomes

$$\begin{aligned} W &= |a_1 + a_2|^2 \\ &= |a_1|^2 + |a_2|^2 + |a_1||a_2| \cos(\phi_1 - \phi_2) \end{aligned} \tag{3}$$

where ϕ_1 and ϕ_2 are now the phases of the two transition amplitudes. While this expression strongly resembles that of **Eq. (1)**, here ϕ_1 and ϕ_2 are influenced by the phase properties of the perturbation Hamiltonian that governs the transition process; these phase factors arise, for example, if light fields constitute the perturbation. But overall it is clear that the change in the state of the system is affected by both the amplitude and phase properties of the perturbation. Analogous with the classical interferometer, equality of the transition amplitudes leads to maximum contrast in the transition rate. Although it has been known since the early days of quantum mechanics that the phase of a perturbation can influence the evolution of a system, generally phase has only been discussed in a passive role. Only in recent years has phase been used as a control parameter for a system, on the same level as the strength of the perturbation itself.

Coherence Control

Coherence control, or quantum control as it is sometimes called, refers to the active process through which one can use phase-dependent perturbations originating with, for example, coherent light waves, to control one or more properties of a quantum system, such as state population, momentum, or spin. This more general perspective of interference leads to a picture of interference of matter waves, rather than simply light waves, and one can now speak of an effective 'matter interferometer'. Laser beams, in particular, are in a unique position to play a role in such processes, since they offer a macroscopic 'phase handle' with which to create such effects. This was recognized by the community of atomic and molecular scientists who emphasized the active manifestations of quantum interference effects, and proposed that branching ratios in photochemical reactions might be controlled through laser-induced interference processes. The use of phase as a control parameter also represents a novel horizon of applications for the laser since most previous applications had involved only amplitude (intensity).

Of course, just as the 'visibility' of the screen pattern for a classical Young's double slit interferometer is governed by the coherence properties of the two slit source, the evolution of a quantum system reflects the coherence properties of the perturbations, and the degree to which one can exercise phase control via external perturbations is also influenced by the interaction of the system of interest with a reservoir – essentially any other degrees of freedom in the system – which can cause decoherence and reduce the effectiveness of the 'matter interferometer'. Since the influence of a reservoir will generally increase with the complexity of a system, one might think that coherence control can only be effectively achieved in simple atomic and molecular systems. Indeed, some of the earliest suggestions for coherence control of a system involved using interference between single and three photon absorption processes, connecting the same initial and final states with photons of frequency 3ω and ω, respectively, and controlling the population of excited states in atomic or diatomic systems. However, it has been difficult to extend this 'two color' paradigm to more complex molecular systems. Since the reservoir leads to decoherence of the system overall, shorter and shorter pulses must be used to overcome decoherence effects. However, short pulses possess a large bandwidth with the result that, for complex systems, selectivity of the final state can be lost. One must therefore consider using interference between multiple pathways in order to control the system, as dictated by the details of the unperturbed Hamiltonian of the system.

For a polyatomic molecule or a solid, the complete Hamiltonian is virtually impossible to determine exactly and it is therefore equally difficult to prescribe the optimal spectral (amplitude and phase) content of the pulses that should be used for control purposes. An alternative approach has therefore emerged, in which one foregoes knowledge of the eigenstates of the Hamiltonian and details of the different possible interfering transition paths in order to achieve control of the end state of a system. This branch of coherence control has come to be known as optimal control. In optimal control one employs an optical source for which one can (ideally) have complete control over the spectral and phase properties. One then uses a feedback process in which experiments are carried out, the effectiveness of achieving a certain result is determined, and the pulse characteristics are then altered to obtain a new result. A key component is the use of an algorithm to select the new pulse properties as part of the feedback system. In this approach the molecule teaches the external control system what it 'requires' for a certain result to be optimally achieved. While the 'best' pulse properties may not directly reveal details of the multiple interference process required to achieve the optimal result, this technique can nonetheless be used to gain some insight into the properties of the unperturbed Hamiltonian and, regardless of such understanding, achieve a desirable result. It has been used to control chemical reaction rates involving several polyatomic molecules, with considerable enhancement in achieving a certain product relative to what can be done using simple thermodynamics.

Coherence Control in Semiconductors

Since coherence control of chemical reactions involving large molecules has represented a significant challenge, it was generally felt that ultrafast decoherence processes would also make control in solids, in general, and semiconductors, in particular, difficult. Early efforts therefore focused on atomic-like situations in which the electrons are bound and associated with discrete states and long coherence times. In semiconductors, the obvious choice is the excitonic system, defect states, or discrete states offered by quantum wells. Population control of excitons and directional ionization of electrons from quantum wells has been clearly demonstrated, similar to related population control of and directional ionization from atoms. Other manifestations of coherence control of semiconductors include control of electron–phonon interactions and intersub-band transitions in quantum wells.

Among the different types of coherence control in semiconductors is the remarkable result that it is possible to coherently control the properties of free electrons associated with continuum states. Although optimal control might be used for some of these processes, we have achieved clear illustrations of control phenomena based on the use of harmonically related beams and interference of single- and two-photon absorption processes connecting the continuum valence and conduction band states as generically illustrated in Fig. 1(c). Details of the various processes observed can be found elsewhere. Of course, the momentum relaxation time of electrons or holes in continuum states is typically of the order of 100 fs at room temperature, but this time is sufficiently long that it can permit phase-controlled processes. Indeed, with respect to conventional carrier transport, this 'long time' lapse is responsible for the typically high electrical mobilities in crystalline semiconductors such as Si and GaAs. In essence, a crystalline semiconductor with translational symmetry has crystal momentum as a good quantum number, and selection rules for scattering prevent the momentum relaxation time from being prohibitively small. Since electrons or holes in any continuum state can participate in such control processes, one need not be concerned about pulsewidth or bandwidth, unless the pulses were to be so short that carriers of both positive and negative effective mass were generated within one band.

Coherent Control of Electrical Current Using Two Color Beams

We now illustrate the basic principles that describe how the interference process involving valence and conduction band states can be used to control properties of a bulk semiconductor. We do so in the case of using one or two beams to generate and control carrier population, electrical current, and spin current. In pointing out the underlying ideas behind coherence control of semiconductors, we will skip or suppress many of the mathematical details which are required for a full understanding but which might obscure the essential physics. In particular we consider how phase-related optical beams with frequencies ω and 2ω interact with a direct gap semiconductor such that $\hbar\omega < E_g < 2\hbar\omega$ where E_g is the electronic bandgap. For simplicity we consider exciting electrons from a single valence band via single-photon absorption at 2ω and two-photon absorption at ω. As is well known, within an independent particle approximation, the states of electrons and holes in semiconductors can be labeled by their vector crystal momentum $\mathbf{k}$; the energy of states near the conduction or valence bandedges varies quadratically with $|\mathbf{k}|$. For one-photon absorption the transition amplitude can be derived using a perturbation Hamiltonian of the form $H = e/mc\mathbf{A}^{2\omega}\cdot\mathbf{p}$ where $\mathbf{A}^{2\omega}$ is the vector potential associated with the light field and $\mathbf{p}$ is the momentum operator. The transition amplitude is therefore of the form

$$\mathrm{a}^{2\omega} \propto E^{2\omega} e^{i\phi_{2\omega}} p_{\mathrm{cv}} \tag{4}$$

where p_{cv} is the interband matrix element of $\mathbf{p}$ along the field ($E^{2\omega}$) direction; for illustration purposes the field is taken to be linearly polarized. The overall transition rate between two particular states of the same $\mathbf{k}$ can be expressed as $W_1 \propto \mathrm{a}^{2\omega}(\mathrm{a}^{2\omega})^* \propto I^{2\omega}|p_{cv}|^2$ where $I^{2\omega}$ is the intensity of the beam. This rate is independent of the phase of the light beam as well as the sign of $\mathbf{k}$. Hence the absorption of light via single-photon transitions generally populates states of equal and opposite momentum with equal probability or, equivalently, establishes a standing electron wave with zero crystal momentum. This is not surprising since photons possess very little momentum and, in the approximation of a uniform electric field, do not give any momentum to an excited electron. The particular states that are excited depends on the light polarization and crystal orientation. However, the main point is that while single-photon absorption can lead to anisotropic filling of electron states, the distribution in momentum space is not polar (dependent on sign of $\mathbf{k}$). Similar considerations apply to the excited holes, but to avoid repetition we will focus on the electrons only.

For two-photon absorption involving the ω photons and connecting states similar to those connected with single-photon absorption, one must employ the 2nd order perturbation theory using the Hamiltonian $H = e/mc\mathbf{A}^{\omega}\cdot\mathbf{p}$. For two-photon absorption there is a transition from the valence band to an (energy nonallowed) intermediate state followed by a transition from the intermediate state to the final state. To determine the total transition rate one must sum over all possible intermediate states. For semiconductors, by far the dominant intermediate state is the final state itself, so that the associated transition amplitude has the form

$$\mathrm{a}^{\omega} \propto (E^{\omega} e^{i\phi_{\omega}} p_{\mathrm{cv}})(E^{2\omega} e^{i\phi_{\omega}} p_{\mathrm{cc}}) \propto (E^{\omega})^2 e^{2i\phi_{\omega}} p_{\mathrm{cv}} \hbar\mathrm{k} \tag{5}$$

where the matrix element p_{cc} is simply the momentum of the conduction band state ($\hbar k$) along the direction of the field. Note that unlike an atomic system, p_{cc} is nonzero, since Bloch states in a crystalline solid do not possess inversion symmetry. If two-photon absorption acts alone, the overall transition rate between two particular $\mathbf{k}$ states would be $W_2 \propto (I^{\omega})^2|p_{cv}|^2k^2$. As with single-photon absorption, because this transition rate is independent of the sign of k, two-photon absorption leads to production of electrons with no net momentum.

When both single- and two-photon transitions are present simultaneously, the transition amplitude is the sum of the transition amplitudes expressed in Eq. (3). The overall transition rate is then found using Eq. (1) and yields $W = W_1 + W_2 + Int$ where the interference term *Int* is given by

$$\mathrm{Int} \propto E^{2\omega} E^{\omega} E^{\omega} \sin(\phi_{2\omega} - 2\phi_{\omega}) k \tag{6}$$

Note, however, that the interference effect depends on the sign of k and hence can be constructive for one part of the conduction band but destructive for other parts of that band, depending on the value of the relative phase, $\Delta\phi = \phi_{2\omega} - 2\phi_{\omega}$. In principle one can largely eliminate transitions with $+\mathbf{k}$ and enhance those with $-\mathbf{k}$. This effectively generates a net momentum for electrons or holes and hence, at least temporarily, leads to an electrical current in the absence of any external bias. The net momentum of the carriers, which is absent in the individual processes, must come from the lattice. Because the carriers are created with net momentum during the pulses one has a form of current injection that can be written as

$$dJ/dt \propto E^{2\omega} E^{\omega} E^{\omega} \sin(\Delta\phi) \tag{7}$$

where J is the current density. This type of current injection is allowed in both centrosymmetric and noncentrosymmetric materials. The physics and concepts behind the quantum interference leading to this form of current injection are analogous with the Young's double slit experiment. Note as well that if this 'interferometer' is balanced (single- and two-photon transition rates are similar), then it is possible to control the overall transition rate with great contrast. Roughly speaking, one balances the two arms of the 'effective interferometer' involved in the interference process, one 'arm' corresponding to the one-photon process and the other to the two-photon process. As an example, let's consider the excitation of GaAs, which has a room-temperature bandgap of 1.42 eV (equivalent to 870 nm). For excitation of GaAs using 1550 and 775 nm light under balanced conditions, the electron cloud is injected with a speed close to 500 km s^{-1}. Under 'balanced' conditions, nearly all the electrons are moving in the same direction. Assuming that the irradiance of the 1550 nm beam is 100 MW cm^{-2}, while that of the second harmonic beam is only 15 kW cm^{-2} (satisfying the 'balance' condition), with Gaussian pulse widths of 100 fs, one obtains a surprisingly large peak current of about 1 kA cm^{-2} for a carrier density of only 10^{14} cm^{-3} if scattering effects are ignored. When scattering is taken into account, the value of the peak current is reduced and the transient current decays on a time-scale of the momentum relaxation time. Fig. 2(a) shows an experimental setup which can be used to demonstrate coherence control of electrical current using the above parameters. Fig. 2(a) shows the region between a pair of electrodes on GaAs being illuminated by a train of harmonically related pulses. Fig. 2(b)

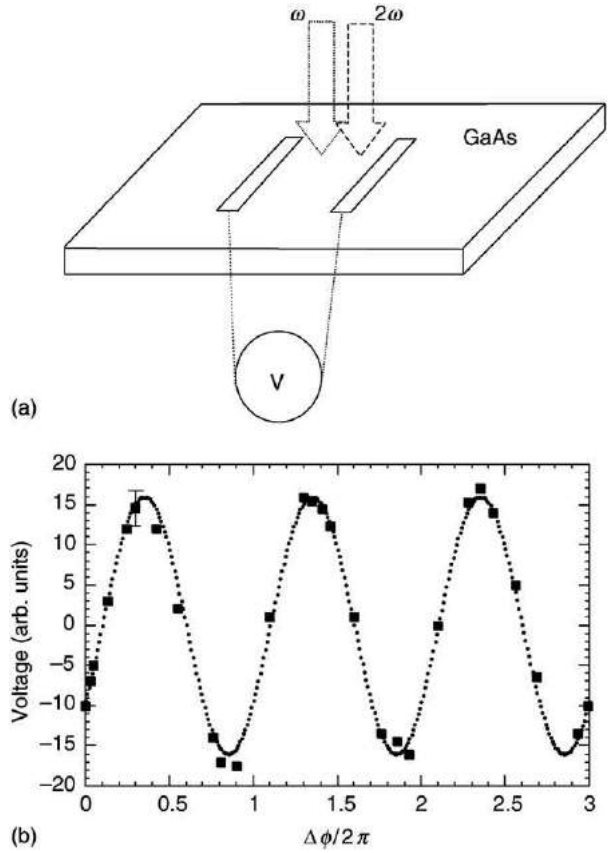


Fig. 2 (a) Experimental setup to measure steady-state voltage (V) across a pair of electrodes on GaAs with the intervening region illuminated by phased radiation at ω and 2ω; (b) induced voltage across a pair of electrodes on a GaAs semiconductor as a function of the phase parameter associated with two harmonically related incident beams.

illustrates how the steady-state experimental voltage across the capacitor changes as the phase parameter $\Delta\phi$ is varied. Transient electrical currents, generated though incident femtosecond optical pulses, have also been detected through the emission of the associated Terahertz radiation.

The current injection via coherence control and conventional cases (i.e., under a DC bias) differs with respect to their evolution. The current injected via coherent control has an onset determined by the rise time of the optical pulses. In the case of normal current production, existing carriers are accelerated by an electric field, and the momentum distribution is never far from isotropic. For a carrier density of 10^{14} cm^{-3} a DC field $\sim$80 kVcm^{-1} is required to produce a current density of 1 kA cm^{-2}. In GaAs, with an electron mobility of 8000 cm^2 V^{-1} s^{-1} this current would occur about 1/2 ps after the field is 'instantaneously' turned on. This illustrates that the coherently controlled phenomenon efficiently and quickly produces a larger current than can be achieved with the redirecting of statistically distributed electrons.

A more detailed analysis must take into account the actual light polarization, crystal symmetry, crystal face, and orientation relative to the optical polarization. For given optical intensities of the two beams, the maximum electrical current injection in the case of GaAs occurs for linearly polarized beams both oriented along the (111) or equivalent direction. However, large currents can also be observed with the light beams polarized along other high symmetry directions.

Coherent Control of Carrier Density, Spin Population, and Spin Current Using Two Color Beams

The processes described above do not exhaust the coherent control effects that can be observed in bulk semiconductors using harmonic beams. Indeed, for noncentrosymmetric materials, certain light polarizations and crystal orientations allow one to coherently control the total carrier generation rate with or without generating an electrical current. In the case of the cubic material GaAs, provided the ω beam has electric field components along two of the three principal crystal axes, with the 2ω beam having a component along the third direction, one can coherently control the total carrier density. However, the overall degree of control here is determined by how 'noncentrosymmetric' the material is.

The spin degrees of freedom of a semiconductor can also be controlled using two color beams. Due to the spin–orbit interaction, the upper valence bands of a typical semiconductor have certain spin characteristics. To date, optical manipulation of electron spin has been largely based on the fact that partially spin-polarized carriers can be injected in a semiconductor via one-photon absorption of circularly polarized light from these upper valence bands. In such carrier injection – where in fact two-photon absorption could be used as well – spins with no net velocity are injected, and then are typically dragged by a bias voltage to produce a spin-polarized current. However, given the protocols discussed above it should not come as a surprise that the two color coherence scheme, when used with certain light polarizations, can coherently control the spin polarization, making it dependent on $\Delta\phi$. Furthermore, for certain polarization combinations it is also possible to generate a spin current with or without an electrical current. Given the excitement surrounding the field of spintronics, where the goal is the use of the spin degree of freedom for data storage and processing, the control of quantities involving the intrinsic angular momentum of the electron is of particular interest.

Various polarization and crystal geometries can be examined for generating spin currents with or without electrical current. For example, with reference to Fig. 3(a), for both beams propagating in the **z** (*normal*) direction of a crystal and possessing the same circular polarization, an electrical current can be injected in the (*xy*) plane, at an angle from the crystallographic **x** direction dependent on the relative phase parameter, $\Delta\phi$; this current is spin-polarized in the **z** direction. As well, the injected carriers have a $\pm$**z** component of their velocity, as many with one component as with the other; but those going in one direction are preferentially spin-polarized in one direction in the (*xy*) plane, while those in the other direction are preferentially spin-polarized in the opposite direction. This is an example of a spin current in the absence of an electrical current.

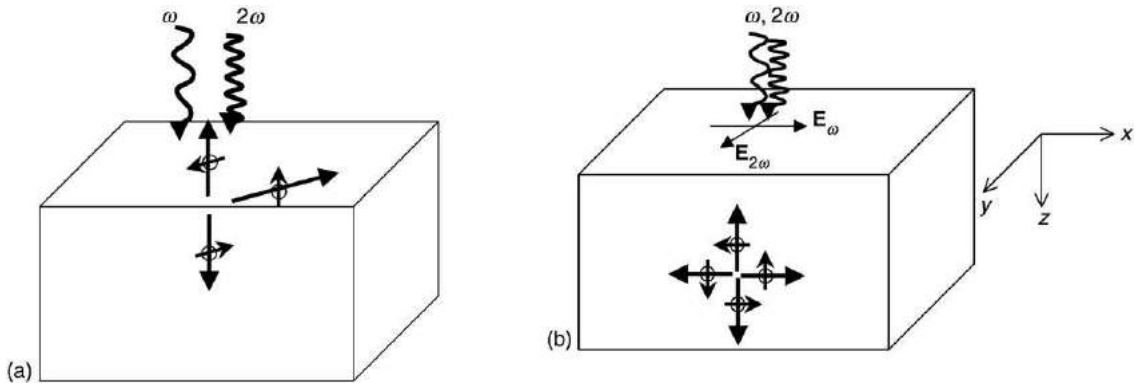


Fig. 3 (a) Excitation of a semiconductor by co-circularly polarized ω and 2ω pulses. Arrows indicate that a spin polarized electrical current is generated in the *x*–*y* plane in a direction dependent on $\Delta\phi$ while a pure spin current is generated in the beam propagation (*z*) direction. (b) Excitation of a semiconductor by orthogonally, linearly polarized ω and 2ω pulses. Arrows indicate that a spin polarized electrical current is generated in the direction of the fundamental beam polarization as well as along the beam propagation (*z*) direction.

Such a pure spin current that is perhaps more striking is observable with the two beams cross linearly polarized, for example with the fundamental beam in the **x** direction and the second harmonic beam in the **y** direction as shown in Fig. 3(b). Then there is no net spin injection; the average spin in any direction is zero. And for a vanishing phase parameter $\Delta\phi$ there is no net electrical current injection. Yet, for example, the electrons injected with a $+\mathbf{x}$ velocity component will have one spin polarization with respect to the **z** direction, while those injected with a $-\mathbf{x}$ component to their velocity will have the opposite spin polarization with respect to the **z** direction.

The examples given above are the spin current analog of the two-color electrical current injection discussed above. There should also be the possibility of injecting a spin current with a single beam into crystals lacking center-of-inversion symmetry.

The simple analysis presented above relies on simple quantum mechanical ideas and calculations using essentially nothing more than Fermi's Golden Rule. Yet the process involves a mixing of two frequencies, and can therefore be thought of as a nonlinear optical effect. Indeed, one of the key ideas to emerge from the theoretical study of such phenomena is that an interpretation of coherence control effects alternate to that provided by the simple quantum interference picture that is provided by the usual susceptibilities of nonlinear optics. These susceptibilities are, of course, based on quantum mechanics, but the macroscopic viewpoint allows for the identification and classification of the effects in terms of 2nd order nonlinear optical effects, 3rd order optical effects, etc. Indeed, one can generalize many of the processes we have discussed above to a hierarchy of frequency mixing effects or high-order nonlinear processes involving multiple beams with frequency $n\omega,m\omega,p\omega\ldots$ with n,m,p being integers. Many of theses schemes require high intensity of one or more of the beams, but can occur in a simple semiconductor.

Coherent Control of Electrical Current Using Single Color Beams

It is also possible to generate coherence control effects using beams at a single frequency (ω), if one focuses on the two orthogonal components (e.g., x and y) polarization states and uses a noncentrosymmetric semiconductor of reduced symmetry such as a strained cubic semiconductor or a wurtzite material such as CdS or CdSe. In this case, interference between absorption pathways associated with the orthogonal components can lead to electrical current injection given by

$$dJ/dt \propto E^{\omega}E^{\omega}\sin(\phi^{x}_{\omega} - \phi^{y}_{\omega}) \tag{8}$$

Since in this process current injection is linear in the beam's intensity, the high intensities necessary for two-photon absorption are not necessary. Nonetheless, the efficacy of this process is limited by the fact that it relies on the breaking of center-of-inversion symmetry of the underlying crystal. It is also clear that the maximum current occurs for circularly polarized light and that right and left circularly polarized light lead to a difference in sign of the current injection. Finally, for this particular single beam scheme, when circularly polarized light is used the electrical current is partially spin polarized.

Conclusions

Through the phase of optical pulses it is possible to control electrical and spin currents, as well as carrier density, in bulk semiconductors on a time-scale that is limited only by the rise time of the optical pulse and the intrinsic response of the semiconductor. A few optically based processes that allow one to achieve these types of control have been illustrated here. Although at one level one can understand these processes in terms of quantum mechanical interference effects, at a macroscopic level one can understand these control phenomena as manifestations of nonlinear optical phenomena. For fundamental as well as applied reasons, our discussion has focused on coherence control effects using the continuum states in bulk semiconductors, although related effects can also occur in quantum dot, quantum well, and superlattice semiconductors. Applications of these control effects will undoubtedly exploit the all-optical nature of the process, including the speed at which the effects can be turned on or off. The turn-off effects, although not discussed extensively here, are related to transport phenomena as well as momentum scattering and related dephasing processes.

See also: Foundations of Coherent Transients in Semiconductors

Further Reading

Brumer, P., 1995. Laser control of chemical reactions. Scientific American 272, 56.
Rabitz, H., de Vivie-Riedle, R., Motzkus, M., Kompa, K., 2000. Whither the future of controlling quantum phenomena? Science 288, 824.
Shah, J., 1999. Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures. New York: Springer.
van Driel, H.M., Sipe, J.E., 2001. Coherent control of photocurrents in semiconductors. In: Tsen, T. (Ed.), Ultrafast Phenomena in Semiconductors. New York: Springer Verlag, p. 261.