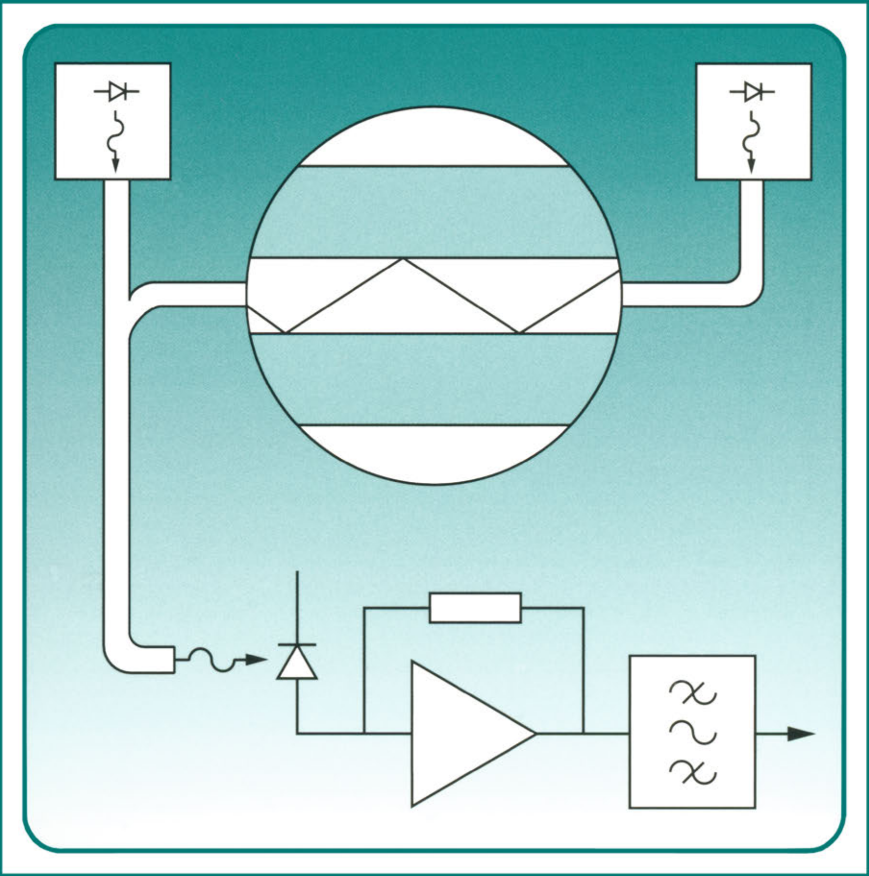
OPTICAL COMMUNICATIONS

Second Edition



M. J. N. SIBLEY

MACMILLAN NEW ELECTRONICS
INTRODUCTIONS TO ADVANCED TOPICS

Optical Communications

Macmillan New Electronics Series
Series Editor: Paul A. Lynn

G. J. Awcock and R. Thomas, *Applied Image Processing*
Rodney F. W. Coates, *Underwater Acoustic Systems*
M. D. Edwards, *Automatic Logic Synthesis Techniques for Digital Systems*
P. J. Fish, *Electronic Noise and Low Noise Design*
W. Forsythe and R. M. Goodall, *Digital Control*
C. G. Guy, *Data Communications for Engineers*
Paul A. Lynn, *Digital Signals, Processors and Noise*
Paul A. Lynn, *Radar Systems*
R. C. V. Macario, *Cellular Radio – Principles and Design*
A. F. Murray and H. M. Reekie, *Integrated Circuit Design*
F. J. Owens, *Signal Processing of Speech*
Dennis N. Pim, *Television and Teletext*
M. Richharia, *Satellite Communications Systems – Design Principles*
M. J. N. Sibley, *Optical Communications,* second edition
P. M. Taylor, *Robotic Control*
G. S. Virk, *Digital Computer Control Systems*
Allan Waters, *Active Filter Design*

Optical Communications

Components and Systems

M. J. N. Sibley

Division of Electronics and Communications
The University of Huddersfield

Second Edition

Macmillan New Electronics
Introductions to Advanced Topics

MACMILLAN

First edition 1990
Second edition 1995

Published by
THE MACMILLAN PRESS LTD
Houndmills, Basingstoke, Hampshire RG21 2XS
and London
Companies and representatives
throughout the world

ISBN 978-0-333-61792-2 ISBN 978-1-349-13524-0 (eBook)
DOI 10.1007/978-1-349-13524-0

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

Contents

Series Editor's Foreword

The rapid development of electronics and its engineering applications ensures that new topics are always competing for a place in university and college courses. But it is often difficult for lecturers to find suitable books for recommendation to students, particularly when a topic is covered by a short lecture module, or as an 'option'.

Macmillan New Electronics offers introductions to advanced topics. The level is generally that of second and subsequent years of undergraduate courses in electronic and electrical engineering, computer science and physics. Some of the authors will paint with a broad brush; others will concentrate on a narrower topic, and cover it in greater detail. But in all cases the titles in the Series will provide a sound basis for further reading of the specialist literature, and an up-to-date appreciation of practical applications and likely trends.

The level, scope and approach of the Series should also appeal to practising engineers and scientists encountering an area of electronics for the first time, or needing a rapid and authoritative update.

Paul A. Lynn

Preface to First Edition

Since the mid 1970s, the field of optical communications has advanced considerably. Optical fibre attenuations have been reduced from over 1000 dB/km to below 0.5 dB/km, and light sources are now available that can launch several milli-watts of power into a fibre. Optical links are now to be found in short-haul industrial routes, as well as in long-haul telecommunications routes. In order to design and maintain these links, it is important to understand the operation of the individual system components, and it is my hope that this book will provide the relevant information.

I have tried to aim the level of this text so that it is suitable for students on the final year of undergraduate courses in Electrical and Electronic Engineering, and Physics, as well as for practising engineers requiring a knowledge of optical communications. The text should also serve as an introduction for students studying the topic at a higher level. The work presented here assumes that the reader is familiar with Maxwell's equations, and certain aspects of communications theory. Such information can be readily found in relevant textbooks.

The information presented has come from a wide variety of sources – many of which appear in the Bibliography at the end of the book. In order to keep the list of references down to manageable proportions, I have only selected certain key papers and books. In order to obtain further information, the interested reader should examine the references that these works themselves give. Most of the journals the papers appear in will be available from any well-equipped library; otherwise they can be obtained through an inter-library loan service. Because of the length of this book, the information obtained from these sources has been heavily condensed. In view of this, I regret any errors or omissions that may have arisen, and hope that they will not detract from this text.

I wish to acknowledge the assistance of Dr K. Fullard of the Joint European Tours project at Culham, for supplying the information about the optical LAN that appears in chapter 7. I also wish to thank the publisher, Mr M. J. Stewart of Macmillan Press, for his guidance, and the series editor, Dr Paul A. Lynn, for his valuable comments on the text. During the compilation of this text, many of my colleagues at the Department of Electrical and Electronic Engineering, The Polytechnic of Huddersfield were party to several interesting discussions. In particular, I wish to acknowledge the help of Dr R. T. Unwin, who read the draft manuscript, offered constructive criticism, and showed a great deal of patience throughout the preparation of this book.

Finally, I would like to thank my family and friends for their continued support and encouragement.

Preface to Second Edition

Since the first edition of this book appeared, optical communications has come of age; system designers are now able to choose from a wide variety of components operating at various wavelengths and bit-rates. Although this is indicative of a mature technology, the drive to send faster and faster pulses over longer and longer distances continues. Work is currently progressing on links that operate at 10 Gbit/s. The use of soliton pulses and fibre amplifiers means that transmission distances can be vast. All these factors serve to increase the number of areas that a student of optical communications has to study, as reflected in the extended breadth of this book.

In revising this text, I have tried to keep the mathematical work to a minimum. However, in this edition, chapter 2 contains the solution of Maxwell's equations as applied to planar and cylindrical waveguides. I am conscious that some readers may not think that this is an improvement, and so I have included the ray-path analysis used in the first edition!

The other major change is that chapter 3 now deals with semiconductor physics in some detail. Knowledge of this area will help the reader to understand the operation of semiconductor light sources, and lasers in general. Solid-state and gas lasers are also examined in this chapter, together with external modulators. This reflects the move towards running a laser continuously while using a Mach–Zehnder modulator to turn the light on and off.

Some of the advanced components I described in chapter 7 of the first edition are now being routinely used, and so they appear in the main body of the text. In common with the first edition, new techniques and components are described in this chapter. In particular, I have included quantum-well lasers, and sub-carrier multiplex systems. Of these two, quantum-well lasers are likely to have a major impact over the next few years.

I have not included any problems in this book. Instead I have introduced worked examples throughout the main body of the text, with the hope that they will aid in the understanding of the subject. Hopefully lecturers who adopt this text will be able to adapt these examples to suit their own purposes.

It is my hope that readers will find this text interesting and informative, and that they too will find the area of optical communications as fascinating as I have since 1981.

M. J. N. Sibley

List of Symbols

α	attenuation constant/absorption coefficient
α_{e}	electron ionisation coefficient
α_{h}	hole ionisation coefficient
$\boldsymbol{a}_x$, $\boldsymbol{a}_y$, $\boldsymbol{a}_z$	unit vectors
A_0	total preamplifier voltage gain
$A(\omega)$	total voltage gain of receiver system
β	phase constant
b	binding parameter of modes in a waveguide
B	bit-rate in digital or bandwidth in analogue systems
B_{eq}	noise equivalent bandwidth
c	velocity of light in a vacuum (3×10^8 m s^{-1})
C_π	base–emitter capacitance
C_{c}	collector–base capacitance
C_{d}	total diode capacitance
C_{f}	feedback resistance parasitic capacitance
C_{gd}	gate–drain capacitance
C_{gs}	gate–source capacitance
C_{in}	input capacitance of preamplifier
C_{j}	junction capacitance
C_{s}	stray input capacitance
C_{T}	total receiver input capacitance
δn	fractional refractive index difference
δE_{c}	conduction band step
δE_{v}	valence band step
D_{mat}	material dispersion coefficient
D_{n}	diffusion coefficient for electrons
D_{p}	diffusion coefficient for holes
D_{wg}	waveguide dispersion coefficient
ϵ_0	permittivity of free-space (8.854×10^{-12} F/m)
ϵ_{r}	relative permittivity
E	electric field strength
E_{c}	conduction band energy level
E_{f}	Fermi level
E_{g}	band-gap difference
E_{v}	valence band energy level

$F(M)$	excess noise factor
γ	propagation coefficient
g	gain per unit length
g_{m}	transconductance
h	Planck's constant (6.624×10^{-34} J s)
$h_{\mathrm{f}}(t)$	pre-detection filter impulse response
$h_{\mathrm{out}}(t)$	output pulse shape
$h_{\mathrm{p}}(t)$	input pulse shape
H	magnetic field strength
$H_{\mathrm{eq}}(\omega)$	equalising network transfer function
$H_{\mathrm{f}}(\omega)$	pre-detection filter transfer function
$H_{\mathrm{out}}(\omega)$	Fourier transform (FT) of output pulse
$H_{\mathrm{p}}(\omega)$	FT of received pulse
$H_{\mathrm{T}}(\omega)$	normalised transimpedance
$<i_{\mathrm{n}}^{2}>_{0}$	mean square (m.s.) noise current for logic 0 signal
$<i_{\mathrm{n}}^{2}>_{1}$	m.s. noise current for logic 1 signals
$<i_{\mathrm{n}}^{2}>_{\mathrm{c}}$	m.s. equivalent input noise current of preamplifier
$<i_{\mathrm{n}}^{2}>_{\mathrm{DB}}$	m.s. photodiode bulk leakage noise current
$<i_{\mathrm{n}}^{2}>_{\mathrm{DS}}$	m.s. photodiode surface leakage noise current
$<i_{\mathrm{n}}^{2}>_{\mathrm{pd}}$	m.s. photodiode noise current
$<i_{\mathrm{n}}^{2}>_{\mathrm{Q}}$	quantum noise
$<i_{\mathrm{n}}^{2}>_{\mathrm{T}}$	total signal-independent m.s. noise current
$<i_{\mathrm{s}}^{2}>$	m.s. photodiode signal current
$i_{\mathrm{s}}(t)$	photodiode signal current
I_{2}, I_{3}	bandwidth type integrals
I_{b}	base current
I_{c}	collector current
I_{d}	total dark current
I_{diode}	total diode current
I_{g}	gate leakage current
I_{m}	multiplied diode current
I_{max}	maximum signal diode current
I_{min}	minimum signal diode current
I_{s}	signal-dependent, unmultiplied photodiode current
$<I_{\mathrm{s}}>$	average signal current
$<I_{\mathrm{s}}>_{0}$	average signal current for a logic 0
$<I_{\mathrm{s}}>_{1}$	average signal current for a logic 1
I_{th}	threshold current
I_{DB}	photodiode bulk leakage current
I_{DS}	photodiode surface leakage current
ISI	inter-symbol interference
J	current density
J_{th}	threshold current density
k	Boltzmann's constant (1.38×10^{-23} J/K)

k_0	free-space propagation constant of a propagating mode
λ	wavelength
L_n	diffusion length in n-type material
L_p	diffusion length in p-type material
m	modulation depth
M	multiplication factor
M_{opt}	optimum avalanche gain
η	quantum efficiency
n	refractive index
n_{eff}	effective refractive index
n_i	total intrinsic carrier density
n_n	electron density in n-type material
n_p	electron density in p-type material
$\langle n^2\rangle_T$	total m.s. output noise voltage
N_a	acceptor atom density
N_c	density of electrons in the conduction band
N_d	donor atom density
N_v	density of holes in the valence band
μ_0	permeability of free-space ($4\pi \times 10^{-7}$ H/m)
μ_r	relative permeability
N	mode number (integer)
N_D	donor atom doping level
N_g	group refractive index
N_{max}	maximum number of modes
NA	numerical aperture
p_n	hole density in n-type material
p_p	hole density in p-type material
P	average received power
P_e	probability of error
q	electron charge (1.6×10^{-19} C)
Q	signal-to-noise parameter
r_π	base–emitter resistance
$r_{bb'}$	base-spreading resistance
r_e	reflection coefficient
R_1, R_2	mirror reflectivity in resonator
R_b	photodiode load resistor
R_f	feedback resistor
R_{in}	preamplifier input resistance
R_j	photodiode shunt resistance
R_L	load resistor
R_o	photodiode responsivity (A/W)
R_s	photodiode series resistance
R_T	low-frequency transimpedance
σ	r.m.s. width of Gaussian distribution (line-width, etc.)

σ_{mat}	material dispersion per unit length
σ_{mod}	modal dispersion per unit length
σ_{off}	r.m.s. output noise voltage for logic 0
σ_{on}	r.m.s. output noise voltage for logic 1
σ_{wg}	waveguide dispersion per unit length
S	instantaneous power flow (Poynting vector)
S_{av}	average power flow
S_{E}	series noise generator (V^2/Hz)
$S_{\mathrm{eq}}(f)$	equivalent input noise current spectral density (A^2/Hz)
S_{I}	shunt noise generator (A^2/Hz)
S/N	signal-to-noise ratio
τ	time constant
τ_{n}	electron lifetime in p-type material
τ_{nr}	non-radiative recombination time
τ_{p}	hole lifetime in n-type material
τ_{ph}	stimulated photon lifetime
τ_{r}	radiative recombination time
τ_{sp}	spontaneous photon lifetime
t_{e}	transmission coefficient
T	absolute temperature (Kelvin)
v_{g}	group velocity
v_{max}	maximum output signal voltage
v_{min}	minimum output signal voltage
v_{p}	phase velocity
V	normalised frequency in a waveguide
V_{br}	reverse breakdown voltage
V_{s}	output signal voltage
V_{T}	threshold voltage
y	normalised frequency variable
Z	impedance of dielectric to TEM waves
$Z_{\mathrm{c}}(s)$	closed-loop transimpedance
$Z_{\mathrm{f}}(s)$	feedback network transimpedance
Z_{in}	total input resistance
Z_0	impedance of free space
$Z_{\mathrm{o}}(s)$	open-loop transimpedance
$Z_{\mathrm{T}}(\omega)$	transimpedance
θ_{i}	angle of incidence
θ_{r}	angle of reflection
θ_{t}	angle of transmission

1 Introduction

Although the subject of this book is optical communications, the field encompasses many different aspects of electronic engineering: electromagnetic theory, semiconductor physics, communications theory, signal processing, and electronic design. In a book of this length, we could not hope to cover every one of these different fields in detail. Instead, we will deal with some aspects in-depth, and cover others by more general discussion. Before we start our studies, let us see how modern-day optical communications came about.

1.1 Historical background

The use of light as a means of communication is not a new idea; many civilisations used sunlight reflected off mirrors to send messages, and communication between warships at sea was achieved using Aldis lamps. Unfortunately, these early systems operated at very low data-rates, and failed to exploit the very large bandwidth of optical communications links.

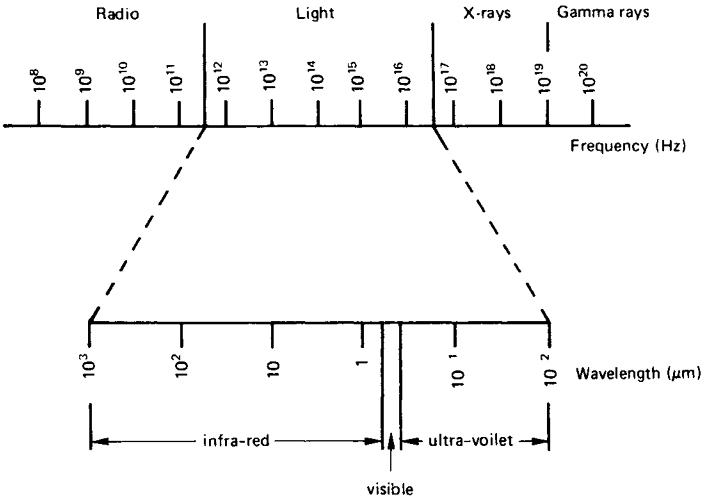


Figure 1.1 The electromagnetic spectrum

A glance at the electromagnetic spectrum shown in figure 1.1, reveals that visible light extends from 0.4 to 0.7 μm which converts to a bandwidth of 320 THz (1 THz = 10^{12} Hz). Even if only 1 per cent of this capability were available, it would still allow for 80 billion, 4 kHz voice channels! (If we could transmit these channels by radio, they would occupy the whole of the spectrum from d.c. right up to the far infra-red. As well as not allowing for any radio or television broadcasts, the propagation characteristics of the transmission scheme would vary tremendously.) The early optical systems used incandescent white light sources, the output of which was interrupted by a hand-operated shutter. Apart from the obvious disadvantage of a low transmission speed, a white light source transmits all the visible, and some invisible, wavelengths at once. If we draw a parallel with radio systems, this is equivalent to a radio transmitter broadcasting a single programme over the whole of the radio spectrum – very inefficient! Clearly, the optical equivalent of an oscillator was needed before light-wave communications could develop.

A breakthrough occurred in 1960, with the invention of the ruby laser by T. H. Maimon [1], working at Hughes Laboratories, USA. For the first time, an intense, coherent light source operating at just one wavelength was made available. It was this development that started a flurry of research activity into optical communications.

Early experiments were carried out with line-of-sight links; however, it soon became apparent that some form of optical waveguide was required. This was because too many things can interefere with light-wave propagation in the atmosphere: fog, rain, clouds, and even the occasional flock of pigeons. Hollow metallic waveguides were initially considered but, because of their impracticality, they were soon ruled out. By 1963, bundles of several hundred glass fibres were already being used for small-scale illumination. However, these early fibres had very high attenuations (>1000 dB/km) and so their use as a transmission medium for optical communications was not considered.

It was in 1966 that C. K. Kao and G. A. Hockman [2] (working at the Standard Telecommunications Laboratories, UK) postulated the use of glass fibres as optical communications waveguides. Because of the high attenuation of the glass, the idea was initially treated with some scepticism; in order to compete with existing co-axial cable transmission lines, the glass fibre attenuation had to be reduced to less than 20 dB/km. However, Kao and Hockman studied the loss mechanisms and, in 1970, workers at the Corning glass works, USA, produced a fibre with the required attenuation. This development led to the first laboratory demonstrations of optical communications with glass fibre, in the early 1970s. A study of the spectral response of glass fibres showed the presence of low-loss transmission windows at 850 nm, 1.3 μm, and 1.55 μm. Although the early optical links used the 850 nm window, the longer wavelength windows exhibit lower losses, typically 0.2 dB/km, and so most modem links use 1.3 and 1.55 μm wavelength light.

While work progressed on reducing fibre attenuation, laser development continued apace. Ruby lasers have to be 'pumped' with the light from a flash lamp, and so the modulation speed is very low. The advent of the semiconductor laser, in 1962, meant that a fast light source was available. The material used was gallium–arsenide, *GaAs*, which emits light at a wavelength of 870 nm. With the discovery of the 850 nm window, the wavelength of emission was reduced by doping the GaAs with aluminium, *Al*. Later modifications included different laser structures to increase device efficiency and lifetime. Various materials were also investigated, to produce devices for operation at 1.3 and 1.55 μm. Unfortunately lasers are quite expensive, and so low-cost light emitting diodes, *LEDs*, have also been developed. Semiconductor sources are now available which emit at any one of many wavelengths, with modulation speeds of several Gbit/s being routinely achieved in the laboratory.

At the receiver, a photodetector converts the optical signal back into an electrical one. The early optical links used avalanche photodiodes, *APDs*, which exhibit current multiplication, that is, the single electron–hole pair produced by the detection of a photon of light generates more electron–hole pairs, so amplifying the signal. In 1973, S. D. Personick [3] (working at Bell Laboratories in the USA) analysed the performance of an optical PCM receiver. This theoretical study showed that an APD feeding a high input impedance preamplifier, employing an FET input stage, would result in the best receiver sensitivity. Unfortunately, the early APDs required high bias voltages, typically 200–400 V, and this made them unattractive for use in terminal equipment.

It was in 1978 that D. R. Smith, R. C. Hooper and I. Garrett [4] (all working at British Telecom Research Laboratories, Martlesham Heath, UK) published a comparison between an APD and a PIN photodiode followed by a low-capacitance, microwave FET input preamplifier (the so-called *PINFET* receiver). They showed that PINFET receivers using a hybrid thick-film construction technique could achieve a sensitivity comparable to that of an APD receiver. They also indicated that PIN receivers for the 1.3 and 1.55 μm transmission windows would out-perform an equivalent APD receiver. (The reasons for this will become clear when we discuss photodiodes in chapter 4). So, the use of PINFET receivers operating in the long-wavelength transmission windows meant that signals could be sent over very long distances – ideal for trunk route telephone links.

The work on long-haul routes aided the development of short-haul industrial links. From an industrial viewpoint, the major advantage of an optical link is that it is immune to electromagnetic interference. Hence optical fibre links can operate in electrically noisy environments, which would disrupt a hard-wire system. For short-haul applications, expensive low-loss glass fibres, lasers and very sensitive receivers are not required. Instead, all-plastic fibres, LEDs and low-cost bipolar preamplifiers are often used. These components are readily available on the commercial market, and are usually supplied with connectors attached for ease of use.

1.2 The optical communications link

An optical communications link is similar to other links in that it consists of a transmitter, a communications channel and a receiver. A more detailed examination (figure 1.2) shows that the communications channel is an optical fibre. In order for the fibre to guide light, it must consist of a *core* of material whose refractive index is greater than that of the surrounding medium – known as the *cladding*. Depending on the design of the fibre, light is constrained to the core by either *total internal reflection* or *refraction*. We can describe the propagation of light in glass with the aid of ray optics; however, in chapter 2, we shall make use of Maxwell's equations. We do this because it will give us a valuable insight into certain effects that cannot be easily explained with ray optics. Also presented in this chapter is a discussion of attenuation mechanisms and fibre fabrication methods.

In optical links, the transmitter is a light source whose output acts as the carrier wave. Although frequency division multiplexing, *fdm*, techniques are used in analogue broadcast systems, most optical communications links use digital time division multiplexing, *tdm*, techniques. The easiest way to modulate a carrier wave with a digital signal is to turn it on and off, so-called *on–off keying*, or amplitude shift keying, *ASK*. In optical systems this is commonly achieved by varying the source drive current directly, so causing a proportional change in optical power. The most common light sources in use at present are semiconductor laser diodes and LEDs, and we shall deal with these devices in chapter 3. Also included in this chapter is a study of solid-state and gas lasers, together with an examination of external modulators.

At the receiving end of an optical link, a PIN photodiode, or an APD, converts the modulated light back into an electrical signal. The photodiode

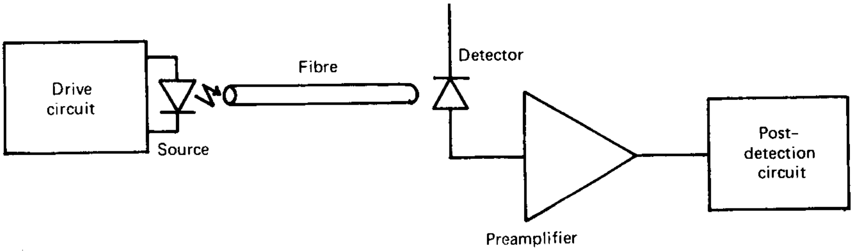


Figure 1.2 A basic optical communications link

current is directly proportional to the incident optical power. (If we draw a parallel with radio receivers, this detection process is similar to the very simple direct detection radio receiver.) Depending on the wavelength of operation, photodiodes can be made out of silicon, germanium or an alloy of indium, gallium and arsenic. We shall consider PIN photodiodes and APDs in chapter 4.

Ultimately, for a limited transmitter power and wide bandwidth channel, it is the receiver noise that limits the maximum transmission distance, and hence repeater spacings. The receiver noise depends upon bandwidth – a low bandwidth receiver results in low noise. However, if the bandwidth is too low, the received signal will be distorted. Therefore, as shown in chapter 5, receiver design is often a compromise between minimising the noise, while maintaining an acceptable degree of signal corruption.

Low-noise preamplifiers are used to boost the small amplitude signal appearing at the output of the photodetector. At present there are two main types: the high-input impedance FET design, *PINFET*, and the *transimpedance feedback* design. Of the two designs, the PINFET preamplifier is currently the most sensitive design available and, as such, finds applications in long-haul telecommunications routes. Transimpedance designs are usually fabricated with bipolar transistors and, although they are noisier than PINFET designs, they are generally cheaper to produce and find applications in short-to-medium haul routes. Bipolar transistors are generally more reliable than FETs, and so bipolar transimpedance preamplifiers are also used in the repeaters in submarine optical links. Preamplifier design is discussed in chapter 6.

In chapter 7, we shall consider the design of several optical transmission links in current use. The examples covered include a long-haul telecommunications link, and a short-haul computer communications link operating in an electrically hostile environment. As with many developing technologies, new advances are being made at a very rapid pace, and so chapter 7 will also consider some of the latest developments. The topics we will examine include the use of very low-loss glass fibres operating with 2.3 μm wavelength light; novel laser designs, and coherent detection receivers which have a far higher sensitivity than direct detection receivers. As well as increasing receiver sensitivity, this last advance can increase the capacity of the optical channel. Although optical fibres exhibit a very large bandwidth, time division multiplexing techniques do not make use of the available capacity – any increase in transmission speed places great strain on the speed of the digital processing circuits. Radio systems use frequency division multiplexing techniques, with each station being allocated a different frequency. Optical coherent receiver systems can operate on the same principle, with separate optical frequencies carrying high-speed data. In this way, the effective data-rate of an optical link will not be set by the speed at which the digital ICs can process the data.

2 Optical Fibre

In most optical communication links, it is the optical fibre that provides the transmission channel. The fibre consists of a solid cylinder of transparent material, the *core*, surrounded by a *cladding* of similar material. Light waves propagate down the core in a series of plane wavefronts, or *modes*; the simple light ray path used in elementary optics is an example of a mode. For this propagation to occur, the refractive index of the core must be larger than that of the cladding and there are two basic structures which have this property; *step-index* and *graded-index* fibres. Of the step-index types, there are multi-mode, *MM*, fibres (which allow a great many modes to propagate) and single-mode, *SM*, fibres (which only allow one mode to propagate). Although graded-index fibres are normally MM, some SM fibres are available.

The three fibre types, together with their respective refractive index profiles, are shown in figure 2.1. (In this figure, *n* is the refractive index of the material.) The cross-hatched area represents the cladding, the diameter of which ranges from 125 μm to a typical maximum of 1 mm. The core diameter can range from 8 μm, for SM fibres, up to typically 50 μm for large core MM fibres.

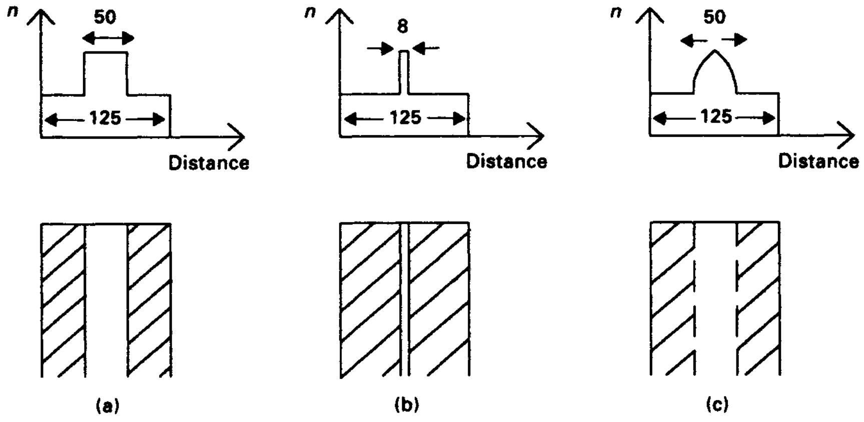


Figure 2.1 Typical refractive index profiles of (a) step-index multimode, (b) step-index single-mode, and (c) graded-index multimode fibres. (All dimensions are in μm.)

Most of the optical fibres in use today are made of either silica glass (SiO_2) or plastic. The change in refractive index, between the core and cladding, is achieved by the addition of certain dopants to the glass; all-plastic fibres use different plastics for the core and cladding. All-glass, SM fibres exhibit very low losses and high bandwidths, which make them ideal for use in long-haul telecommunications routes. Unfortunately such fibres are expensive to produce and so are seldom found in short-haul (less than 500 m length) industrial links.

Large core fibres for use in medical and industrial applications are generally made of plastic, making them more robust than the all-glass types and much cheaper to manufacture. However, the very high attenuation and low bandwidth of these fibres tend to limit their uses in communications links. Medium-haul routes, between 500 m and 1 km lengths, generally use plastic cladding/glass core fibre, otherwise known as *plastic clad silica*, or *PCS*. All-plastic and PCS fibres are almost exclusively step-index, multi-mode types.

The attenuation and bandwidth of an optical fibre will determine the maximum distance that signals can be sent. Attenuation is usually expressed in dB/km, while bandwidth is usually quoted in terms of the *bandwidth length product*, which has units of GHz km, or MHz km. Attenuation depends on impurities in the core, and so the fibre must be made from very pure materials. To some extent the bandwidth also depends on the core impurities; however, as we shall see later, the bandwidth is usually limited by the number of propagating modes. This explains why single-mode fibres have a very large bandwidth.

In this chapter we shall examine the properties and design of optical fibres. Initially we will solve Maxwell's equations in an infinite block of dielectric material (glass), and then consider propagation in a planar dielectric waveguide. When we come to examine propagation in optical fibre, we will solve Maxwell's equations as applied to a cylindrical waveguide. This involves some rather complicated mathematics which some readers may prefer to omit at a first reading. In view of this, important results from the full analysis are quoted in the relevant sections. (The work in this chapter assumes that the reader is familiar with Maxwell's equations. Most books on electromagnetism cover the derivation of these equations [1–3].)

2.1 Propagation of light in a dielectric

In some instances it is convenient to treat light as a stream of particles, or *photons*; in others as an electromagnetic wave. Here we will treat light as a wave, and apply Maxwell's equations to study light wave propagation. We will consider a plane wavefront, of arbitrary optical frequency, travelling in an infinite block of dielectric (glass). This will give us a valuable insight into certain fibre characteristics which we cannot easily explain in terms of simple geometric ray optics.

2.1.1 The wave equation

In order to study the variation of the E and H fields in a dielectric, we need to derive the relevant wave equations. We take as our starting point, the following Maxwell's equations:

$$\nabla \times \boldsymbol{E} = -\mu \frac{\partial \boldsymbol{H}}{\partial t} \tag{2.1a}$$

and

$$\nabla \times \boldsymbol{H} = \epsilon \frac{\partial \boldsymbol{E}}{\partial t} + \sigma \boldsymbol{E} \tag{2.1b}$$

If $\boldsymbol{E}$ and $\boldsymbol{H}$ vary sinusoidally with time at the frequency of the light we are transmitting, then we can use the phasor forms of $\boldsymbol{E}$ and $\boldsymbol{H}$. Thus we can write

$$\boldsymbol{E} = E_x \exp(\mathrm{j}\omega t)\boldsymbol{a}_x \qquad \text{and} \qquad \boldsymbol{H} = H_y \exp(\mathrm{j}\omega t)\boldsymbol{a}_y$$

where $\boldsymbol{a}_x$ and $\boldsymbol{a}_y$ are the x- and y-axes unit vectors. We can now write (2.1a) and (2.1b) as

$$\frac{\partial \boldsymbol{E}}{\partial z} = -\mathrm{j}\omega\mu \boldsymbol{H} \quad (2.2\text{a}) \qquad \text{and} \qquad -\frac{\partial \boldsymbol{H}}{\partial z} = \mathrm{j}\omega\epsilon \boldsymbol{E} + \sigma \boldsymbol{E} \quad (2.2\text{b})$$

We can manipulate these two equations to give the wave equations which describe the propagation of a plane transverse electromagnetic (TEM) wave in the material. Thus if we differentiate (2.2a) with respect to z, and substitute from (2.2b), we get

$$\frac{\partial^2 \boldsymbol{E}}{\partial z^2} = -\omega^2\mu\epsilon \boldsymbol{E} + \mathrm{j}\omega\mu\sigma \boldsymbol{E} \tag{2.3}$$

and, if we differentiate (2.2b) with respect to z, and substitute from (2.2a), we get

$$\frac{\partial^2 \boldsymbol{H}}{\partial z^2} = -\omega^2\mu\epsilon \boldsymbol{H} + \mathrm{j}\omega\mu\sigma \boldsymbol{H} \tag{2.4}$$

If we now let $\gamma^2 = -\omega^2\mu\epsilon + \mathrm{j}\omega\mu\sigma$, one possible solution to these equations is

$$\boldsymbol{E} = E_{x0}\exp(\mathrm{j}\omega t)\exp(-\gamma z)\boldsymbol{a}_x \tag{2.5}$$

and

$$\boldsymbol{H} = H_{y\mathrm{o}}\exp(\mathrm{j}\omega t)\exp(-\gamma z)\boldsymbol{a}_y \tag{2.6}$$

where the subscript o denotes the values of E and H at the origin of a right-handed cartesian co-ordinate set, and γ is known as the *propagation coefficient*. Writing $\gamma = \alpha + \mathrm{j}\beta$, where α and β are the *attenuation* and *phase constants* respectively, we get

$$\boldsymbol{E} = E_{x\mathrm{o}}\exp(-\alpha z)\cos(\omega t - \beta z)\boldsymbol{a}_x \tag{2.7}$$

and

$$\boldsymbol{H} = H_{y\mathrm{o}}\exp(-\alpha z)\cos(\omega t - \beta z)\boldsymbol{a}_y \tag{2.8}$$

These equations describe a TEM wave travelling in the positive z direction, undergoing attenuation as $\exp(-\alpha z)$. The E and H fields are orthogonal to each other and, as figure 2.2 shows, perpendicular to the direction of propagation.

2.1.2 Propagation parameters

Equations (2.7) and (2.8) form the starting point for a more detailed study of light wave propagation. However, before we proceed much further, we

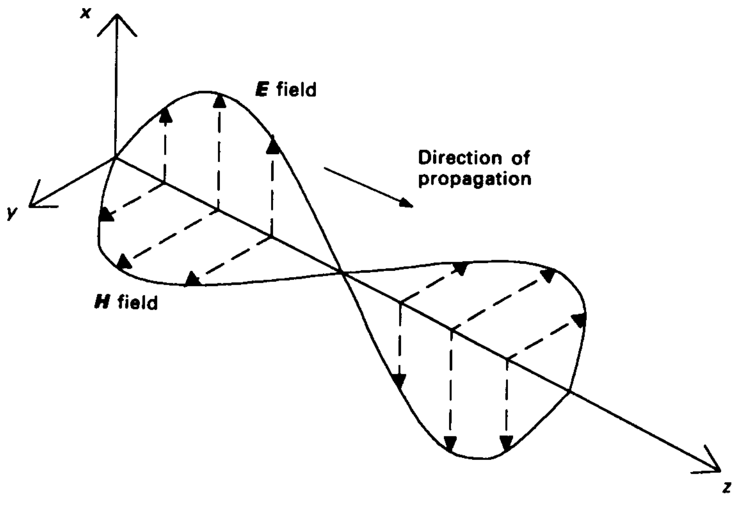


Figure 2.2 Variation of E and H for a TEM wave propagating in the z direction

will find it useful to derive some propagation parameters.

From the previous section, the propagation coefficient, γ, is given by

$$\gamma = \alpha + j\beta \qquad \text{and} \qquad \gamma^2 = -\omega^2\mu\epsilon + j\omega\mu\sigma$$

Hence it is a simple matter to show that

$$\alpha^2 - \beta^2 = -\omega^2\mu\epsilon \tag{2.9}$$

and

$$2\alpha\beta = \omega\sigma\mu \tag{2.10}$$

As glass is an insulator, the conductivity is very low, $\sigma << 0$, and the relative permeability is approximately unity, $\mu_r \approx 1$. Over a distance of a few wavelengths this results in $\alpha \approx 0$ (which implies zero attenuation) and so we can write the E and H fields as

$$\boldsymbol{E} = E_{x0}\cos(\omega t - \beta z)\boldsymbol{a}_x \tag{2.11}$$

and,

$$\boldsymbol{H} = H_{y0}\cos(\omega t - \beta z)\boldsymbol{a}_y \tag{2.12}$$

where $\beta \approx \omega\sqrt{\mu_0\epsilon}$. We can study the propagation of these fields by considering a point on the travelling wave as t and z change.

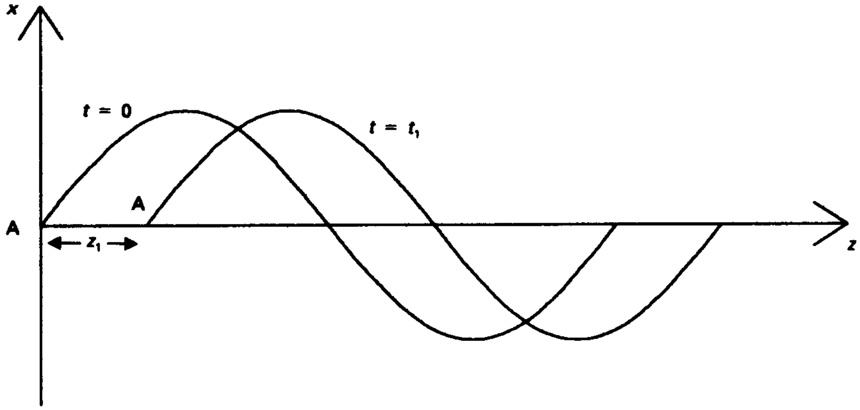


Figure 2.3 Illustrative of the phase velocity of a constant phase point, A, on the E field of a TEM wave

Figure 2.3 shows the sinusoidal variation of the E field with time and distance. If we consider the point A then, at time $t = 0$ and distance $z = 0$, the amplitude of the wave is zero. At time $t = t_1$, the point A has moved a distance z_1 and, as the amplitude of the wave is still zero, we can write

$$\sin(\omega t_1 - \beta z_1) = 0 \qquad \text{and so} \qquad \omega t_1 = \beta z_1 \tag{2.13}$$

Thus we can see that the constant phase point, A, propagates along the z-axis at a certain velocity. This is the *phase velocity*, v_p, given by

$$v_p = \frac{z_1}{t_1} = \frac{\omega}{\beta} = \frac{\omega}{\omega\sqrt{\mu_0\epsilon}} = \frac{1}{\sqrt{\mu_0\epsilon}} \tag{2.14}$$

If the dielectric is free space, then v_p is $\approx 3 \times 10^8$ m s^{-1} (the velocity of light in a vacuum, c). This leads us directly on to the definition of refractive index. For the dielectric, $\epsilon = \epsilon_0\epsilon_r$ and so

$$v_p = \frac{1}{\sqrt{\mu_0\epsilon_0}} \times \frac{1}{\sqrt{\epsilon_r}} = \frac{c}{n} \tag{2.15}$$

where n is the *refractive index* of the dielectric.

It is a simple matter to show that the wavelength of the light in the dielectric, λ, is

$$\lambda = \frac{2\pi}{\beta} = \frac{2\pi v_p}{\omega} = \frac{2\pi c}{n\omega} = \frac{\lambda_0}{n} \tag{2.16}$$

where λ_0 is the wavelength of the light in free-space. This leads us to an alternative definition for β:

$$\beta = \frac{2\pi n}{\lambda_0} \tag{2.17}$$

or

$$\beta = nk_0$$

where k_0 is the *free-space propagation constant.* (It is interesting to note that if ϵ_r, and hence n, vary with frequency, then β and v_p will also vary. Thus if we have two light-waves of slightly different frequencies, the two waves will travel at different velocities and the signal is said to be *dispersed.* We shall return to this point in the next section.)

The impedance of the dielectric to TEM waves, Z, equals E/H, and we can find E as a function of H by using (2.2a). Thus

$$\frac{\partial E}{\partial z} = -\mathrm{j}\mu_0\omega \boldsymbol{H} \quad \text{becomes} \quad -\mathrm{j}\beta \boldsymbol{E} = -\mathrm{j}\mu_0\omega \boldsymbol{H} \quad \text{and so}$$

$$Z = \mu_0\omega/\beta = \sqrt{\mu_0/\epsilon_0\epsilon_r} = Z_0/n \tag{2.18}$$

where Z_0 is the impedance of free-space, 377 Ω.

One final useful parameter is the power flow. If we take the cross product of $\boldsymbol{E}$ and $\boldsymbol{H}$, we will get a third vector, acting in the direction of propagation, with units of W m^{-2}, that is, power flow per unit area. This vector is known as the *Poynting vector*, S, and its *instantaneous* value is given by

$$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H} \tag{2.19}$$

We can find the *average* power flow, S_{av}, in the usual way by integrating (2.19) over one period, and then dividing by the period. In phasor notation form, S_{av}, will be given by

$$S_{av} = R_e\left\{\frac{1}{2}\ \boldsymbol{E} \times \boldsymbol{H}^*\right\} \tag{2.20}$$

where R_e denotes 'the real part of . . .', $\boldsymbol{H}^* = \boldsymbol{H}\exp(-\mathrm{j}\ [\omega t - \phi])$, and ϕ is the temporal phase angle between the E and H fields. (In geometric optics, ray-paths are drawn in the direction of propagation and normal to the plane of $\boldsymbol{E}$ and $\boldsymbol{H}$. Thus the ray-path has the direction of power flow.)

Example

Light of wavelength 600 nm is propagating in a block of transparent material which has the following characteristics:

$$\mu_r = 1;\ \epsilon_r = 5;\ \sigma = 3 \times 10^{-4}\ \text{S/m}$$

Determine the following parameters:

(a) attenuation and phase coefficients;
(b) phase velocity;
(c) refractive index;
(d) impedance to TEM waves.

If the light has an electric field strength of 5 kV/m, determine the magnetic field strength, and the average power per unit area.

(a) We can find the attenuation and phase coefficients from equations (2.9) and (2.10). So

$$\alpha^2 - \beta^2 = -\omega^2\mu\epsilon$$

and

$$2\alpha\beta = \omega\sigma\mu$$

Thus

$$\begin{aligned} \alpha^2 - \beta^2 &= -\left[\frac{2\pi c}{\lambda}\right]^2 \mu_0\epsilon_0\epsilon_r \\ &= -5.5 \times 10^{14} \end{aligned}$$

and

$$2\alpha\beta = 1.2 \times 10^6$$

Hence

$$\alpha^2 - \left[\frac{1.2 \times 10^6}{2\alpha}\right]^2 = -5.5 \times 10^{14}$$

and so $\alpha \approx 0$ or imaginary. Thus the light experiences negligible attenuation. By following a similar procedure, we find that $\beta = 2.34 \times 10^7$ rad/s.

(b) The phase velocity is given by $v_p = \omega/\beta$. Thus

$$\begin{aligned} v_p &= \frac{3.14 \times 10^{15}}{2.34 \times 10^7} \\ &= 1.34 \times 10^8 \text{ m/s} \end{aligned}$$

(c) The refractive index is given by

$$n = \frac{c}{v_p} = 2.24$$

or

$$n = \sqrt{\epsilon_r} = 2.24$$

or

$$n = \frac{\beta}{k_0} = 2.24$$

(d) The impedance of the material is

$$\begin{aligned} Z &= \frac{Z_0}{n} \\ &= \frac{377}{2.24} \\ &= 168\ \Omega \end{aligned}$$

Now, the magnitude of the E field is 5 kV/m. As $Z = E/H$, we can write

$$\begin{aligned} H &= \frac{E}{Z} \\ &= \frac{5 \times 10^3}{168} \\ &= 29.8\ \text{A/m} \end{aligned}$$

We can find the average power per m^2 by using

$$S_{\text{av}} = \frac{1}{2}\ E \times H$$

Hence

$$S_{\text{av}} = 74.5\ \text{kW/m}^2$$

This represents a power of 1.86 W in a 25 mm^2 area.

2.1.3 Group velocity and material dispersion

As we have seen, the velocity of light in a dielectric depends upon the refractive index. However, because of the atomic interactions between the material and the optical signal, refractive index varies with wavelength and so any light consisting of several different wavelengths will be dispersed. (A familiar example of dispersion is the spectrum produced when white light passes through a glass prism.) To examine the effect of dispersion on an optical communication link, we will consider intensity, or amplitude, modulation of an optical signal.

If a light source of frequency ω_c is amplitude modulated by a single frequency, ω_m, then the electric field intensity, e_{AM}, at a certain point in space will be

$$\begin{aligned} e_{\mathrm{AM}} &= E_{x\mathrm{o}}(1 + m\cos\omega_{\mathrm{m}}t)\cos\omega_{\mathrm{c}}t \\ &= E_{x\mathrm{o}}\{\cos\omega_{\mathrm{c}}t + \frac{m}{2}\,[\cos(\omega_{\mathrm{c}} + \omega_{\mathrm{m}})t + \cos(\omega_{\mathrm{c}} - \omega_{\mathrm{m}})t]\} \qquad (2.21) \end{aligned}$$

where m is the depth of modulation. Thus there are three individual frequency components: the carrier signal, ω_{c}; the upper-side frequency, $\omega_{\mathrm{c}} + \omega_{\mathrm{m}}$; and the lower-side frequency $\omega_{\mathrm{c}} - \omega_{\mathrm{m}}$. As $\beta = 2\pi/\lambda$, each of these components will have their own value of β. So, the variation of e_{AM} with distance, z, can be written as

$$\begin{aligned} e_{\mathrm{AM}} &= E_{x\mathrm{o}}(\cos\omega_{\mathrm{c}}t - \beta z) + E_{x\mathrm{o}}\,\frac{m}{2}\cos[(\omega_{\mathrm{c}} + \delta\omega)t - (\beta + \delta\beta)z] \\ &\quad + E_{x\mathrm{o}}\,\frac{m}{2}\cos[(\omega_{\mathrm{c}} - \delta\omega)t - (\beta - \delta\beta)z] \qquad (2.22) \end{aligned}$$

where $\delta\omega$ has replaced ω_{m}, and we have assumed that the variation of β with ω is linear around ω_{c} (figure 2.4).

The last two terms in (2.22) – the two side frequencies – can be written as

$$\begin{aligned} &E_{x\mathrm{o}}\frac{m}{2}\cos[(\omega_{\mathrm{c}}t - \beta z) + (\delta\omega t - \delta\beta z)] + E_{x\mathrm{o}}\frac{m}{2}\cos[(\omega_{\mathrm{c}}t - \beta z) - (\delta\omega t - \delta\beta z)] \\ &\quad = E_{x\mathrm{o}}m\cos(\omega_{\mathrm{c}}t - \beta z)\cos(\delta\omega t - \delta\beta z) \qquad (2.23) \end{aligned}$$

and so we can describe the total wave by

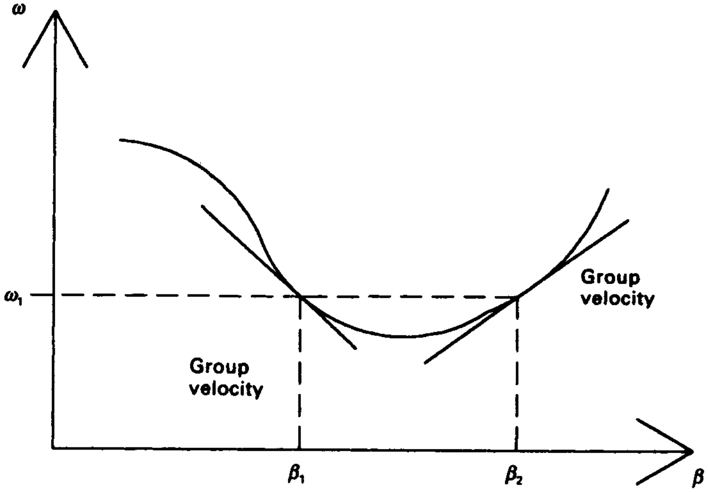


Figure 2.4 Showing the relationship between the phase and group velocity of a modulated TEM wave

$$E_{xo}(\cos\omega_c t - \beta z) + E_{xo}m\cos(\omega_c t - \beta z)\cos(\delta\omega t - \delta\beta z) \qquad (2.24)$$

Examination of this equation shows that the first term, the carrier wave, propagates at the familiar phase velocity. However, the second term, the modulation envelope, travels at a velocity of $\delta\omega/\delta\beta$ known as the *group velocity*, v_g. Thus we can write,

$$v_p = \frac{\omega}{\beta} \qquad (2.25) \qquad \text{and} \qquad v_g = \frac{d\omega}{d\beta} \qquad (2.26)$$

From these equations it should be evident that v_g is the gradient of a graph of ω against β, as in figure 2.4.

Examination of figure 2.4 shows that the group velocity is dependent on frequency. Thus different frequency components in a signal will travel at different group velocities, and so will arrive at their destination at different times. For digital modulation of the carrier, this results in smearing, or *dispersion*, of the pulses, which affects the maximum rate of modulation. The variation of refractive index with frequency is dependent on the glass material, and so this form of dispersion is known as *material dispersion.*

To observe the effect of material dispersion, let us derive the difference in propagation times, $\delta\tau$, for the two sidebands previously considered. We can express $\delta\tau$ as

$$\delta\tau = \frac{d\tau}{d\lambda} \times \delta\lambda \qquad (2.27)$$

where $\delta\lambda$ is the wavelength difference between the lower and upper sideband, and $d\tau/d\lambda$ is known as the *material dispersion coefficient*, D_{mat}. If we consider a *unit length*, then $\tau = 1/v_g$, and so

$$D_{mat} = \frac{d\tau}{d\lambda} = \frac{d}{d\lambda} \times \frac{1}{v_g} \qquad (2.28)$$

Now

$$\begin{aligned}\frac{1}{v_g} &= \frac{d\beta}{d\omega} = \frac{d\lambda}{d\omega} \times \frac{d\beta}{d\lambda} = -\frac{\lambda_0^2}{2\pi c} \times \frac{d\beta}{d\lambda} = -\frac{\lambda_0^2}{2\pi c} \times \frac{d}{d\lambda}k_0 n \\ &= -\frac{\lambda_0^2}{2\pi c} \times \frac{d}{d\lambda}\left(\frac{2\pi n}{\lambda_0}\right) = -\frac{\lambda_0^2}{c} \times \frac{d}{d\lambda}\left(\frac{n}{\lambda_0}\right) \\ &= \frac{1}{c}\left[n - \lambda_0\frac{dn}{d\lambda}\right] = \frac{N_g}{c} \qquad (2.29)\end{aligned}$$

where N_g is the *group refractive index* – compare with the definition of refractive index given by equation (2.15). So

$$\begin{aligned} D_{\text{mat}} &= \frac{d\tau}{d\lambda} = \frac{d}{d\lambda} \times \frac{N_g}{c} = \frac{1}{c}\left[\frac{dn}{d\lambda} - \frac{\lambda_0 d^2n}{d\lambda^2} - \frac{dn}{d\lambda}\right] \\ &= -\frac{\lambda_0}{c} \times \frac{d^2n}{d\lambda^2} \end{aligned} \tag{2.30}$$

(The negative sign shows that the upper sideband signal, the lowest wavelength, arrives before the lower sideband, the highest wavelength.) The units of D_{mat} are normally ns/nm/km (remember that we are considering a unit length of material). So, in order to find the dispersion in ns, we need to multiply D_{mat} by the wavelength difference between the minimum and maximum spectral components, and the length of the optical link. As the link length is variable, the material dispersion is usually expressed in units of time per unit length – symbol σ_{mat}.

Before we consider an example, it is worth noting that if the group velocity is the same as the phase velocity, as with air, then the material will be dispersionless. We should also note that, when we consider the planar waveguide, we should resolve the material dispersion along the horizontal axis. However, with multimode waveguides, modal dispersion is generally more significant than the material dispersion.

Example

A 100 MHz sinewave causes amplitude modulation of a 600 nm wavelength light source. The resultant light propagates through a dispersive medium with D_{mat} = 50 ps/nm/km. Determine the material dispersion.

As the light source is amplitude modulated, the modulated light will consist of the carrier wave and two sidebands spaced 100 MHz either side of the carrier. Thus the spread in wavelength is

$$\delta\lambda = 2.4 \times 10^{-4} \text{ nm}$$

and so the material dispersion is

$$\sigma_{\text{mat}} = 0.012 \text{ ps/km}$$

This calculation assumes that the source is monochromatic, that is the source generates light at a frequency 5×10^{14} Hz and *no other frequencies*. In practice, most light sources generate a range of wavelengths with the spread being known as the *linewidth* of the source. If we take a linewidth of 10 nm (that is, the light consists of a range of frequencies from 4.96 × 10^{14} Hz to 5.04 × 10^{14} Hz) we find that

$$\delta\lambda \approx 10 \text{ nm}$$

and so

$$\sigma_{\text{mat}} = 500 \text{ ps/km}$$

which is far higher than the dispersion introduced by the modulating signal alone.

From these calculations we can see that the spectral purity of the source can dominate σ_{mat} if the source linewidth is large. So, for high data-rate or long-haul applications, it is important to use narrow linewidth sources (dealt with in the next chapter).

Figure 2.5 shows the variation of D_{mat} with λ for three typical *glass* fibres. Because D_{mat} passes through zero at wavelengths around 1.3 μm, which happens to coincide with one of the transmission windows, this was the most popular wavelength for long-haul links. This situation is now changing with the introduction of *dispersion shifted* fibres, dealt with in section 2.3.4, in which

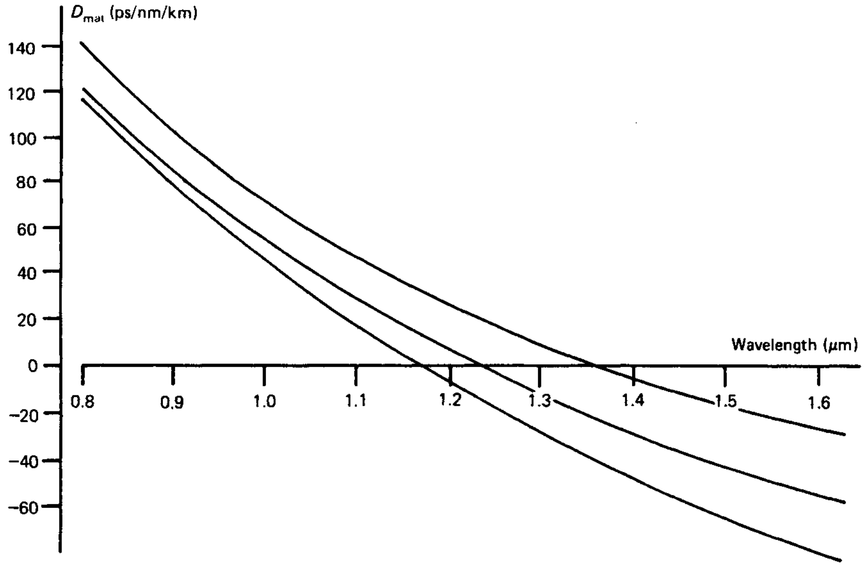


Figure 2.5 Variation of material dispersion, D_{mat}, with wavelength for three different glass fibres

the zero dispersion point is at 1.55 μm – a transmission window which offers lower attenuation.

We have seen that the composition of the fibre causes dispersion of the signal due to the variation of group refractive index with wavelength. There are, however, two further forms of dispersion – *modal* and *waveguide* – and we can examine these by considering propagation in a dielectric slab waveguide.

2.2 Propagation in a planar dielectric waveguide

In this section we shall consider propagation in a simple planar optical waveguide. In particular, we shall examine reflection and refraction of a light wave at the waveguide boundaries. This will lead to the conditions we must satisfy before successful propagation can occur, and introduce us to modal and waveguide dispersion. Although the values of dispersion we will calculate will seem large, we should remember that *planar* optical waveguides are generally quite short in length, and so dispersion effects are usually insignificant. In spite of this, the work presented here will be useful when we consider optical fibre.

2.2.1 Reflection and refraction at boundaries

Figure 2.6 shows a transverse electric, *TE*, wave, $\boldsymbol{E}_\mathrm{i}$ and $\boldsymbol{H}_\mathrm{i}$, incident on a boundary between two dissimilar, non-conducting, dielectrics (the waveguide boundary). As can be seen, some of the wave undergoes reflection, $\boldsymbol{E}_\mathrm{r}$ and $\boldsymbol{H}_\mathrm{r}$, while the rest is transmitted (or *refracted*), $\boldsymbol{E}_\mathrm{t}$ and $\boldsymbol{H}_\mathrm{t}$. In order to determine the optical power in both waves, we can resolve the waves into their x, y, and z components, and then apply the boundary conditions.

Let us initially consider the E field as it crosses the boundary. We can express the incident field as the combination of a field propagating in the negative x-direction, and another field travelling in the negative y-direction. Thus the propagation constant *associated with the propagating mode*, *k*, will be given by

$$k^2 = \beta_x^{\,2} + \beta_y^{\,2}$$

and so $\boldsymbol{E}_\mathrm{i}$ can be written as

$$\boldsymbol{E}_\mathrm{i} = \boldsymbol{a}_z E_0 \exp(\mathrm{j}\beta_1[x\sin\theta_\mathrm{i} + \mathrm{y}\cos\theta_\mathrm{i}]) \tag{2.31}$$

Similarly, we can write the reflected and transmitted fields as

$$\boldsymbol{E}_\mathrm{r} = \boldsymbol{a}_z E_\mathrm{r} \exp(\mathrm{j}\beta_1[x\sin\theta_\mathrm{r} + y\cos\theta_\mathrm{r}]) \tag{2.32}$$

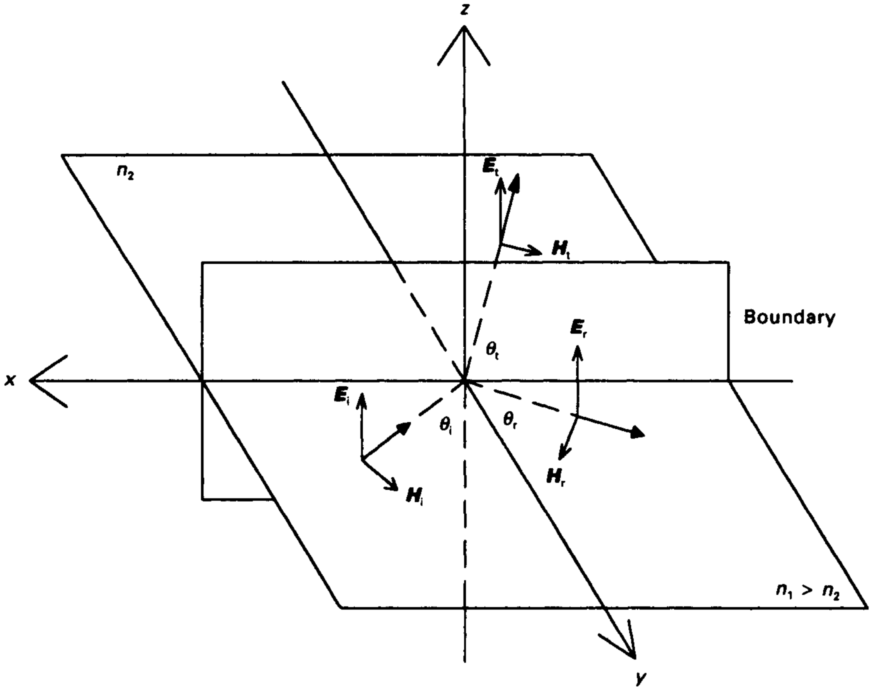


Figure 2.6 Reflection and refraction of a TEM wave, at the boundary between two dielectric materials

and

$$\boldsymbol{E}_t = \boldsymbol{a}_z E_t \exp(j\beta_2[x\sin\theta_t + y\cos\theta_t]) \tag{2.33}$$

(Here x and y refer to the *distances travelled along the respective axes.* This explains the absence of any negative signs in these equations.)

Now, the boundary conditions at the interface require the tangential components of the E and H fields in both media to be continuous. If we initially consider the continuity of the E field then, as these fields are already parallel to the interface, we can write

$$E_i + E_r = E_t \tag{2.34}$$

Dividing by E_i yields

$$1 + r_e = t_e \tag{2.35}$$

where r_e is the *reflection coefficient*, $r_e = E_r/E_i$, and t_e is the *transmission coefficient*, $t_e = E_t/E_i$. Thus we can write E_r and E_t as

$$\boldsymbol{E}_r = \boldsymbol{a}_z r_e E_0 \exp(j\beta_1[x\sin\theta_r + y\cos\theta_r]) \tag{2.36}$$

and

$$\boldsymbol{E}_t = \boldsymbol{a}_z t_e E_0 \exp(j\beta_2[x\sin\theta_t + y\cos\theta_t]) \tag{2.37}$$

We can substitute the expressions for the E fields at the interface, $y = 0$, back into (2.34) to give

$$E_0\exp(\,j\beta_1 x\sin\theta_i) + r_e E_0\exp(j\beta_1 x\sin\theta_r) = t_e E_0\exp(j\beta_2 x\sin\theta_t) \tag{2.38}$$

In order to satisfy (2.35), the exponential terms in (2.38) must be equal to each other, that is

$$\beta_1\sin\theta_i = \beta_1\sin\theta_r = \beta_2\sin\theta_t$$

The first of these equalities yields

$$\theta_i = \theta_r \tag{2.39}$$

which is *Snell's law of reflection*, that is, the angle of reflection equals the angle of incidence. The second equality gives

$$\sin\theta_t = \frac{\beta_1}{\beta_2}\sin\theta_i = \frac{k_0 n_1}{k_0 n_2}\sin\theta_i = \frac{n_1}{n_2}\sin\theta_i \tag{2.40}$$

known as *Snell's law of refraction*, or simply Snell's Law. (These equations should be familiar from geometric optics.)

In order to find an expression for r_e, let us now consider the second boundary relation – the continuity of the tangential H field. As the H fields act at right angles to the directions of propagation, we can write

$$\boldsymbol{H}_i = (-\boldsymbol{a}_x\cos\theta_i + \boldsymbol{a}_y\sin\theta_i)\frac{E_i}{Z_1} \tag{2.41}$$

$$\boldsymbol{H}_r = (\boldsymbol{a}_x\cos\theta_r + \boldsymbol{a}_y\sin\theta_r)\frac{E_r}{Z_1} \tag{2.42}$$

$$\boldsymbol{H}_t = (-\boldsymbol{a}_x\cos\theta_t + \boldsymbol{a}_y\sin\theta_t)\frac{E_t}{Z_2} \tag{2.43}$$

Now, the tangential H field boundary relation gives, at $y = 0$

$$\frac{E_0}{Z_1}\cos\theta_i\exp(j\beta_{1x}x) - r_e\frac{E_0}{Z_1}\cos\theta_r\exp(j\beta_{1x}x)$$

$$= t_e\frac{E_0}{Z_2}\cos\theta_t\exp(j\beta_{2x}x) \tag{2.44}$$

where we have substituted for E_r and E_t. The new parameters, β_{1x} and β_{2x}, are the phase constants for media 1 and 2 *resolved onto the x-axis*, defined by

$$\beta_{1x} = \beta_1\sin\theta_i = \beta_1\sin\theta_r \tag{2.45}$$

and

$$\beta_{2x} = \beta_2\sin\theta_t \tag{2.46}$$

As we have seen from (2.38), the exponential terms in (2.44) are all equal. Therefore (2.44) becomes

$$\frac{\cos\theta_i}{Z_1}(1 - r_e) = \frac{t_e\cos\theta_t}{Z_2} \tag{2.47}$$

Since $1 + r_e = t_e$, we can eliminate t_e from (2.47) to give

$$r_e = \frac{Z_2\cos\theta_i - Z_1\cos\theta_t}{Z_2\cos\theta_i + Z_1\cos\theta_t} = \frac{n_1\cos\theta_i - n_2\cos\theta_t}{n_1\cos\theta_i + n_2\cos\theta_t} \tag{2.48}$$

and, by using Snell's law, we can eliminate θ_t from (2.48) to give

$$r_e = \frac{\cos\theta_i - \sqrt{(n_2/n_1)^2 - \sin^2\theta_i}}{\cos\theta_i + \sqrt{(n_2/n_1)^2 - \sin^2\theta_i}} \tag{2.49}$$

Close examination of (2.49) reveals that r_e is unity if the term under the square root is zero, that is $\sin^2\theta_i = (n_2/n_1)^2$. Under these conditions, the reflected E field will equal the incident E field, and this is *total internal reflection*. The angle of incidence at which this occurs is the *critical angle*, θ_c, given by

$$\sin^2\theta_c = \left[\frac{n_2}{n_1}\right]^2 \text{ or } \sin\theta_c = \frac{n_2}{n_1} \tag{2.50}$$

Substitution of this result into Snell's Law gives the angle of refraction to be 90°, and so a transmitted ray travels along the interface. If the angle of incidence is greater than θ_c (that is, $\sin\theta_i > n_2/n_1$) then r_e will be complex, but $|r_e|$ will be unity and total internal reflection still takes place. However, there will also be a transmitted wave. In order to study this in greater detail, let us consider the expression for the transmitted E field, reproduced here as (2.51)

$$\boldsymbol{E}_t = \boldsymbol{a}_z t_e E_0 \exp(\mathrm{j}\beta_{2x}x + \mathrm{j}\beta_{2y}y) \tag{2.51}$$

where β_{2y} is the phase constant in medium 2 resolved onto the y-axis.

To evaluate $\boldsymbol{E}_t$ we need to find β_{2x} and β_{2y} or, by implication, $\sin\theta_t$ and $\cos\theta_t$. If the incident ray hits the boundary at an angle greater than the critical angle, that is $\theta_i > \theta_c$, then $\sin\theta_i > n_2/n_1$. If we substitute this into Snell's law, we find that $\sin\theta_t > 1$, which is physically impossible. We could work with hyperbolic functions at this point, but if we let $\sin\theta_t > 1$, then $\cos\theta_t$ will be imaginary, that is

$$\cos\theta_t = \sqrt{1-\sin^2\theta_t} = \mathrm{j}A \tag{2.52}$$

where A is a *real* number given by

$$A = \sqrt{(n_1/n_2)^2\sin^2\theta_i - 1} \tag{2.53}$$

Thus the transmitted wave can be written as

$$\begin{aligned}\boldsymbol{E}_t &= \boldsymbol{a}_z t_e E_0 \exp(\mathrm{j}\beta_{2x}x + \mathrm{j}^2 A\beta_2 y) \\ &= \boldsymbol{a}_z t_e E_0 \exp(-\beta_2 A y)\exp(\mathrm{j}\beta_{2x}x)\end{aligned} \tag{2.54}$$

This equation shows that an E field exists in the lower refractive index material *even though* total internal reflection takes place. As equation (2.54) shows, this field propagates without loss in the negative x-direction, but undergoes attenuation as $\exp(-\beta_2 A y)$ along the y-axis, *at right angles to its direction of propagation.*

In order to find the transmitted power, we must also find the transmitted H field, previously given by equation (2.43):

$$\boldsymbol{H}_t = (-\boldsymbol{a}_x\cos\theta_t + \boldsymbol{a}_y\sin\theta_t)\frac{E_t}{Z_2}$$

where $E_t = t_e E_0 \exp(-\beta_2 A y)\exp(\mathrm{j}\beta_{2x}x)$. Substitution for $\cos\theta_t$ yields

$$\boldsymbol{H}_t = (-\boldsymbol{a}_x \mathrm{j}A + \boldsymbol{a}_y\sin\theta_t)\frac{E_t}{Z_2} \tag{2.55}$$

As this equation shows, there are two components to the transmitted H field: a component along the x-axis that has a 90° phase-shift, time-wise, with respect to the E field; and a component along the y-axis. As the transmitted H field has two components, the transmitted power will also have two components.

As we have already seen in section 2.1.2, the average power is given by

$$S_{\mathrm{av}} = \frac{1}{2} \boldsymbol{E} \times \boldsymbol{H}^*$$

where $\boldsymbol{H}^* = \boldsymbol{H}\exp\left(-\mathrm{j}[\omega t - \phi]\right)$, and ϕ is the temporal phase angle between the individual E and H field components (90° for the x-axis component of H_{t}). So, with the fields given by (2.54) and (2.55), we have

$$\begin{aligned} S_{\mathrm{av}} &= -\boldsymbol{a}_x \frac{1}{2} \times \frac{E_{\mathrm{t}}^2}{Z_2} \sin\theta_{\mathrm{t}} - \boldsymbol{a}_y \frac{1}{2} \times \frac{AE_{\mathrm{t}}^2}{Z_2} \exp(\mathrm{j}\pi/2) \\ &= -\boldsymbol{a}_x \frac{1}{2} \times \frac{E_{\mathrm{t}}^2}{Z_2} \sin\theta_{\mathrm{t}} - \boldsymbol{a}_y \mathrm{j} \frac{1}{2} \times \frac{AE_{\mathrm{t}}^2}{Z_2} \qquad (2.56) \end{aligned}$$

Thus it can be seen that the transmitted power has two components: an x-axis component (along the boundary) with the same properties as the transmitted E field; and an imaginary component at right angles to the interface. This imaginary component is due to the 90° temporal phase shift between the E and H fields. (A similar relationship occurs between the voltage and current in reactive circuits.) The physical interpretation of this is that for the first quarter cycle, the E field is positive and the H field is negative, giving negative power flow. In the next quarter cycle, the E field is still positive, but the H field is also positive, and this gives positive power flow. Thus power flows to and from the boundary four times per complete cycle of the E or H field, and so no *net* power flows across the boundary along the y-axis.

The real part of the Poynting vector, from (2.56), is

$$S_{\mathrm{av}} = -\boldsymbol{a}_x \frac{1}{2} \times \frac{E_{\mathrm{t}}^2}{Z_2} \sin\theta_{\mathrm{t}} \qquad (2.57)$$

which shows that power flows *along* the boundary. So, although total internal reflection takes place, there is still a TEM wave propagating along the boundary – the *evanescent wave*. As we shall see in the following example, this wave is very tightly bound to the interface between the two media.

(We should note that the evanescent wave is why a cladding surrounds the core of an optical fibre. If air surrounds the core, total internal reflection will still take place, but the air cladding will contain the evanescent wave.

Thus if we place another optical fibre close to the first fibre, the evanescent wave will cause coupling of power from one fibre to the other. This effect is desirable in optical couplers, but is clearly undesirable when the fibres are tightly bound in a cable. By surrounding the core with a cladding of similar material, the power flow in the core is protected, both from coupling with adjacent fibres and from environmental effects.)

Example

A TE wave, with a free-space wavelength of 600 nm, is propagating in a dielectric of refractive index 1.5. The wave hits a boundary with a second dielectric of refractive index 1.4, at an angle of 75° to a normal drawn perpendicular to the boundary. Determine the average transmitted power, and calculate the attenuation of the evanescent wave at a distance of one wavelength from the boundary.

Let us initially calculate the reflection coefficient, from which we can find the transmission coefficient. Now, r_e is given by, equation (2.49)

$$r_e = \frac{\cos\theta_i - \sqrt{(n_2/n_1)^2 - \sin^2\theta_i}}{\cos\theta_i + \sqrt{(n_2/n_1)^2 - \sin^2\theta_i}}$$

$$= \frac{0.26 - \sqrt{0.87 - 0.93}}{0.26 + \sqrt{0.87 - 0.93}}$$

$$= \frac{0.26 - \mathrm{j}0.25}{0.26 + \mathrm{j}0.25} = 1\underline{/-2\phi}$$

where $\phi = \tan^{-1}(0.25/0.26) = 0.75$ rad. (This angle is the spatial phase-change experienced by the E field on reflection.) Since $t_e = 1 + r_e$

$$t_e = \frac{2 \times 0.26}{0.26 + \mathrm{j}0.25} = 0.54\underline{/-\phi}$$

Thus, $\boldsymbol{E}_t$ (equation 2.54) will be

$$\boldsymbol{E}_t = \boldsymbol{a}_z 0.54 E_0 \exp(-3.93 \times 10^6 y)\exp(\mathrm{j}1.57 \times 10^7 x)\ \underline{/-\phi}$$

Hence the average transmitted power is

$$S_{av} = 5.4 \times 10^{-4} E_0^2 \exp(-7.86 \times 10^6 y)$$

We should note that, as the magnitude of r_e is unity, the reflected power is the same as the incident power. (The angle associated with r_e is simply the spatial phase shift experienced at the boundary.) Thus we can say that the evanescent wave couples with, but takes no power from, the light travelling in the dielectric. It can, however, deliver power and most SM couplers rely on this property.

If we consider the average power at y equals one wavelength in the second dielectric, we find

$$S_{av} = 5.4 \times 10^{-4} E_0^2 \exp(-3.38) \qquad \text{for } y = 430 \text{ nm}$$
$$= 18.4 \times 10^{-6} E_0^2$$

This power is 30 times less than that transmitted across the boundary – an attenuation of roughly 15 dB at a distance of one wavelength from the boundary. Clearly the evanescent wave is tightly bound to the interface between the two materials.

2.2.2 Propagation modes – ray-path analysis

In the previous section, we considered the reflection of a light ray at a single dielectric boundary. We showed that, provided the angle of incidence was greater than θ_c, total internal reflection would occur. It might be thought that any ray satisfying this requirement must propagate without loss. However, as we will shortly see, a light ray must satisfy certain conditions before it can successfully propagate.

Figure 2.7 shows the situation we will analyse. In this diagram, the E field is drawn at right angles to the ray-path. In order for the ray to propagate, the E field at A should be in phase with the E field at B, that is *the ray must constructively interfere with itself.* If this is not the case, the fields will destructively interfere with each other, and the ray will simply die out. So, in order to maintain constructive interference at point B, the change in

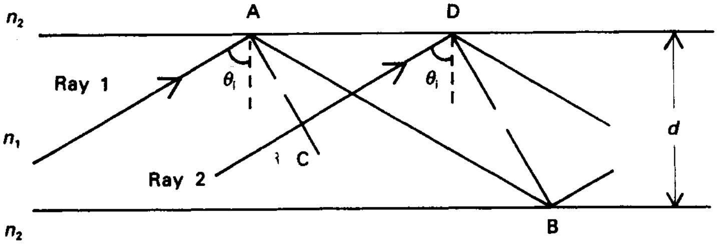


Figure 2.7 Illustrative of the requirement for successful propagation of two TEM waves in a planar optical waveguide (CD = a, AB = b)

phase that the ray undergoes as it travels from A to B must be an integral number of cycles.

In going from A to B, the ray crosses the waveguide twice, and is reflected off the boundary twice. We can find the phase change due to reflection off the boundary from the reflection coefficient. From (2.49), r_e is

$$r_e = \frac{\cos\theta_i - \sqrt{(n_2/n_1)^2 - \sin^2\theta_i}}{\cos\theta_i + \sqrt{(n_2/n_1)^2 - \sin^2\theta_i}}$$

or, from the example at the end of the previous section

$$r_e = 1\angle{-2\phi} \text{ where } \phi = \tan^{-1}\frac{\sqrt{(n_1{}^2\sin^2\theta_i - n_2{}^2)}}{n_1\cos\theta_i} \tag{2.58}$$

Therefore, successful propagation occurs provided

$$2 \times 2d \times \beta_{1y} + 2 \times 2\phi = 2\pi N \tag{2.59}$$

where β_{1y} is the phase coefficient resolved onto the y-axis and N is a positive integer, known as the *mode number*. (Although we have taken upward travelling rays in figure 2.7, downward travelling rays will result in identical equations. These two rays make up a single *waveguide mode*.)

Now, if we substitute for ϕ, (2.58) becomes

$$2d \times \beta_{1y} - 2 \times \tan^{-1}\frac{\sqrt{(n_1{}^2\sin^2\theta_i - n_2{}^2)}}{n_1\cos\theta_i} = \pi N$$

or

$$\tan\left(\beta_{1y}d - \frac{\pi}{2}N\right) = \frac{\sqrt{n_1{}^2\sin^2\theta_i - n_2{}^2}}{n_1\cos\theta_i} \tag{2.60}$$

Now, $\beta_{1y} = \beta_1\cos\theta_i$ and $\beta_1 = k_0n_1$, and so we can write (2.60) as

$$\tan\left(\beta_{1y}d - \frac{\pi}{2}N\right) = \frac{2\pi\sqrt{n_1{}^2\sin^2\theta_i - n_2{}^2}}{\beta_{1y}\lambda_0} \tag{2.61}$$

From our previous discussions, the evanescent field undergoes attenuation as $\exp(-\alpha_2 y)$ and, from (2.54), we can write the attenuation factor as

$$\begin{aligned}\alpha_2 &= \beta_2 A \\ &= \beta_2\sqrt{(n_1/n_2)^2\sin^2\theta_i - 1}\end{aligned}$$

or

$$\alpha_2 = \frac{2\pi\sqrt{n_1^2\sin^2\theta_i - n_2^2}}{\lambda_0} \tag{2.62}$$

Therefore we can write (2.61) as

$$\tan\left(\beta_{1y}d - \frac{\pi}{2}N\right) = \frac{\alpha_2}{\beta_{1y}} \tag{2.63}$$

Both (2.60) and (2.63) are known as *eigenvalue* equations. Solution of (2.63) will yield the values of β_{1y}, the *eigenvalues*, for which light rays will propagate, while solution of (2.60) will yield the permitted values of θ_i. Unfortunately we can only solve these equations using graphical or numerical methods as the following example shows.

Example

Light of wavelength 1.3 μm is propagating in a planar waveguide of width 200 μm, depth 10 μm and refractive index 1.46, surrounded by material of refractive index 1.44. Find the permitted angles of incidence.

We can find the angle of incidence for each propagating mode by substituting these parameters into (2.60), and then solving the equation by graphical means. This is shown in figure 2.8, which is a plot of the left- and right-hand sides of (2.60), for varying angles of incidence, θ_i.

This graph shows that approximate values of θ_i are 88°, 86°, 84° and 82°, for mode numbers 0–3 respectively. Taking these values as a starting point, we can use numerical iteration to find the values of θ_i to any degree of accuracy. Thus the values of θ_i, to two decimal places, are 87.83°, 85.67°, 83.55° and 81.55°.

This analysis has shown that only those modes that satisfy the eigenvalue equation can propagate in the waveguide. We can estimate the number of modes by noting that the highest order mode will propagate at the lowest angle of incidence. As this angle will have to be greater than or equal to the critical angle, we can use θ_c in (2.60) to give

$$\tan\left[\frac{2\pi n_1}{\lambda_0}d\cos\theta_c - N_{\max}\frac{\pi}{2}\right] = 0$$

or

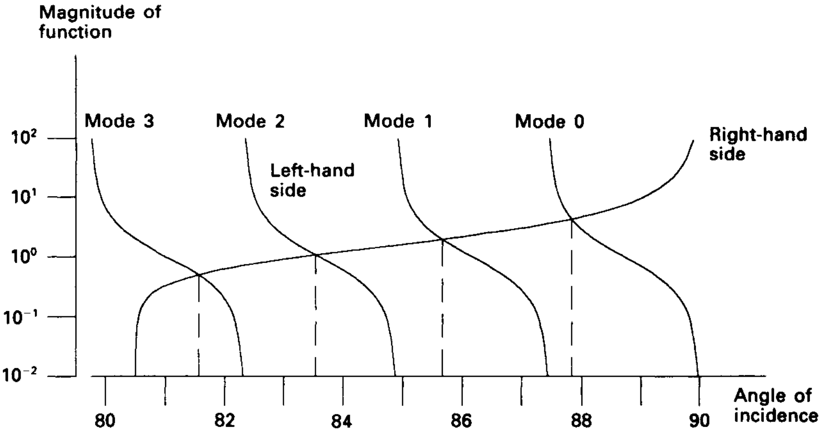


Figure 2.8 Eigenvalue graphs for a planar dielectric waveguide

$$\frac{2\pi d(n_1^2 - n_2^2)^{\frac{1}{2}}}{\lambda_0} = N_{\text{max}}\frac{\pi}{2}$$

If we define a normalised frequency variable, V, as

$$V = \frac{2\pi d(n_1^2 - n_2^2)^{\frac{1}{2}}}{\lambda_0} \tag{2.64}$$

then the maximum number of modes will be

$$N_{\text{max}} = \frac{2V}{\pi} = \frac{4d(n_1^2 - n_2^2)^{\frac{1}{2}}}{\lambda_0} \tag{2.65}$$

Equation (2.65) shows that the value of N_{max} is unlikely to be an integer, and so we must round it up to the nearest whole number. If we take the previous example, then $V = 5.82$ and so the number of propagating modes is 4. As can be seen from figure 2.8, there are only four solutions to the eigenvalue equation, so confirming the accuracy of (2.65). It should be noted that we can find the maximum propagating frequency, or wavelength, from V. We can also find the condition for single-mode operation from V. If $N_{\text{max}} = 1$, V must be $\pi/2$, and we can find the waveguide depth from (2.65).

In the next section we will apply Maxwell's equations to the planar dielectric waveguide. Although this will give us the same results as presented in this section, the treatment is more thorough, and it will help us when we come to consider propagation in optical fibre. (As the following analysis involves some complex mathematics, some readers may wish to neglect it on a first reading.)

2.2.3 Propagation modes – modal analysis

To examine propagation in a planar dielectric waveguide thoroughly, we need to solve Maxwell's equations. If we assume that the dielectric is ideal, $\sigma = 0$ and we can write Maxwell's equations as

$$\nabla \times \boldsymbol{E} = -\frac{\mu \partial \boldsymbol{H}}{\partial t} \qquad \text{(2.66a) and } \nabla \times \boldsymbol{H} = \frac{\epsilon \partial \boldsymbol{E}}{\partial t} \qquad \text{(2.66b)}$$

If we consider a transverse electric field propagating as in figure 2.9, we get

$$\nabla \times \boldsymbol{E} = \begin{vmatrix} \boldsymbol{a}_x & \boldsymbol{a}_y & \boldsymbol{a}_z \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ 0 & 0 & E_z \end{vmatrix} = \frac{\partial}{\partial y} E_z \boldsymbol{a}_x - \frac{\partial}{\partial x} E_z \boldsymbol{a}_y = -\mu\left(\frac{\partial}{\partial t} H_x \boldsymbol{a}_x + \frac{\partial}{\partial t} H_y \boldsymbol{a}_y\right)$$

and

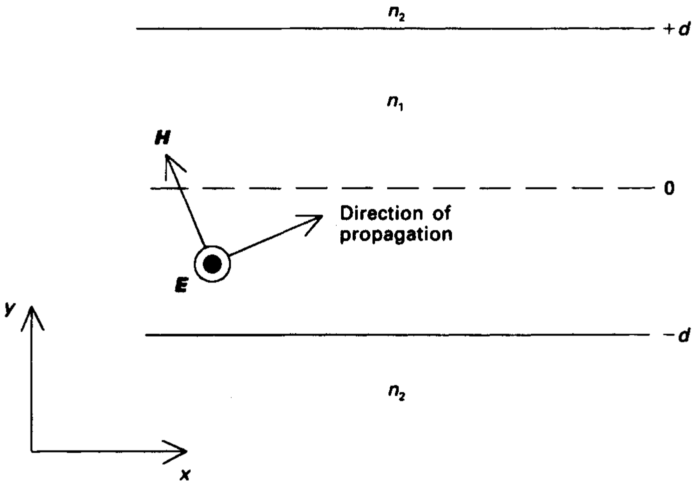


Figure 2.9 A transverse electric wave propagating in a planar dielectric waveguide

$$\nabla \times \boldsymbol{H} = \begin{vmatrix} \boldsymbol{a}_x & \boldsymbol{a}_y & \boldsymbol{a}_z \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ H_x & H_y & 0 \end{vmatrix} = \left(\frac{\partial}{\partial x}H_y - \frac{\partial}{\partial y}H_x\right)\boldsymbol{a}_z = \epsilon\frac{\partial}{\partial t}E_z\boldsymbol{a}_z$$

Now, the fields are propagating in the x direction and so we can write the *x component* of the E field as

$$\boldsymbol{E} = E\exp(\mathrm{j}\omega t)\exp(-\mathrm{j}\beta_x x)\boldsymbol{a}_z$$

with a similar expression for the H field. (Here β_x is the propagation coefficient of each individual mode along the x-axis.) Thus differentiation with respect to time results in multiplying the field strengths by jω, and differentiation with respect to x gives multiplication by $-\mathrm{j}\beta_x$. (A similar argument applies to the y components of the E and H fields.) With this in mind, and by equating the vector coefficients, Maxwell's equations result in the following

$$\frac{\partial}{\partial y}E_z = -\mathrm{j}\omega\mu H_x \tag{2.67a}$$

$$-\mathrm{j}\beta_x E_z = \mathrm{j}\omega\mu H_y \tag{2.67b}$$

$$-\mathrm{j}\beta_x H_y - \frac{\partial}{\partial y}H_x = \mathrm{j}\omega\epsilon E_z \tag{2.67c}$$

Rather than derive an expression for the y component of the E field, we can rearrange (2.67a) and (2.67b) to give H_x and H_y in terms of E_z, and substitute the results into (2.67c) to give

$$\frac{\mathrm{j}\beta_x^2 E_z}{\omega\mu} + \frac{1}{\mathrm{j}\omega\mu}\frac{\partial^2 E_z}{\partial y^2} = \mathrm{j}\omega\epsilon E_z$$

Hence we can write

$$\frac{\partial^2 E_z}{\partial y^2} = (\beta_x^2 - \omega^2\mu\epsilon)E_z \tag{2.68}$$

which is valid in both the slab and the surrounding material.

We should note that the E field, as described by (2.68), is independent of time. So, the solution to (2.68) describes the *stationary distribution* of the *tangential E* field in the *vertical y direction*. This distribution is independent of time because we are considering a waveguide, and reflections off the boundary walls results in a *standing wave* pattern that must satisfy (2.68).

Now, we have three regions of interest: the surrounding material below $y = -d$; the surrounding material above $y = d$; and the slab material between $-d$ and $+d$. If we initially consider the solution to (2.68) in the surrounding material below $y = -d$, the E field must decrease exponentially away from the boundary (it is an evanescent wave). So, a possible solution is

$$E_z = K\exp(\alpha_2 y) \qquad y < -d \tag{2.69a}$$

and, by using (2.67a)

$$H_x = \frac{-\alpha_2}{\mathrm{j}\omega\mu} K\exp(\alpha_2 y) \qquad y < -d \tag{2.69b}$$

where K is a constant, and α_2 is the attenuation constant given by

$$\alpha_2^2 = \beta_x^2 - n_2^2 k_0^2 \tag{2.70}$$

(We can derive this equation by substituting equation 2.69a into equation 2.68.)

If we take $y > d$, we must have an exponential decrease in E_z with increasing y. So, another solution to (2.68) is

$$E_z = L\exp(-\alpha_2 y) \qquad y > d \tag{2.71a}$$

and so

$$H_x = \frac{+\alpha_2}{\mathrm{j}\omega\mu} L\exp(-\alpha_2 y) \qquad y > d \tag{2.71b}$$

where L is another constant. If we consider the slab material, (2.68) becomes

$$\frac{\partial^2 E_z}{\partial y^2} = (\beta_{1x}^2 - n_1^2 k_0^2) E_z$$

with a possible solution given by

$$E_z = M\sin(\beta_{1y} y + \phi) \qquad -d < y < +d \tag{2.72a}$$

where M is a constant, β_{1y} is the phase coefficient in the slab material *resolved along the y-axis*, and ϕ is a spatial phase shift that we have yet to determine. Thus H_x will be given by

$$H_x = \frac{M\beta_{1y}}{j\omega\mu}\cos(\beta_{1y}y + \phi) \qquad -d < y < +d \qquad (2.72b)$$

where

$$\beta_{1y}{}^2 = n_1{}^2k_0^2 - \beta_{1x}{}^2 \qquad (2.73)$$

(This equation can be derived by substituting equation 2.72a into 2.68. We should also note that β_{1y}, β_{1x} and n_1k_0 form a right-angled triangle. The proof of this is left as an exercise for the reader.)

So, we have six equations that describe the E_z and H_x fields in the slab and surrounding material. All these equations have a number of constants that we can find using the boundary relations for the E and H fields. If we initially consider the lower interface, we have, from (2.69a) and (2.72b)

$$K\exp(-\alpha_2 d) = M\sin(-\beta_{1y}d + \phi) \qquad \text{at } y = -d$$

for the continuity of the E_z field. After some simple manipulation, we get

$$K = M\sin(-\beta_{1y}d + \phi)\exp(\alpha_2 d)$$

The continuity of the H_x field gives, using (2.69b) and (2.72b)

$$\frac{-\alpha_2}{j\omega\mu}K\exp(-\alpha_2 d) = -\frac{M\beta_{1y}}{j\omega\mu}\cos\,(-\beta_{1y}d + \phi)$$

and so

$$K = \frac{M\beta_{1y}}{\alpha_2}\cos(-\beta_{1y}d - \phi)\exp(\alpha_2 d)$$

If we equate the two equations for K, we get

$$M\sin(-\beta_{1y}d + \phi)\exp(\alpha_2 d) = \frac{M\beta_{1y}}{\alpha_2}\cos(-\beta_{1y}d + \phi)\exp(\alpha_2 d)$$

After some rearranging we get

$$\tan(-\beta_{1y} + \phi) = \frac{\beta_{1y}}{\alpha_2}$$

or

$$-\beta_{1y}d + \phi = \tan^{-1}(\beta_{1y}/\alpha_2) + m'\pi$$

where m' is an integer. Thus the phase angle ϕ is given by

$$\phi = \tan^{-1}(\beta_{1y}/\alpha_2) - \beta_{1y}d + m'\pi \tag{2.74}$$

Now, if we consider the upper interface, $y = +d$, we have

$$L\exp(-\alpha_2 d) = M\sin(\beta_{1y}d + \phi)$$

for the continuity of the E_z field. After some rearranging we get

$$L = M\sin(\beta_{1y}d + \phi)\exp(\alpha_2 d)$$

The continuity of the H_x fields gives

$$\frac{+\alpha_2}{j\omega\mu}L\exp(-\alpha_2 d) = -\frac{M\beta_{1y}}{j\omega\mu}\cos(\beta_{1y}y + \phi)$$

and so

$$L = -\frac{M\beta_{1y}}{\alpha_2}\cos(\beta_{1y}y + \phi)\exp(\alpha_2 d)$$

By equating the two values for L we get

$$M\sin(\beta_{1y}d + \phi)\exp(\alpha_2 d) = -\frac{M\beta_{1y}}{\alpha_2}\cos(\beta_{1y}y + \phi)\exp(\alpha_2 d)$$

Hence

$$\tan(\beta_{1y}d + \phi) = -\frac{\beta_{1y}}{\alpha_2}$$

and so another expression for ϕ is

$$\phi = -\tan^{-1}(\beta_{1y}/\alpha_2) - \beta_{1y}d + m''\pi \tag{2.75}$$

where m'' is another integer.

If we equate (2.74) and (2.75), we get

$$\tan^{-1}(\beta_{1y}/\alpha_2) + \beta_{1y}d + m'\pi = -\tan^{-1}(\beta_{1y}/\alpha_2) - \beta_{1y}d + m''\pi$$

or

$$2\tan^{-1}(\beta_{1y}/\alpha_2) = -2\beta_{1y}d + m''\pi - m'\pi \tag{2.76}$$

We should note that m' and m'' are both integers that cover the range 0 to ∞, and so we can arbitrarily set m' to zero. Now

$$\tan^{-1}(\beta_{1y}/\alpha_2) = \frac{\pi}{2} - \tan^{-1}(\alpha_2/\beta_{1y})$$

and so (2.76) becomes

$$\frac{\pi}{2} - \tan^{-1}(\alpha_2/\beta_{1y}) = -\beta_{1y}d + \frac{m''\pi}{2}$$

which can be written as

$$\tan\left(\beta_{1y}d - N\frac{\pi}{2}\right) = -\frac{\alpha_2}{\beta_{1y}} \tag{2.77}$$

where N is an integer. The solution of this equation yields the values of β_{1y} for which light rays will propagate. Comparison with equation (2.63) shows that this method, based on Maxwell's equations, results in an equation that is identical to that derived using the ray-model. (Although this derivation is more complicated than the ray-path model, it has introduced us to several important parameters as we shall see in the following sections.)

Let us now examine the cut-off condition for the planar waveguide. The maximum angle of incidence for any particular mode is the critical angle θ_c. If we can satisfy this condition, we find from our ray-path analysis that α_2 (equation 2.62) is zero. Thus the E_z component is not attenuated as it passes through the surrounding material, and the wave is not closely bound to the slab. If we take $\alpha_2 = 0$, we find from (2.70) that $\beta_x = \beta_2$, and so the phase coefficient of the propagating mode is identical to that of the surrounding material. This is known as the *cut-off condition*, and it represents the minimum value of β for which any mode will propagate. If the waveguide is operating well away from the cut-off condition, the propagating modes will be tightly bound to the slab, and so we can intuitively reason that $\beta_x = \beta_1$. Thus we can see that each propagating mode must have $\beta_2 < \beta_x < \beta_1$. We can define a binding parameter, b, as

$$b = \frac{\beta_x^2 - n_2^2k_0^2}{n_1^2k_0^2 - n_2^2k_0^2} \tag{2.78}$$

and so, with $n_2k_0 < \beta_x < n_1k_0$, the range of b will be 0 for loosely bound modes, to 1 for tightly bound modes.

[An alternative way of viewing this binding parameter is to let each mode have an *effective refractive index*, n_{eff}. Thus (2.78) will become

$$b = \frac{n_{\text{eff}}^2 - n_2^2}{n_1^2 - n_2^2}$$

Re-casting (2.78) in this form can be useful in that it shows that the velocity of the propagating mode is bounded by the velocity in the core, for tightly bound modes, and the velocity in the cladding material, for loosely bound modes. We will use the term effective refractive index very shortly.]

We can find the cut-off wavelength of any guide by substituting $\alpha_2 = 0$ into (2.77) to give

$$\tan\left(\beta_{1y}d - N\frac{\pi}{2}\right) = 0$$

which implies

$$\beta_{1y}d = N\frac{\pi}{2} \tag{2.79}$$

As $\beta_{1y} = \beta_1\cos\theta_c$, $\cos^2\theta_c = 1 - \sin^2\theta_c$, and $\sin\theta_c = n_2/n_1$, we can write

$$\beta_1 d\left(1 - \frac{n_2^2}{n_1^2}\right)^{\frac{1}{2}} = N\frac{\pi}{2}$$

or

$$\frac{2\pi n_1 d}{\lambda_{co}}\left(\frac{n_1^2 - n_2^2}{n_1^2}\right)^{\frac{1}{2}} = N\frac{\pi}{2}$$

Thus the cut-off wavelength for a particular mode is given by

$$\lambda_{co} = \frac{4d}{N}\left(n_1^2 - n_2^2\right)^{\frac{1}{2}} \tag{2.80}$$

If we consider the lowest order mode ($N = 0$), we have a λ_{co} of infinity, and so there is, theoretically, no cut-off wavelength for the lowest order mode. If we take $N = 1$, we can determine the waveguide depth that just results in the first-order mode being cut-off.

Example

Light of wavelength 1.3 μm is propagating in a planar waveguide of width 200 μm and refractive index 1.46, surrounded by material of refractive index 1.44. Determine the waveguide depth for single-mode operation.

Let us consider the $N = 1$ mode, and allow this mode to be just cut-off. Now, the cut-off condition for this mode is

$$1.3 \times 10^{-6} = 4d(1.46^2 - 1.44^2)^{\frac{1}{2}}$$

and so the maximum waveguide depth is 2.7 µm. Thus single-mode planar waveguides are clearly very small devices!

Before we finish this particular section, let us return to the cut-off condition for any particular mode (2.79):

$$\beta_{1y}d = N\frac{\pi}{2}$$

This can be written as

$$V = N\frac{\pi}{2} \tag{2.81}$$

where

$$V = \frac{2\pi d(n_1{}^2 - n_2{}^2)^{\frac{1}{2}}}{\lambda_0} \tag{2.82}$$

is known as the *V value* of the waveguide. Equations (2.81) and (2.82) are identical to those we derived using the ray-path analysis (equations 2.64 and 2.65). We can also express V as

$$V^2 = (\alpha_2 d)^2 + (\beta_{1y}d)^2 \tag{2.83}$$

This can be easily proved by noting that

$$\alpha_2{}^2 = \beta_x{}^2 - n_2{}^2k_0{}^2$$

and

$$\beta_{1y}{}^2 = n_1{}^2k_0{}^2 - \beta_x{}^2$$

Thus

$$\begin{aligned} V^2 &= d^2(\beta^2 - \beta_2{}^2 + \beta_1{}^2 - \beta^2) \\ &= d^2(n_1{}^2k_0{}^2 - n_2{}^2k_0{}^2) \end{aligned}$$

and so

$$V = \frac{2\pi d(n_1{}^2 - n_2{}^2)^{\frac{1}{2}}}{\lambda_0}$$

which is identical to (2.64).

In this section we have applied Maxwell's equations to a planar dielectric waveguide. We finished by showing that we can reduce the number of propagating modes by decreasing the waveguide thickness. In particular, if the V value of the waveguide is less than $\pi/2$, then only the zero-order mode can propagate (so-called monomode or *single-mode*, SM, operation). As we shall see in the next section, single-mode operation helps to reduce the total dispersion.

2.2.4 Modal dispersion – ray-path analysis

We have already seen that only a certain number of modes can propagate. Each of these modes carries the modulation signal and, as each one is incident on the boundary at a different angle, they will each have their own individual propagation times. In a digital system, the net effect is to smear out the pulses, and so this is a form of dispersion – *modal dispersion.*

The difference in arrival time, δt, between the fastest and slowest modes will be given by

$$\delta t = t_{\max} - t_{\min} \tag{2.84}$$

where $t_{\max}$ and $t_{\min}$ are the propagation times of the highest and lowest order modes respectively. As it is the modulation envelope that carries the information, we can find $t_{\max}$ and $t_{\min}$ by dividing the waveguide length by the *axial* components of the group velocities. As this requires knowledge of θ_i for the various waveguide modes, it may not be a practical way of estimating δt.

We can obtain an indication of the dispersion by approximating the angle of incidence for the highest order mode to θ_c, and that of the zero-order mode to 90°. Thus

$$t_{\min} = \frac{LN_{g1}}{c} \quad \text{and} \quad t_{\max} = \frac{LN_{g1}}{c\sin\theta_c} = \frac{LN_{g1}{}^2}{cN_{g2}}$$

where L is the length of the waveguide, and we have used the group refractive indices N_{g1} and N_{g2}. Therefore δt will be

$$\delta t = \frac{LN_{g1}}{cN_{g2}}(N_{g1} - N_{g2}) \tag{2.85}$$

Now, if we take $N_{g1}/n_1 \approx N_{g2}/n_2$, then the dispersion *per unit length* will be

$$\frac{\delta t}{L} = \frac{N_{g1}}{cN_{g2}}(N_{g1} - N_{g2}) = \frac{N_{g1}}{cn_2}(n_1 - n_2)$$
$$= \frac{N_{g1}\delta n}{c}$$

or

$$\sigma_{\mathrm{mod}} = \frac{N_{g1}}{c} \times \delta n \qquad (2.86)$$

where δn is the *fractional refractive index difference*, and σ_{mod} is the dispersion per unit length. We should note that each individual mode will also suffer from material dispersion. Thus, when we resolve the transit time of each mode on to the fibre axis, we should also take into account the material dispersion. Fortunately, in most multimode waveguides, the material dispersion is far less than the modal dispersion, and so we can generally ignore its effects.

Example

Light of wavelength 850 nm is propagating in a waveguide of 10 μm depth, with refractive indices n_1 = 1.5 and n_2 = 1.4, and group refractive indices N_{g1} = 1.64 and N_{g2} = 1.53. Determine the modal dispersion.

If we use equation (2.86), we find

$$\sigma_{\mathrm{mod}} = \frac{1.64}{3 \times 10^8} \times \frac{1.5 - 1.4}{1.4}$$
$$= 0.39 \text{ ns/km}$$

If we use modal analysis, we find that there are thirteen modes, and the angles of incidence for the zero and thirteenth order modes are 89.21° and 69.96° respectively. This method results in σ_{mod} = 0.35 ns/km, and so we can see that the error in using (2.86) is small. If the waveguide is single-mode, then σ_{mod} reduces to zero.

2.2.5 Modal dispersion – modal analysis

The analysis just presented used the ray-path approximation to determine the modal dispersion. Although the error in using this approximation is very small, a more rigorous treatment using mode theory will aid us when we consider optical fibre. To help us in our analysis, we will use the binding parameter, b, previously defined by (2.78):

$$b = \frac{\beta^2 - n_2^2 k_0^2}{n_1^2 k_0^2 - n_2^2 k_0^2} \tag{2.87}$$

We can rearrange (2.87) to give

$$\begin{aligned} \beta &= [n_2^2 k_0^2 + b(n_1^2 k_0^2 - n_2^2 k_0^2)]^{\frac{1}{2}} \\ &= \frac{2\pi n_2}{\lambda_0}\left[1 + b\left(\frac{n_1^2 - n_2^2}{n_2^2}\right)\right]^{\frac{1}{2}} \end{aligned} \tag{2.88}$$

Most optical waveguides are fabricated with $n_1 \approx n_2$ (a condition known as *weakly guiding* [4]) and so (2.88) becomes

$$\begin{aligned} \beta &= \frac{2\pi n_2}{\lambda_0}(1 + 2b\delta n)^{\frac{1}{2}} \\ &= \frac{\omega n_2}{c}(1 + 2b\delta n)^{\frac{1}{2}} \\ &= \frac{\omega n_2}{c}(1 + b\delta n) \end{aligned} \tag{2.89}$$

where we have used the binomial expansion. Now, modes travel at the group velocity, v_g, and so the time taken for a mode to travel a unit length, τ, is given by

$$\tau = \frac{1}{v_g} = \frac{d\beta}{d\omega} = (1 + b\delta n)\frac{d}{d\omega}\left[\frac{\omega n_2}{c}\right] + \left[\frac{\omega n_2}{c}\right]\frac{d}{d\omega}(1 + b\delta n) \tag{2.90}$$

If we ignore the last term in (2.90) we get

$$\begin{aligned} \tau &= \frac{1}{c}(1 + b\delta n)\left[n_2 + \omega\frac{dn_2}{d\omega}\right] \\ &= \frac{N_{g2}}{c}(1 + b\delta n) \end{aligned} \tag{2.91}$$

We have already seen that the propagation constant for a particular mode has a range given by $\beta_2 < \beta < \beta_1$, and so b must lie in the range $0 < b < 1$. Thus the difference in propagation time between the highest and lowest order modes is

$$\begin{aligned}\frac{\delta t}{L} &= \frac{N_{g2}}{c}(1 + b_1\delta n - 1 - b_0\delta n)\\ &= \frac{N_{g2}}{c}(b_1\delta n - b_0\delta n)\\ &= \frac{N_{g2}\delta n}{c}(b_1 - b_0)\\ &= \frac{\delta n N_{g2}}{c}\end{aligned} \tag{2.92}$$

where we have taken $b_1 = 1$, and $b_0 = 0$. This should be compared to equation (2.86) obtained using ray-path analysis. One obvious difference is that (2.86) uses the group refractive index in the slab material, whereas (2.92) uses the group refractive index in the surrounding material. However, we should remember that this analysis uses the weakly guiding approximation, and so $N_{g1} \approx N_{g2}$ and (2.92) and (2.86) become equivalent.

2.2.6 Waveguide dispersion – ray-path and modal analysis

As well as suffering from modal and material dispersion, a propagating signal will also undergo *waveguide dispersion.* As we will see, waveguide dispersion results from the variation of propagation coefficient, and hence allowed angle of incidence, with wavelength. (When we considered propagation in section 2.2.2, we found that only those modes that satisfy the eigenvalue equation (2.60) will propagate successfully. As the refractive index of the waveguide material is present in (2.60), θ_i will vary if n changes with wavelength.) In common with the previous sections, we will initially use a ray-path analysis. However, this will limit us somewhat, and so we must resort to a modal analysis. With this in mind, some readers can neglect the latter part of this section on a first reading.

By following an analysis similar to that used in section 2.1.3, the *axial* transit time per unit length per unit of source line width, is

$$\begin{aligned}\tau = \frac{d\beta_{1x}}{d\omega} &= -\frac{\lambda_0^2}{2\pi c} \times \frac{d}{d\lambda}\beta_{1x}\\ &= -\frac{\lambda_0^2}{2\pi c} \times \frac{d}{d\lambda}\left[\frac{2\pi n_1 \sin\theta_i}{\lambda_0}\right]\\ &= \frac{\sin\theta_i}{c}\left[n_1 - \lambda_0\frac{dn_1}{d\lambda}\right] - \frac{n_1\lambda_0}{c} \times \frac{d}{d\lambda}\sin\theta_i\end{aligned} \tag{2.93}$$

The first term in (2.93) leads to the material dispersion resolved on to the x-axis (refer to equation 2.29). We should expect this because each individual mode will suffer from material dispersion. However, the second term in (2.93) leads to the waveguide dispersion, σ_{wg}. This is due to the waveguide propagation constants, and hence the permitted angles of incidence, varying with wavelength.

Let us now turn our attention to a modal analysis. In the previous section we expressed the propagation coefficient, of a particular mode, as (equation 2.89)

$$\beta = \frac{\omega n_2}{c}(1 + b\delta n)$$

or

$$\begin{aligned}\beta &= n_2 k_0(1 + b\delta n)\\ &= \beta_2(1 + b\delta n)\end{aligned}$$

Now, material dispersion results from the variation of *group refractive index* with wavelength. We have already defined k_0 as the free-space propagation coefficient, and so the group refractive index of a particular mode can be written as

$$\begin{aligned}N_g &= \frac{d\beta}{dk_0} = \frac{d\beta_2}{dk_0} + \frac{d}{dk_0}(\beta_2 b\delta n)\\ &= \frac{d\beta_2}{dk_0} + \frac{d}{d\beta_2}(\beta_2 b\delta n)\frac{d\beta_2}{dk_0} \qquad (2.94)\end{aligned}$$

Now, the V value of the guide is given by, equation (2.82)

$$\begin{aligned}V &= \frac{2\pi d(n_1{}^2 - n_2{}^2)^{\frac{1}{2}}}{\lambda_0}\\ &\approx \beta_2 d\sqrt{2\delta n} \qquad (2.95)\end{aligned}$$

where we have used the weakly guiding approximation ($n_1 \approx n_2$) and the binomial expansion. If we take the variation of δn with β_2 to be small, we can use (2.95) in (2.94) to give

$$\begin{aligned}N_g &= \frac{d\beta_2}{dk_0} + \frac{d}{dV}(Vb\delta n)\frac{d\beta_2}{dk_0}\\ &= \frac{d\beta_2}{dk_0}\left[1 + \frac{d}{dV}(Vb\delta n)\right]\\ &= \frac{d\beta_2}{dk_0}\left[1 + \delta n\frac{d}{dV}(Vb)\right]\end{aligned}$$

where we have neglected the variation of δn with V. The term outside the bracket is the group refractive index of the surrounding material, and so we can write the group refractive index of a propagating mode as

$$N_g = N_{g2}\left[1 + \delta n \frac{\mathrm{d}}{\mathrm{d}V}(Vb)\right] \tag{2.96}$$

Dispersion results from the variation of this group refractive index with wavelength. We can see from (2.96) that the dispersion will depend on two components: the first yields the material dispersion; however, the second term involves the mode-dependent term $\mathrm{d}(Vb)/\mathrm{d}V$. This is the ***waveguide dispersion.*** We should note that even if the guide is single-mode, waveguide dispersion will still be present. This is because both V and b are wavelength dependent, and so the linewidth of the optical source will contribute to waveguide and material dispersion.

We have now completed our study of dispersion in planar optical waveguides. We have found that optical signals are distorted by three mechanisms: modal dispersion caused by the dimensions of the waveguide allowing many modes to propagate; material dispersion caused by the group refractive index of the waveguide varying with wavelength; and waveguide dispersion caused by the waveguide propagation parameters being dependent on wavelength. It is worth remembering at this point that planar optical waveguides are usually very short in length, and so the dispersion effects we have been studying are not normally significant. However, these studies have introduced us to some very important concepts that will help our investigation of propagation in optical fibre.

2.2.7 Numerical aperture

Figure 2.10 shows two light rays entering a planar waveguide. Refraction of both rays occurs on entry; however ray1 fails to propagate in the guide

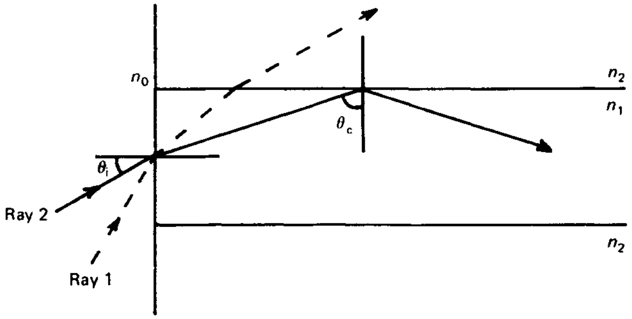


Figure 2.10 Construction for the determination of the numerical aperture

because it hits the boundary at an angle less than θ_c. On the other hand, ray2 enters the waveguide at an angle θ_i and then hits the boundary at θ_c; thus it will propagate successfully. If θ_i is the maximum angle of incidence, then the *numerical aperture, NA*, of the waveguide is equal to the sine of θ_i.

We can find the NA by applying Snell's law to ray2. Thus

$$\begin{aligned} n_0 \sin\theta_i &= n_1 \sin(90° - \theta_c) = n_1 \cos\theta_c \\ &= n_1\left[1 - \frac{n_2^2}{n_1^2}\right]^{\frac{1}{2}} \end{aligned}$$

Therefore

$$\sin\theta_i = \frac{1}{n_0}(n_1^2 - n_2^2)^{\frac{1}{2}} \tag{2.97}$$

If the guide is in air, then

$$NA = \sin\theta_i = (n_1^2 - n_2^2)^{\frac{1}{2}} \tag{2.98}$$

A large NA results in efficient coupling of light into the waveguide. However, a high NA implies that $n_1 >> n_2$ which results in a large amount of modal dispersion, so limiting the available bandwidth.

2.3 Propagation in optical fibres

So far we have only considered propagation in an infinite dielectric block and a planar dielectric waveguide. In this section we shall consider a cylindrical waveguide – the optical fibre. Light rays propagating in the fibre core fall into one of two groups. The first group consists of those light rays which pass through the axis of the core. Such rays are known as *meridional rays*, and figure 2.11a shows the passage of two of these rays propagating in a step-index fibre. With a little thought, it should be apparent that we can regard meridional rays as equivalent to the rays we considered in the planar dielectric.

The second group consists of those rays that never pass through the axis, known as *skew rays*. As figure 2.11b shows, these rays do not fully utilise the area of the core. As skew rays travel significantly farther than meridional rays, they generally undergo higher attenuation.

In the following section, we will apply Maxwell's equations to a cylindrical waveguide. Unfortunately this will involve us in some rather complex mathematics which some readers can neglect on a first reading. However, the full solution does yield some very important results, and these are quoted

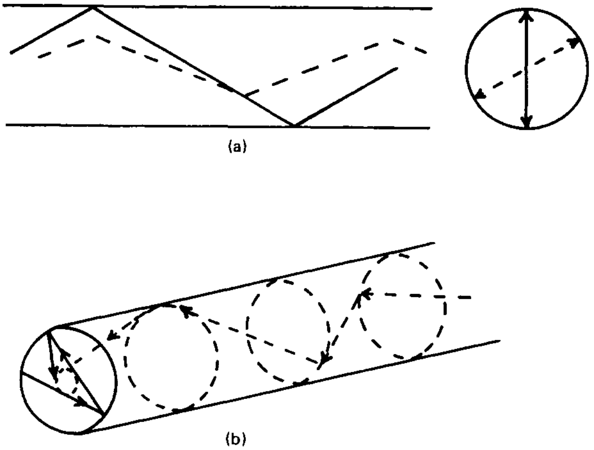


Figure 2.11 Propagation of (a) meridional and (b) skew rays in the core of a step-index, MM, optical fibre

in later sections. In common with the planar waveguide, we will find the condition for single-mode operation. We will then go on to study the dispersion characteristics of a cylindrical waveguide.

2.3.1 Propagation in step-index optical fibres

When we considered the planar waveguide, we solved Maxwell's equations using Cartesian coordinates. Now that we are considering a cylindrical waveguide, we must use *cylindrical co-ordinates*, as shown in figure 2.12.

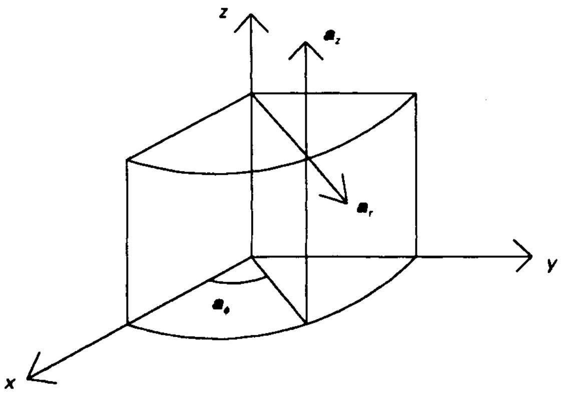


Figure 2.12 Definition of cylindrical co-ordinate set

In the following derivations, mention is made of *waveguide modes.* As we saw when we considered the planar waveguide, only certain modes (ray-paths) could propagate successfully. A similar situation exists in the cylindrical waveguide, as we will see when we come to solve the relevant eigenvalue equation.

Maxwell's equations, as applied to an ideal insulator such as glass, can be written as

$$\nabla \times \boldsymbol{E} = -\mu\frac{\partial \boldsymbol{H}}{\partial t} \qquad \text{(2.99a) and,} \quad \nabla \times \boldsymbol{H} = \epsilon\frac{\partial \boldsymbol{E}}{\partial t} \qquad \text{(2.99b)}$$

If we assume that the $\boldsymbol{E}$ and $\boldsymbol{H}$ fields propagate in the positive z direction with negligible attenuation, we can write

$$\boldsymbol{E} = E_0(r, \phi)\mathrm{e}^{\mathrm{j}(\omega t - \beta z)} \qquad \text{(2.100a)}$$

and

$$\boldsymbol{H} = H_0(r, \phi)\mathrm{e}^{\mathrm{j}(\omega t - \beta z)} \qquad \text{(2.100b)}$$

where we have used the phasor representation of $\boldsymbol{E}$ and $\boldsymbol{H}$, and β is the phase constant of any particular mode. If we substitute (2.100a) into (2.99a) we get

$$\nabla \times \boldsymbol{E} = \begin{vmatrix} \frac{1}{r}\boldsymbol{a}_r & \boldsymbol{a}_\phi & \frac{1}{r}\boldsymbol{a}_z \\ \frac{\partial}{\partial r} & \frac{\partial}{\partial \phi} & \frac{\partial}{\partial z} \\ E_r & rE_\phi & E_z \end{vmatrix} = -\mu\left(\frac{\partial}{\partial t}H_r\boldsymbol{a}_r + \frac{\partial}{\partial t}H_\phi\boldsymbol{a}_\phi + \frac{\partial}{\partial t}H_z\boldsymbol{a}_z\right)$$

By equating unit vectors, and performing the differentiation with respect to time, we get

$$\frac{1}{r}\left(\frac{\partial}{\partial \phi}E_z + \mathrm{j}\beta rE_\phi\right) = -\mathrm{j}\omega\mu H_r \qquad \text{(2.101a)}$$

$$-\left(\frac{\partial}{\partial r}E_z + \mathrm{j}\beta E_r\right) = -\mathrm{j}\omega\mu H_\phi \qquad \text{(2.101b)}$$

and

$$\frac{1}{r}\left(\frac{\partial}{\partial r}rE_\phi - \frac{\partial}{\partial \phi}E_r\right) = -\mathrm{j}\omega\mu H_z \qquad \text{(2.101c)}$$

By following a similar procedure with the *H* field, we get

$$\frac{1}{r}\left(\frac{\partial}{\partial\phi}H_z + \mathrm{j}\beta r H_\phi\right) = \mathrm{j}\epsilon E_r \tag{2.102a}$$

$$-\left(\frac{\partial}{\partial r}H_z + \mathrm{j}\beta H_r\right) = \mathrm{j}\epsilon E_\phi \tag{2.102b}$$

and

$$\frac{1}{r}\left(\frac{\partial}{\partial r}rH_\phi - \frac{\partial}{\partial\phi}H_r\right) = \mathrm{j}\epsilon E_z \tag{2.102c}$$

Now, if we rearrange (2.102b) to give H_r in terms of E_ϕ and H_z, and substitute the result into (2.101a), we can write

$$E_\phi = -\frac{\mathrm{j}}{\omega^2\mu\epsilon - \beta^2}\left(\frac{\beta}{r} \times \frac{\partial}{\partial\phi}E_z - \mu \times \frac{\partial}{\partial r}H_z\right) \tag{2.103a}$$

Similarly

$$E_r = -\frac{\mathrm{j}}{\omega^2\mu\epsilon - \beta^2}\left(\beta \times \frac{\partial}{\partial r}E_z + \frac{\omega\mu}{r} \times \frac{\partial}{\partial\phi}H_z\right) \tag{2.103b}$$

$$H_\phi = -\frac{\mathrm{j}}{\omega^2\mu\epsilon - \beta^2}\left(\frac{\beta}{r} \times \frac{\partial}{\partial\phi}H_z + \omega\epsilon \times \frac{\partial}{\partial r}E_z\right) \tag{2.103c}$$

and

$$H_r = -\frac{\mathrm{j}}{\omega^2\mu\epsilon - \beta^2}\left(\beta \times \frac{\partial}{\partial r}H_z - \frac{\omega\epsilon}{r} \times \frac{\partial}{\partial\phi}E_z\right) \tag{2.103d}$$

We can now substitute (2.103d) and (2.103c) into (2.102c) to give

$$\frac{\partial}{\partial r^2}E_z + \frac{1}{r} \times \frac{\partial}{\partial r}E_z + \frac{1}{r^2} \times \frac{\partial}{\partial\phi^2}E_z + (\omega^2\mu\epsilon - \beta^2)E_z = 0 \tag{2.104}$$

and, by substituting (2.103a) and (2.103b) into (2.102c) we get

$$\frac{\partial}{\partial r^2}H_z + \frac{1}{r} \times \frac{\partial}{\partial r}H_z + \frac{1}{r^2} \times \frac{\partial}{\partial\phi^2}H_z + (\omega^2\mu\epsilon - \beta^2)H_z = 0 \tag{2.105}$$

Equations (2.104) and (2.105) are the wave equations for the E and H fields as applied to a circular waveguide. We should note that these equations apply equally to the core and cladding of step-index optical fibre. As E and H are travelling waves, a general solution to these equations is

$$E_z(t, r, \phi, z) = Ax(r)y(\phi)e^{j(\omega t - \beta z)} \tag{2.106}$$

and

$$H_z(t, r, \phi, z) = Bx(r)y(\phi)e^{j(\omega t - \beta z)} \tag{2.107}$$

where A and B are constants, and $x(r)$ and $y(\phi)$ are yet to be determined.

To find $y(\phi)$, let us consider the variation of E_z with ϕ. A complete rotation of E_z occurs as ϕ goes from 0 to 2π. Hence E_z must be the same at all multiples of 2π, that is *E_z must be periodic with ϕ*. Thus we can say

$$y(\phi) = e^{j\nu\phi}$$

where ν is an integer. We can now substitute for $y(\phi)$ into (2.104) and use the result in the wave equation for E_z to give (after some cancellation)

$$\frac{\partial}{\partial r^2}x(r) + \frac{1}{r} \times \frac{\partial}{\partial r}x(r) + \left[(\omega^2\mu\epsilon - \beta^2) - \frac{\nu^2}{r^2}\right]x(r) = 0 \tag{2.108}$$

This equation is known as Bessel's differential equation, and the solution uses Bessel functions with the bracketed term in (2.108) as the argument.

To solve (2.108) we must consider two regions of interest – $r < a$ (the core of the fibre) and $r > a$ (the cladding of the fibre) where a is the fibre core radius. If we first consider the core and let $a \to 0$, there must be finite solutions to (2.108). (After all, light propagates in the core, and so there must be solutions to (2.108).) The functions that satisfy this condition are *Bessel functions of the first kind.* The solution to (2.108) then becomes

$$x(r) = J_\nu(ur)$$

and so

$$E_z(t, r, \phi, z) = AJ_\nu(ur)e^{j\nu\phi}e^{j(\omega t - \beta z)} \tag{2.109}$$

and

$$H_z(t, r, \phi, z) = BJ_\nu(ur)e^{j\nu\phi}e^{j(\omega t - \beta z)} \tag{2.110}$$

where $u^2 = \omega^2\mu\epsilon - \beta^2$ and ν is the order of the Bessel function. By noting that $1/\sqrt{\mu\epsilon} = n/c$, we can write

$$u^2 = \left[\frac{2\pi n_1}{\lambda_0}\right]^2 - \beta^2$$

or,

$$u^2 = \beta_1^2 - \beta^2$$

Let us now consider the cladding modes. Outside the core the E and H fields must decay away to zero for large radius. (This is a necessary condition for evanescent waves which we first encountered in section 2.2.1.) Now, if we let the radius tend to infinity, the Bessel functions of the first kind are finite and *not zero.* Hence we must use the *modified Bessel functions of the second kind* which give zero field at large radius. Thus

$$x(r) = K_v(wr)$$

and so

$$E_z(t, r, \phi, z) = CK_v(wr)e^{jv\phi}e^{j(\omega t - \beta z)} \tag{2.111}$$

and

$$H_z(t, r, \phi, z) = DK_v(wr)e^{jv\phi}e^{j(\omega t - \beta z)} \tag{2.112}$$

where the argument of the function is wr. We can find an expression for w by noting that as $a \rightarrow \infty$, $K_v(wa) \rightarrow e^{-wa}$. To ensure that $K_v(wa)$ tends to zero, we must have $w > 0$ and so

$$w^2 = \beta^2 - \left[\frac{2\pi n_2}{\lambda_0}\right]^2$$

or

$$w^2 = \beta^2 - \beta_2^2$$

We can immediately see that w^2 is different from u^2 in that the order of the subtraction is reversed, that is

$$w^2 = -(\beta_2^2 - \beta^2) \qquad \text{whereas } u^2 = \beta_1^2 - \beta^2$$

(This is of importance when we use the boundary conditions for $E_{\phi 2}$ and $H_{\phi 2}$.) We can also perform a simple check on the values of u^2 and w^2 by noting

$$\beta_2^2 < \beta^2 < \beta_1^2 \qquad \text{provided } n_2 < n_1.$$

Thus we can see that $J_v(ur)$ and $K_v(wr)$ satisfy the conditions placed on the E and H fields in the core and cladding. In passing, it is interesting to note that the lowest order mode will have a phase coefficient of β_1, that is, it will be totally confined to the core, whereas the highest order mode will propagate with β_2, that is, it will travel in the cladding – an evanescent wave. This is an identical situation to that encountered when we considered the planar optical waveguide.

The constants A, B, C and D can be found by applying boundary conditions to the E and H fields, that is the tangential component of the $E_{z,\phi}$ and $H_{z,\phi}$ fields are continuous across the boundary between the core and the cladding. So, for the tangential E_z field, we have

$$E_{z1} = E_{z2}$$

or

$$E_{z1} - E_{z2} = 0$$

Hence

$$AJ_v(ua) - CK_v(wa) = 0 \tag{2.113}$$

By a similar process, the continuity of the H_z field yields

$$BJ_v(ua) - DK_v(wa) = 0 \tag{2.114}$$

As regards the E_ϕ component, we can substitute (2.109) and (2.110) into (2.113) to give E_ϕ for $r < a$, and substitute (2.111) and (2.112) into (2.113) to give E_ϕ for $r > a$. Thus, the boundary condition, $E_{\phi 1} - E_{\phi 2} = 0$ becomes

$$\begin{aligned} &-\frac{\mathrm{j}}{u^2}\left[A\frac{\mathrm{j}v\beta}{a}J_v(ua) - B\omega\mu uJ'_v(ua)\right] \\ &+\frac{\mathrm{j}}{w^2}\left[C\frac{\mathrm{j}v\beta}{a}K_v(wa) - D\omega\mu wK'_v(wa)\right] = 0 \end{aligned} \tag{2.115}$$

where the prime symbols – ′ – denotes differentiation with respect to radius. (The apparent reversal in the sign of $E_{\phi 2}$ results from the difference between u^2 and w^2 noted earlier.) By applying the same procedure to the H_ϕ field, we can write

$$\begin{aligned} &-\frac{\mathrm{j}}{u^2}\left[B\frac{\mathrm{j}v\beta}{a}J_v(ua) + A\omega\epsilon_1 uJ'_v(ua)\right] \\ &-\frac{\mathrm{j}}{w^2}\left[D\frac{\mathrm{j}v\beta}{a}K_v(wa) + C\omega\epsilon_2 wK'_v(wa)\right] = 0 \end{aligned} \tag{2.116}$$

So, we have four equations (2.113), (2.114), (2.115) and (2.116) and four unknowns A, B, C and D. If we express these equations in the form of a matrix, the determinant will yield the permitted values of the propagation constant β. Thus

$$\begin{vmatrix} J_\nu(ua) & 0 & -K_\nu(wa) & 0 \\ \dfrac{\beta\nu}{au^2}J_\nu(ua) & \dfrac{j\omega\mu}{u}J'_\nu(ua) & \dfrac{\beta\nu}{aw^2}K_\nu(wa) & \dfrac{j\omega\mu}{w}K'_\nu(wa) \\ 0 & J_\nu(ua) & 0 & -K_\nu(wa) \\ \dfrac{-j\omega\epsilon_1}{u}J'_\nu(ua) & \dfrac{\beta\nu}{au^2}J_\nu(ua) & \dfrac{-j\omega\epsilon_2}{w}K'_\nu(wa) & \dfrac{\beta\nu}{aw^2}K_\nu(wa) \end{vmatrix} = 0$$

This rather complex determinant results in the following eigenvalue equation

$$\left[\frac{J'_\nu(ua)}{uJ_\nu(ua)} + \frac{K'_\nu(wa)}{wK_\nu(wa)}\right]\left[\beta_1{}^2\frac{J'_\nu(ua)}{uJ_\nu(ua)} + \beta_2{}^2\frac{K'_\nu(wa)}{wK_\nu(wa)}\right] = \left[\frac{\beta\nu}{a}\right]^2\left[\frac{1}{u^2} + \frac{1}{w^2}\right]^2 \tag{2.117}$$

The complete solution to this equation will yield the values of β for which a mode will propagate in the fibre. Unfortunately, (2.117) can only be solved by graphical/numerical techniques similar to those used when we considered the planar waveguide. (We will leave the solution to this equation until we consider a particular fibre type in section 2.3.3.)

Before we go on to examine dispersion in optical fibre, let us return to the four equations that link A, B, C and D – equations (2.113), (2.114), (2.115) and (2.116).

From (2.113) and (2.114) we can write

$$\frac{A}{C} = \frac{K_\nu(wa)}{J_\nu(ua)} \quad \text{and} \quad \frac{B}{D} = \frac{K_\nu(wa)}{J_\nu(ua)}$$

which simply relate the E and H fields inside the core to those in the cladding. However, (2.115) and (2.116) result in

$$\frac{A}{B} = \frac{-j\omega\mu}{\beta\nu}\left[\frac{J'_\nu(ua)}{uaJ_\nu(ua)} + \frac{K'_\nu(wa)}{waK_\nu(wa)}\right]\left[\frac{u^2w^2a^2}{u^2 + w^2}\right]$$

which indicates that the E and H fields inside the core are linked. This is an important result because it shows that, unlike rectangular waveguides, circular waveguides can support *hybrid modes* as well as the more familiar transverse electric and transverse magnetic modes. These hybrid modes are designated as *EH* or *HE* depending on the relative magnitude of the *E* and *H* field components transverse to the fibre axis. (We can visualise these hybrid modes as being the skew rays shown in figure 2.11.)

2.3.2 Dispersion in cylindrical waveguides

Let us now turn our attention to the dispersion characteristics of cylindrical waveguides. When we considered the simple planar waveguide, we saw the propagating signals suffered from modal, material and waveguide dispersion. In that particular instance, we used both the simple ray model and the more complete modal analysis. Unfortunately, now that we are considering the cylindrical waveguide, we must use modal analysis and so the mathematics becomes rather involved. Thus this section can be neglected on a first reading. In common with the previous section, all important results will be quoted later when required.

In section 2.2.3, equation (2.78), we defined a normalised propagation constant, b, as

$$b = \frac{\beta^2 - \beta_2^2}{\beta_1^2 - \beta_2^2} \qquad (2.118)$$

where $\beta_2 = 2\pi n_2/\lambda_0$, $\beta_1 = 2\pi n_1/\lambda_0$ and β is the propagation constant of any particular mode. If we proceed in a similar fashion to that used in section 2.2.5, we can express β as

$$\beta \approx \beta_2 (1 + \delta n b) \qquad (2.119)$$

where we have used the weakly guiding approximation that $n_1 \approx n_2$.

Now, propagating modes travel at their own group velocities given by c/N_g. As we saw in section 2.2.6, N_g can be expressed as (equation 2.94)

$$\begin{aligned} N_g &= \frac{d\beta_2}{dk} + \frac{d}{dk}(\beta_2 b \delta n) \\ &= N_{g2} + N_{g2} b \delta n + \beta_2 \frac{d}{dk}(b \delta n) \end{aligned} \qquad (2.120)$$

If we ignore the last term in (2.120), we get

$$N_g = N_{g2}(1 + b\delta n)$$

and so the group velocity is given by

$$v_g = \frac{c}{N_{g2}(1 + b\delta n)}$$

Thus the transit time of any particular mode is

$$\tau = \frac{1}{v_g} = \frac{N_{g2}}{c}(1 + b\delta n)$$

from which the modal dispersion is given by

$$\sigma_{mod} = \frac{\delta n N_{g2}}{c} \tag{2.121}$$

This expression, and derivation, is identical to that obtained for the planar optical waveguide (equation 2.92). We should note that this equivalence implies that only TE and TM modes propagate in the optical fibre. This is a consequence of the weakly guiding approximation, and so (2.121) is only an approximation.

In order to find the *material dispersion*, we need to find the variation of group velocity with wavelength. This is equivalent to differentiating (2.121) with respect to wavelength. So

$$D_{mat} = \frac{d}{d\lambda}\delta n \frac{N_{g2}}{c}$$

After some simple, but lengthy, manipulation, we find that D_{mat} is given by

$$\begin{aligned} D_{mat} &= \frac{dn_1}{d\lambda}\frac{N_{g2}}{cn_2} - \frac{n_1}{n_2^2}\frac{N_{g2}}{c}\frac{dn_2}{d\lambda} - \lambda_0\frac{\delta n}{c}\frac{d^2n_2}{d\lambda^2} \\ &\approx -\lambda_0\frac{\delta n}{c}\frac{d^2n_2}{d\lambda^2} \end{aligned} \tag{2.122}$$

(This should be compared with the expression for material dispersion in a block of glass given by equation 2.30.)

In order to study waveguide dispersion, we can proceed in an identical manner to that used when we studied the planar optical waveguide. So, we can express the group refractive index as

$$N_g = N_{g2}\left[1 + \delta n\frac{d}{dV}(Vb)\right] \tag{2.123}$$

The first term in this expression is related to the material dispersion, while the second gives rise to the waveguide dispersion, σ_{wg}. In order to find the amount of waveguide dispersion, we need to differentiate the last term in (2.123) with respect to wavelength, and then multiply by the line-width of the source. So

$$\begin{aligned} D_{wg} &= \frac{1}{c}\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[N_{g2}\delta n \frac{\mathrm{d}}{\mathrm{d}V}(Vb)\right] \\ &= \frac{1}{c} N_{g2}\delta n \frac{\mathrm{d}}{\mathrm{d}V}\frac{\mathrm{d}V}{\mathrm{d}\lambda}\frac{\mathrm{d}}{\mathrm{d}V}(Vb) \\ &= -\frac{N_{g2}\delta n}{c\lambda} V \frac{\mathrm{d}^2}{\mathrm{d}V^2}(Vb) \end{aligned} \tag{2.124}$$

where we have neglected the variation of N_{g2} and δn with wavelength. (If we had not made this assumption, the derivation would have yielded extra material dispersion terms.)

Examination of (2.124) shows that waveguide dispersion is caused by the waveguide propagation constants varying with the V value of the waveguide. At this point we could plot $\mathrm{d}(Vb)/\mathrm{d}b$ and get the second differential by graphical means. However, in multimode fibres the waveguide dispersion is generally small when compared with modal dispersion, and so we can neglect it. When we come to consider single-mode fibres, we will find that the waveguide dispersion is of the same order of magnitude as the material dispersion, and cannot be neglected.

2.3.3 Step-index multimode fibre

In section 2.3.1 we solved Maxwell's equations in a cylindrical waveguide. We found that the fibre can support transverse electric, TE, transverse magnetic, TM, and hybrid modes, EH or HE. Here we will apply the results from the previous analysis to a certain type of fibre which is in common use today. (Although most of the parameters used in the following have already been defined, some readers may have omitted the previous section and so they are defined again in this section.)

If the refractive index of the core is very nearly that of the fibre, that is $n_1 \approx n_2$, the fibre is known as *weakly guiding* [4] and it is this type of fibre which is most commonly used in telecommunications links. (Weakly guiding fibres support a small number of modes, and so modal dispersion is reduced.) With this restriction, the eigenvalue equation (2.117) for a fibre with radius a, reduces to

$$\frac{J_{v-1}(ua)}{J_v(ua)} = -\frac{w}{u} \times \frac{K_{v-1}(wa)}{K_v(wa)} \tag{2.125}$$

where $J_{\text{v}}(ua)$ is the Bessel function of the first kind, $K_{\text{v}}(wa)$ is the modified Bessel function of the second kind, and v is the Bessel function order. In obtaining (2.125), the following relationships were used:

$$J'_{\text{v}}(ua) = -J_{\text{v}-1}(ua) + \frac{v}{ua}J_{\text{v}}(ua)$$

and

$$K'_{\text{v}}(wa) = K_{\text{v}-1}(wa) - \frac{v}{wa}K_{\text{v}}(ua)$$

The parameters u and w are defined by

$$u^2 = \left[\frac{2\pi n_1}{\lambda_0}\right]^2 - \beta^2 \qquad \text{and} \qquad w^2 = \beta^2 - \left[\frac{2\pi n_2}{\lambda_0}\right]^2 \tag{2.126}$$

where β is the phase constant of a particular mode. (It should be noted that (2.125) is an approximation in that it, indirectly, predicts that only TE or TM modes propagate in the fibre.)

We can define a normalised frequency variable, similar to that used with the planar waveguide, as

$$V = \frac{2\pi a(n_1{}^2 - n_2{}^2)^{\frac{1}{2}}}{\lambda_0} \tag{2.127}$$

from which it is easy to show that

$$V^2 = (ua)^2 + (wa)^2 \tag{2.128}$$

As with the eigenvalue equation we encountered in section 2.2.2, the solution of equation (2.125) involves graphical techniques. The solutions can be obtained by plotting a graph of the left- and right-hand sides of (2.125) against ua, and then finding the points of intersection – the solutions to (2.125). Unfortunately, we would have to plot graphs for each value of Bessel function order, v, and for each of these plots there will be a certain number of solutions, m. Thus the propagating modes are usually known as LP_{vm}, where LP refers to *linear polarisation.*

Figure 2.13 shows a plot of both sides of (2.125) for a fibre with a normalised frequency of 12.5. The curves that follow a tangent function are the left-hand side of (2.125). Two plots of the right-hand side have been drawn – the upper plot is for a Bessel function order of 0, while the lower plot is for $v = 1$. (Here we have made use of $J_{-1}(x) = -J_1(x)$ and $K_{-1}(x) = K_1(x)$.)

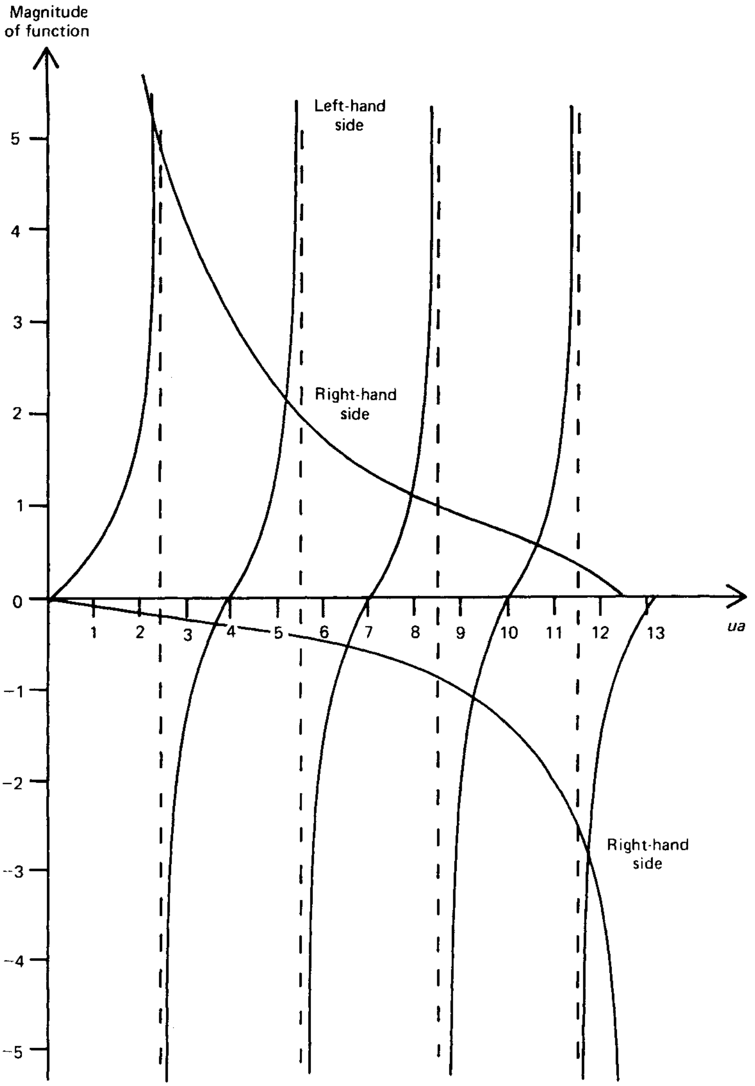


Figure 2.13 Eigenvalue graphs for the zero- and first-order modes in a cylindrical waveguide

An interesting feature of these plots is that there are no eigenvalues for $ua > V$. Thus V is sometimes known as the *normalised cut-off frequency.* From figure 2.13, we can see that, provided the argument ua is less than V, the number of eigenvalues, or the number of modes, is one greater than the number of zeros for the particular Bessel function order.

Thus for the zero-order function, LP_{0m}, the number of zeros with $ua < V$ is

4, and so the number of modes is 5. For $v = 1$, LP_{1m}, the number of modes is 4; note that $ua = 0$ is a possible solution. We could find the total number of modes using this method. However, for a large V, this method would be tedious to say the least. An alternative method of estimating the number of modes is based upon a knowledge of the numerical aperture.

If we ignore skew rays, then the numerical aperture of a fibre will be identical to that of the planar dielectric given by equation (2.98). With a cylindrical fibre however, any light falling within an *acceptance cone* will propagate. The solid acceptance angle of this cone, Ω, will be given by

$$\Omega = \pi\theta_i^2 \tag{2.129}$$

Now, if θ_i is small, then $\theta_i \approx \sin\theta_i$ – the numerical aperture. So, Ω will be given by

$$\Omega = \pi NA^2 = \pi(n_1^2 - n_2^2) \tag{2.130}$$

We can now estimate the number of modes propagating by noting that the number of modes per unit angle is $2A/\lambda_0^2$, where A is the cross-sectional area of the fibre end. (The factor 2 is included because each mode can take on one of two different polarisation states.) Therefore the number of modes, N_{max}, will be

$$N_{max} = \frac{2A}{\lambda_0^2} \times \Omega = \frac{2\pi a^2}{\lambda_0^2} \times \pi(n_1^2 - n_2^2)$$

or

$$N_{max} = \frac{V^2}{2} \tag{2.131}$$

As we have already seen, evanescent waves are present in the cladding, and so not all of the transmitted power is confined to the core. By using the weakly guiding approximation, the proportion of cladding power, P_{clad}, to total power, P, can be approximated to [4]

$$\frac{P_{clad}}{P} = \frac{4}{3}N_{max}^{-\frac{1}{2}} \tag{2.132}$$

Example

Light of wavelength 850 nm is propagating in 62.5 μm core diameter, PCS fibre with $n_1 = 1.5$ and $n_2 = 1.4$. The group refractive index of

the cladding material, N_{g2}, is 1.53. Estimate the modal dispersion, and the proportion of the total power carried in the cladding.

We can estimate the number of 850 nm modes propagating in the core by using equation (2.131). Thus

$$
\begin{aligned}
N_{\max} &= \frac{V^2}{2} \\
&= \frac{2\pi a^2}{\lambda_0^2} \times \pi(n_1^2 - n_2^2) \\
&= 7.74 \times 10^3
\end{aligned}
$$

As the number of propagating modes is quite large, we can approximate σ_{mod} by the expression derived for the planar waveguide, equation (2.92) or (2.121). Therefore

$$
\begin{aligned}
\sigma_{\text{mod}} &= \delta n \frac{N_{g2}}{c} \\
&= \frac{0.071 \times 1.53}{3 \times 10^8} \\
&= 364 \text{ ps/km}
\end{aligned}
$$

Such a large value of modal dispersion will tend to dominate the dispersion characteristic. Thus the bandwidth–length product of step-index, MM fibres varies from less than 1 MHz km to 100 MHz km.

As regards the distribution of power, we can use equation (2.132) to give

$$
\begin{aligned}
\frac{P_{\text{clad}}}{P} &= \frac{4}{3} N_{\max}^{-\frac{1}{2}} \\
&= \frac{4}{3} (7.74 \times 10^3)^{-\frac{1}{2}} \\
&= 1.5 \times 10^{-2}
\end{aligned}
$$

Thus only 1.5 per cent of the total power is carried in the cladding.

2.3.4 Step-index single-mode fibre

At the start of the previous section, we saw that propagating modes had to satisfy the following eigenvalue equation

$$\frac{J_{\nu-1}(ua)}{J_\nu(ua)} = -\frac{w}{u} \times \frac{K_{\nu-1}(wa)}{K_\nu(wa)} \tag{2.133}$$

In a SM fibre, only the lowest order mode can propagate. This corresponds to $V = 0$ and so (2.133) becomes

$$\frac{J_1(ua)}{J_0(ua)} = \frac{w}{u} \times \frac{K_1(wa)}{K_0(wa)} \tag{2.134}$$

As we saw earlier, there are m possible modes for $\nu = 0$. So, for SM operation, m must be equal to one (that is, only the LP_{01} mode can propagate) and this sets a limit to ua. As the maximum value of ua is the normalised cut-off frequency, there will also be a limit to V. Now, the first discontinuity in the zero-order plot drawn in figure 2.13 occurs at $ua = 2.405$, and so V must be less than 2.405 for SM operation. If we use the definition of V (equation 2.127) then the condition for SM operation is

$$\lambda_0 > 2.6a(n_1^2 - n_2^2)^{\frac{1}{2}} \tag{2.135}$$

The term in brackets is the numerical aperture, which for practical SM fibres is usually about 0.1. Thus, for operation at 1.3 μm, the fibre diameter should be less than 10 μm. It is interesting to note that because the condition for SM operation is dependent on wavelength, the linewidth of the source causes waveguide dispersion.

Dispersion in SM fibres is due to material and waveguide effects. As we saw in section 2.3.2, the material dispersion in a weakly guiding optical fibre (2.122) is given by

$$D_{\text{mat}} \approx -\lambda_0 \frac{\delta n}{c} \frac{\mathrm{d}^2 n_2}{\mathrm{d}\lambda^2} \tag{2.136}$$

whereas the waveguide dispersion (2.124) is given by

$$D_{\text{wg}} = -\frac{N_{\text{g}2}\delta n}{c\lambda} V \frac{\mathrm{d}^2}{\mathrm{d}V^2}(Vb) \tag{2.137}$$

In order to find D_{wg} we need to find $\mathrm{d}^2(Vb)/\mathrm{d}V^2$. D. Gloge [4] has derived an expression for $\mathrm{d}(Vb)/\mathrm{d}V$ as

$$\begin{aligned}\frac{\mathrm{d}(Vb)}{\mathrm{d}V} &= b\left[1 - \frac{2J_\nu^2(ua)}{J_{\nu+1}(ua)\,J_{\nu-1}(ua)}\right] \\ &= b\left[1 + \frac{2J_0^2(ua)}{J_1^2(ua)}\right]\end{aligned} \tag{2.138}$$

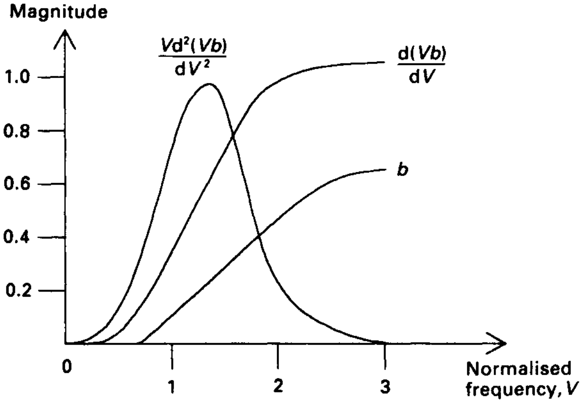


Figure 2.14 Variation of b, $d(Vb)/dV$ and $Vd^2(Vb)/dV^2$ with normalised frequency, V, in an optical fibre

for SM operation. The argument of the Bessel functions in (2.138) is defined by equation (2.126). Unfortunately the second differential can only be obtained by graphical means.

Figure 2.14 shows the variation of $Vd^2(Vb)/dV^2$ with V. Most SM fibres are fabricated with V values between 2.0 and 2.4. Over this range of V, the second differential term can be approximated by

$$\frac{Vd^2(Vb)}{dV^2} \approx -\frac{1.984}{V^2} \tag{2.139}$$

and so (2.137) can be approximated by

$$D_{wg} = \frac{N_{g2}\delta n}{c\lambda_0}\,\frac{1.984}{V} \tag{2.140}$$

In common with the material dispersion, the units of D_{wg} are usually ns/nm/km, and so we can reduce D_{wg} by using narrow line-width sources.

In order to find the total dispersion, we can simply add together the waveguide and material dispersions. However, D_{wg} is positive, while D_{mat} becomes negative for wavelengths above about 1.3 μm. Thus the material and waveguide dispersion will cancel each other out at a certain wavelength. Figure 2.15 shows the theoretical variation with wavelength of D_{wg}, D_{mat} and total dispersion ($D_{mat} + D_{wg}$) for a typical SM fibre. As can be seen, reduction of the core radius moves the dispersion zero to higher wavelengths. (In practice, the zero dispersion point is usually limited to 1.55 μm. This is because it is difficult to manufacture very small core fibres.) Fibres which

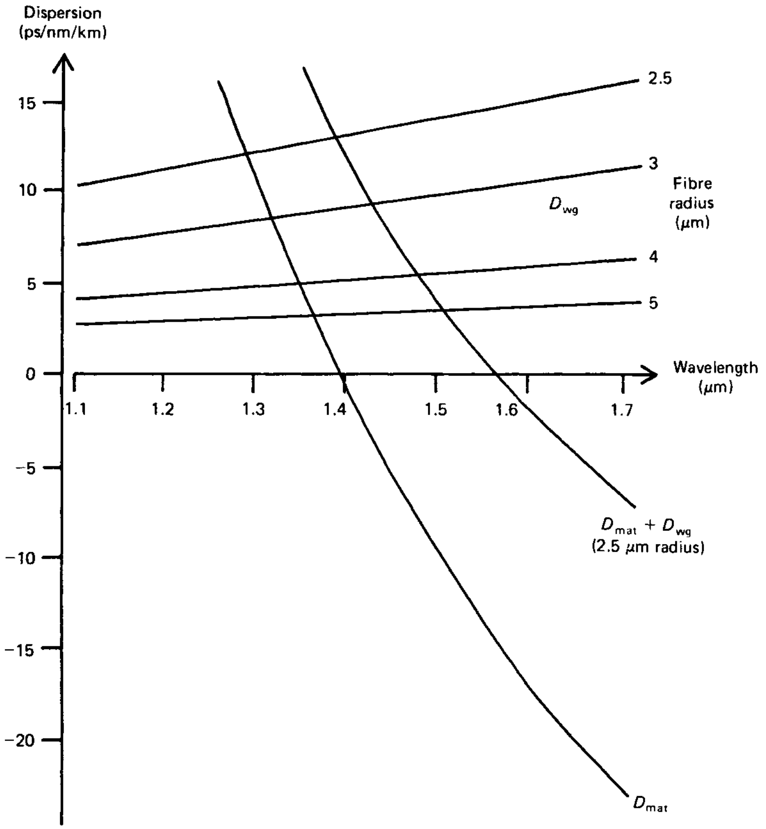


Figure 2.15 Showing the shift in the zero dispersion point, obtained by balancing D_{mat} with D_{wg}

exhibit this characteristic are known as *dispersion shifted* fibres [5]. As the higher wavelength results in lower attenuation, they are of great importance in long-haul, high data-rate routes. Single-mode fibres have a very high information capacity, with a typical bandwidth-length product greater than 40 GHz km.

Let us now consider the distribution of optical power in the fibre. As there is only one mode propagating in the fibre, equation (2.132) becomes invalid, and we must use [4]

$$\frac{P_{core}}{P} = \left(1 - \frac{u^2}{V^2}\right)\left[1 + \frac{J_0(ua)}{J_1^2(ua)}\right] \tag{2.141}$$

and

$$\frac{P_{\text{clad}}}{P} = 1 - \frac{P_{\text{core}}}{P} \tag{2.142}$$

Application of these two equations indicates that as V reduces, more power is carried in the cladding until, in the limit as $V \rightarrow 0$, all the power is in the evanescent wave. This is of great importance when considering fibre couplers.

2.3.5 Graded-index fibre

We have already seen that modal dispersion causes pulse distortion in MM fibres. However, we can use *graded-index* fibres to reduce this effect. The principle behind these fibres is that the refractive index is steadily reduced as the distance from the core centre increases. Thus constant refraction will constrain the propagating rays to the fibre core. With such a profile, the higher order modes travelling in the outer regions of the core, will travel faster than the lower order modes travelling in the high refractive index region. If the index profile is carefully controlled, then the transit times of the individual modes should be identical, so eliminating modal dispersion.

The ideal index profile for these fibres is given by

$$\begin{aligned} n(r) &= n_1 \, [1 - 2 \times (\delta n/n_2) \times (r/a)^{\alpha}]^{\frac{1}{2}} \qquad && 0 \leq r \leq a \\ &= n_2 && r \geq a \end{aligned} \tag{2.143}$$

where n_1 is the refractive index at the centre of the core, and α defines the core profile. As the wave equation is rather complex, we will not consider propagation in any detail. However, analysis shows that the optimum value of α is approximately 2. With α in this region, σ_{mod} is usually less than 100 ps/km. Of course there will still be material and waveguide dispersion effects and, depending on the source, these result in a bandwidth–length product that is typically less than 1 GHz km. (A considerable reduction in bandwidth results if α is not optimal.)

2.4 Calculation of fibre bandwidth

If a very narrow optical pulse propagates down a length of fibre then, because of dispersion, the width of the output signal will be larger than that of the input. If the input pulse width is typically 10 times less than the output pulse width, the output signal will closely approximate the impulse response of the fibre. Depending on the type and length of fibre, this impulse response can take on several different shapes. A Gaussian response results if there is considerable transfer of power between propagating modes – *mode mixing*. Mode mixing results from reflections off imperfections due to *micro-bending* (caused by laying the fibre over a rough surface) and scattering from fusion splices and connectors). An exponential response can also be obtained. Such an impulse response results from considerable modal dispersion in the absence of mode mixing. Of course, these are idealised extremes; in practice, the impulse response is a combination of the two. So, any bandwidth calculations performed with either pulse shape, will only give an indication of the available capacity.

If we consider a Gaussian impulse response, we can write the received pulse shape, h_{out}, as

$$h_{out}(t) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-t^2/(2\sigma^2)} \tag{2.144}$$

Hence the received pulse spectrum, H_{out}, will be given by

$$H_{out}(\omega) = e^{-\omega^2\sigma^2/2} \tag{2.145}$$

where σ is the root mean square (r.m.s.) width of the pulse. The -3 dB bandwidth is equal to the frequency at which the received power is half the d.c. power, that is

$$\frac{H_{out}(\omega)}{H_{out}(0)} = e^{(-\omega^2\sigma^2/2)} = \frac{1}{2} \tag{2.146}$$

and so the 3 dB bandwidth will be

$$\omega_{opt} = 1.18/\sigma \tag{2.147}$$

We should note that this is the *optical* bandwidth. We are more usually concerned with the electrical bandwidth, that is the bandwidth at the output of the detector. The detector converts optical power to an electrical current, and so a 3 dB drop in optical power produces a 6 dB drop in electrical power. Thus the electrical bandwidth, ω_{elec}, is the frequency at which the optical power is $1/\sqrt{2}$ times the d.c. value. Hence ω_{elec} will be given by

$$\omega_{elec} = 0.83/\sigma \tag{2.148}$$

From now on, we shall use the electrical bandwidth whenever we refer to bandwidth.

The three sources of dispersion will determine the r.m.s. width of the received pulse. We have already seen that we can add the material dispersion, D_{mat}, and the waveguide dispersion, D_{wg}, together. However, in order to account for the modal dispersion, we must add σ_{mod} on a mean square basis.

(This is a result of convolving the individual pulse shapes due to the modal and source-dependent dispersion – see the paper by Personick [6]. So, the total dispersion, σ, will be given by

$$\sigma^2 = \sigma_{\text{mod}}{}^2 + (\sigma_{\text{mat}} + \sigma_{\text{wg}})^2 \tag{2.149}$$

Example

An optical fibre link uses 25 μm radius, MM fibre with n_1 = 1.5, N_{g1} = 1.64, n_2 = 1.4, N_{g2} = 1.53 and D_{mat} = 500 ps/nm/km. An 850 nm wavelength LED, with a linewidth of 30 nm, is used as the source. Calculate the bandwidth–length product of the link assuming a Gaussian impulse response.

By applying equation (2.121), we get

$$\begin{aligned}\sigma_{\text{mod}} &= \delta n \frac{N_{g2}}{c}\\ &= 364 \text{ ns/km}\end{aligned}$$

We can find the material dispersion by multiplying D_{mat} by the linewidth of the source to give

$$\begin{aligned}\sigma_{\text{mat}} &= 500 \times 30\\ &= 15 \text{ ns/km}\end{aligned}$$

As the waveguide dispersion is negligible when compared with σ_{mat}, the total fibre dispersion will be given by

$$\begin{aligned}\sigma &= \sqrt{\sigma_{\text{mod}}{}^2 + \sigma_{\text{mat}}{}^2}\\ &\approx 364 \text{ ns/km}\end{aligned}$$

Thus we can see that the modal dispersion is the dominant factor.

We can now find the bandwidth of the link by using equation (2.148) to give

$$\begin{aligned}\omega_{\text{elec}} &= 0.83/\sigma\\ &= 2.3 \times 10^6 \text{ rad/s}\end{aligned}$$

from which the bandwidth–length product is 366 kHz.km.

Repeat the previous if the link uses SM fibre with n_1 = 1.48, N_{g1} = 1.64, n_2 = 1.47, N_{g2} = 1.63, and D_{mat} = −5 ps/nm/km. Assume that a 1 nm linewidth, 1.3 µm wavelength laser is used as the source. If the source is then replaced by a 30 nm linewidth LED, determine the new bandwidth.

As the fibre is single-mode, we need to find the waveguide dispersion from equation (2.140). Thus

$$D_{wg} = \frac{N_{g2}}{c}\frac{\delta n}{\lambda_0}\frac{1.984}{V}$$

where V = 2.405 for single-mode operation. Therefore

$$D_{wg} = \frac{1.63 \times 6.8 \times 10^{-3}}{3 \times 10^5 \times 1.3 \times 10^3} \times \frac{1.984}{2.405}$$
$$= 23.4 \text{ ps/nm/km}$$

As the source has a linewidth of 1 nm, we find that

$$\sigma_{wg} = 23.4 \text{ ps/km}$$

The material dispersion is

$$\sigma_{mat} = 5 \text{ ps/km}$$

and so the total dispersion is 28.4 ps/km. Thus the bandwidth of the link is 4.7 GHz km.

If we use a 30 nm linewidth LED source, we find

$$\sigma_{wg} = 23.4 \times 30$$
$$= 702 \text{ ps/km}$$

and

$$\sigma_{mat} = 5 \times 30$$
$$= 150 \text{ ps/km}$$

Thus the bandwidth with the LED source is 155 MHz km.

2.5 Attenuation in optical fibres

Coupling losses between the source/fibre, fibre/fibre and fibre/detector can cause attenuation in optical links. Losses can also occur due to bending the fibre too far, so that the light ray hits the boundary at an angle less than θ_c. As these loss mechanisms are extrinsic in nature, we can reduce them by taking various precautions. However the fibre itself will absorb some light, and it is this attenuation that concerns us here.

The attenuation/wavelength characteristic of a typical glass fibre is shown in figure 2.16. This figure also shows the relative magnitudes of the four main sources of attenuation: electron absorption, Rayleigh scattering, material absorption and impurity absorption. The first three of these are known as *intrinsic absorption mechanisms* because they are a characteristic of the glass itself. Absorption by impurities is an *extrinsic absorption mechanism*, and we will examine this loss first.

2.5.1 Impurity absorption

In ordinary glass, impurities such as water and transition metal ions, dominate the attenuation characteristic. However, because the glass is usually thin, the attenuation is not of great concern. In optical fibres that are several kilometres long, the presence of any impurities results in very high attenuations which may render the fibre useless; a fibre made of the glass used in lenses would have a loss of several thousand dB per kilometre. By contrast, if we produced a window out of the glass used in the best optical fibres, then we would be able to see through a window 30 km thick!

The presence of water molecules can dominate the extrinsic loss. The OH bond absorbs light at a fundamental wavelength of about 2.7 μm and this, together with interactions from silicon resonances, causes harmonic peaks at 1.4 μm. 950 and 725 nm, as in figure 2.16. Between these peaks are regions of low attenuation – the transmission windows at 850 nm, 1.3 μm and 1.55 μm. As a high water concentration results in the tails associated with the peaks being large, it is important to minimise the OH impurity concentration.

In order to reduce attenuation to below 20 dB/km, a water concentration of less than a few parts per billion (ppb) is required. Such values are being routinely achieved by using the *modified chemical vapour deposition* manufacturing process (examined in section 2.6.2). Different manufacturing methods will produce lower water concentration. For example, the *vapour-phase axial deposition, VAD*, process can produce fibres with OH concentrations of less than 0.8 ppb. With this impurity level, the peaks and valleys in the attenuation curve are smoothed out, and this results in a typical loss of less than 0.2 dB/km in the 1.55 μm window.

The presence of transition metal ions (iron, cobalt, copper, etc.) can cause additional loss. If these metals are present in concentration of 1 ppb,

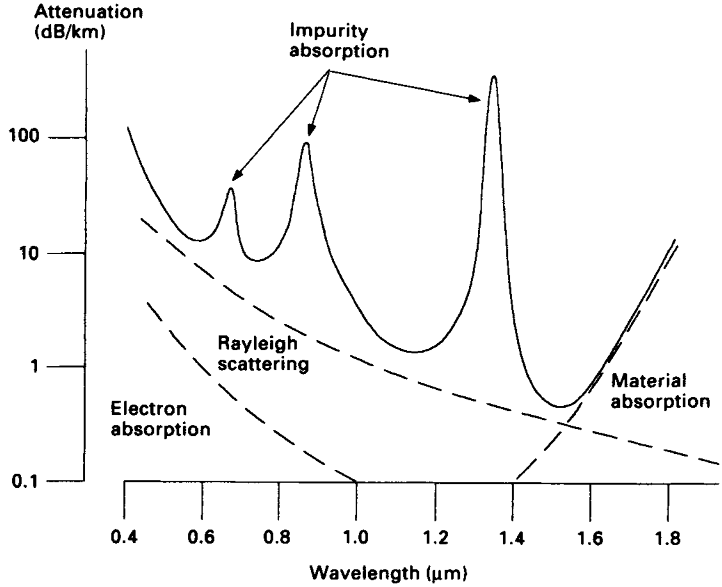


Figure 2.16 Attenuation/wavelength characteristic of a silica-based glass fibre

then the attenuation will increase by about 1 dB/km. In telecommunications grade fibre, the loss due to transition metal ion impurities is usually insignificant when compared to the OH loss.

2.5.2 Rayleigh scattering

Rayleigh scattering results from the scattering of light from small irregularities in the structure of the core. (A similar mechanism makes the sky appear blue, by scattering light off dust particles in the atmosphere.) These irregularities are usually due to density fluctuations which were frozen into the glass at manufacture. Consequently this is a fundamental loss mechanism, which places a lower bound on the fibre attenuation. Rayleigh scattering is only significant when the wavelength of the light is of the same order as the dimensions of the scattering mechanism. In practice, this loss reduces as the fourth power of wavelength, and so operation at long wavelengths is desirable.

2.5.3 Material absorption

It might be thought that operation at longer wavelengths will produce lower losses. In principle this is correct; however the atomic bonds associated with the core material will absorb the long wavelength light – *material absorption*. Although the fundamental wavelengths of the absorption bonds are outside the range of interests, the tails are significant. Thus operation at wavelengths greater than 1.55 μm will not produce a significant drop in attenuation. However, fibres made out of fluoride glasses, for example *ZrF4*, will transmit higher wavelength light.

2.5.4 Electron absorption

In the ultra-violet region, light is absorbed by photons exciting the electrons in a core atom, to a higher energy state. (Although this is a form of material absorption, interaction occurs on the atomic scale rather than the molecular scale.) In silica fibres, the absorption peak occurs in the ultra-violet region at about 0.14 μm; however, the tail of this peak extends through to about 1 μm, so causing attenuation in the transmission windows.

2.5.5 PCS and all-plastic fibres

In PCS fibre, the main absorption peaks are due to the O–H bond resonances, identical to an all-glass fibre, and the C–H bond resonances due to the plastic cladding. The net result is that PCS fibres exhibit a transmission window at 870 nm with a typical attenuation of 8 dB/km, and so PCS links can use relatively cheap near infra-red light sources. In view of the relatively low attenuation, and the fact that PCS fibre is step-index MM, most PCS links are dispersion limited rather than attenuation limited.

All-plastic fibres exhibit very high attenuation due to the presence of C–H bonds in the core material. These bonds result in a transmission window at 670 nm, with a typical attenuation of 200 dB/km. As well as the high attenuation peaks caused by the complex C–H bonds, there is a large amount of Rayleigh scattering in all-plastic fibres. This is due to scattering from the large chain molecules that make up the material. Although plastic fibres exhibit very low bandwidth–length products and very high attenuation, there is considerable interest in using such fibres for localised distribution systems such as computer installations.

2.6 Fibre materials and fabrication methods

2.6.1 Materials

Most of the glass fibres in use today are fabricated out of silica, SiO_2. This has a refractive index of between 1.44 and 1.46, and doping with various chemicals produces glasses of different refractive indices. In order to increase the refractive index, oxides of germanium, GeO_2, or phosphorus, P_2O_5, are commonly used. A decrease in n results from doping with boron oxide, B_2O_3, or fluorine, *F*. The amount of dopant used determines the refractive index of the fibre. For example, a 5 per cent concentration of GeO_2 will increase the refractive index of SiO_2 from 1.46 to 1.465. It should be noted that heavy doping is undesirable, because it can affect both the fibre dispersion and attenuation.

Plastic clad silica, *PCS*, fibres are commonly made from a pure silica core, with a silicone resin cladding. This gives a cladding refractive index of 1.4 at 850 nm, resulting in an acceptance angle of 20°. We can increase the NA by using a Teflon cladding. This material has a refractive index of about 1.3, resulting in an acceptance angle of 70°. As we have seen, the attenuation of these fibres is not as large as for the all-plastic fibres, and so PCS fibres find many applications in medium-haul routes.

All-plastic fibres are commonly made with a polystyrene core, $n_1 = 1.6$, and a methyl methacrylate cladding, $n_1 = 1.5$. These fibres usually have a core radius of 300 μm or more, and so can couple large amounts of power. Unfortunately, because the attenuation is very high and the bandwidth very low, these fibres are only useful in very short communication links, or medical applications.

2.6.2 Modified chemical vapour deposition (MCVD)

Most low-loss fibres are made by producing a glass *preform* which has the refractive index profile of the final fibre, that is MM, SM, or graded-index, but is considerably larger. If the preform is heated and a thin strand is pulled from it, then an optical fibre can be drawn from the preform. The next section describes this process in greater detail; here we will consider preform fabrication.

MCVD is probably the most common way of producing a preform. (An alternative method is *vapour-phase axial deposition, VAD*. However this process is not in common use at present.) The first step in the process is to produce a SiO_2 tube, or *substrate*. This forms the cladding of the final fibre, and so it may need to be doped when formed. As shown in figure 2.17a, the substrate is made by depositing a layer of SiO_2 particles and dopants, called a *soot*, onto a rotating ceramic former, or *mandrel*.

When the soot reaches the required depth, it is vitrified into a clear glass by heating to about 1400°C. The mandrel can then be withdrawn. (A complete preform can also be made by depositing the core glass first, and then depositing the cladding. The mandrel can then be withdrawn, and the resulting tube collapsed to form a preform. This process is known as *outside vapour phase oxidation*, and the first optical fibres with attenuations of less than 20 dB/km were made using this process.)

In the MCVD process, the cladding tube is placed in a lathe, and the gaseous core constituents pass through it (figure 2.17b). As the deposit forms,

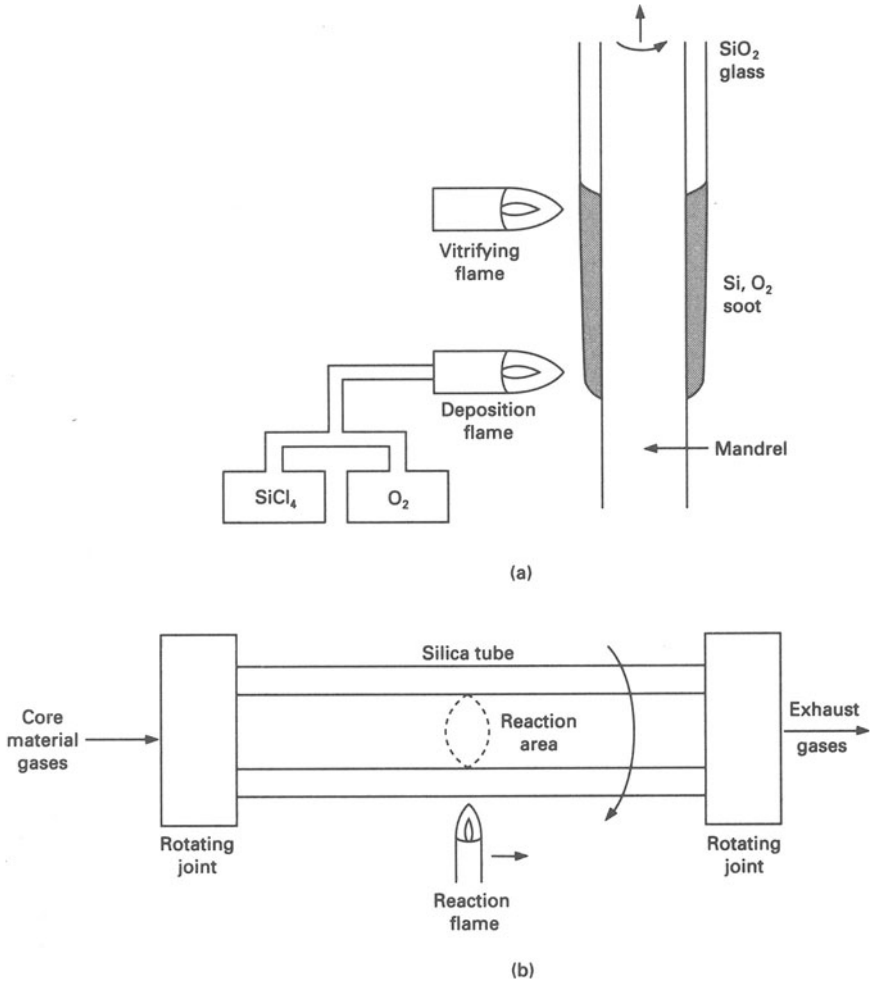


Figure 2.17 (a) Formation of silica cladding tube and (b) deposition of core glasses

an oxyhydrogen torch sinters the core particles into a clear glass. When the required core depth is achieved, the vapour is shut off, and strong heating causes the tube to collapse. The result is a preform with the required refractive index profile. (A graded-index preform can be produced by varying the dopant concentrations during deposition.) The preform is then placed in a *pulling tower* which draws out the fibre.

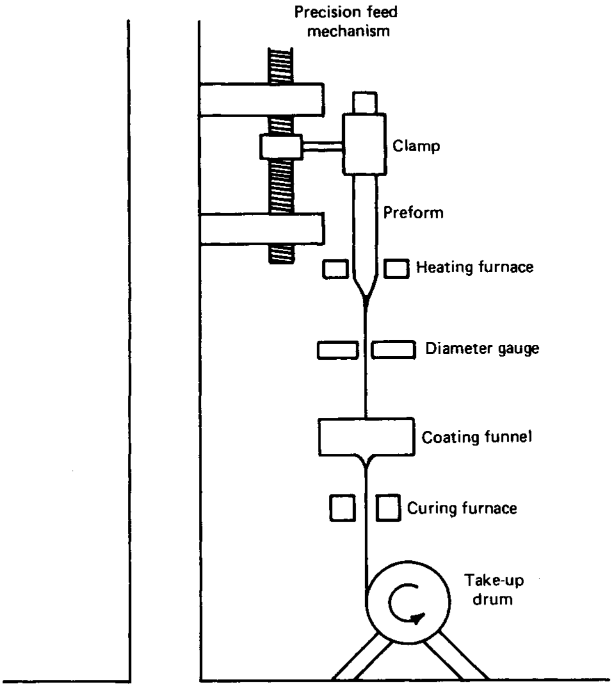


Figure 2.18 Schematic of a fibre pulling tower

2.6.3 Fibre drawing from a perform

Having produced a preform, the fibre is drawn from it in a *fibre pulling tower*, shown schematically in figure 2.18. A clamp at the top of the tower holds the preform in place, and a circular drawing furnace softens the tip.

A filament of glass is drawn from the tip, and attached to a take-up drum at the base of the tower. As the drum rotates, it pulls the fibre from the preform. The rate of drum rotation determines the thickness of the fibre, and so a non-contact thickness gauge regulates the drum speed by means of a feedback loop.

Below the gauge, the fibre passes through a funnel containing a plastic coating which helps to protect the fibre from impurities and structural damage. A curing lamp ensures that the coating is a solid before the fibre reaches the take-up drum. A typical preform with a diameter of 2 cm, and a length of 1 m, will produce several kilometres of 125 μm diameter fibre.

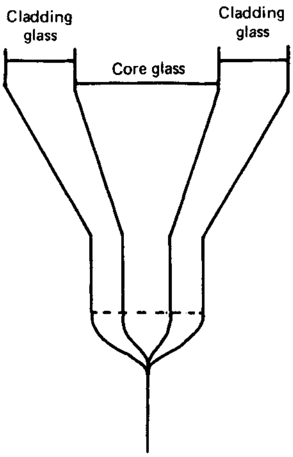


Figure 2.19 Double crucible method of optical fibre production

2.6.4 Fibre drawing from a double crucible

A major disadvantage of fibre-pulling from a preform is that the process does not lend itself to continuous production. However, if the fibre can be drawn directly from the core and cladding glasses, then a continuous process results, making the fibre cheaper to produce. Such a process is the *double crucible* method of fibre manufacture (also known as the *direct melt* technique).

In a *double crucible* pulling tower, two concentric funnels, the double crucible, replace the preform. As figure 2.19 shows, the outer funnel contains the cladding material, while the inner funnel contains the core glass. In order to reduce contamination, the crucibles are usually made of platinum. The crucibles are heated to melt the glasses, and the fibre can then be drawn as previously described. Rods of the core and cladding material can be made by melting mixtures of the purified glass constituents, and a continuous drawing process results from feeding these into the crucibles. It should be obvious that this method of manufacture is only suitable for the production of step-index glass, PCS or all-plastic fibres.

2.7 Connectors and couplers

2.7.1 Optical fibre connectors

When we wish to join two optical fibres together, we must use some form of connector. We could simply butt the two fibre ends together, and use an

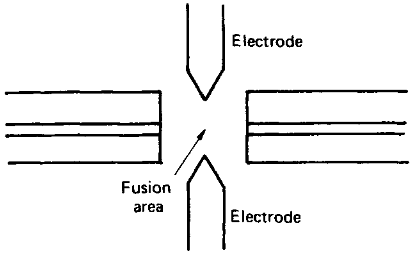


Figure 2.20 Schematic diagram of a fusion splicer

epoxy resin to hold them in place. However, if the fibres move slightly while the epoxy is setting, then a considerable amount of power can be lost. One solution to the problem is to fuse the two fibre ends together, so making a stable, low loss joint. This method, known as *fusion splicing*, is shown in schematic form in figure 2.20.

The two fibre ends are viewed through a microscope, and butted together using micro-positioners. When they are correctly aligned, an electric arc is struck across the join, causing the two ends to melt and fuse. Inspection with a microscope reveals whether the joint is satisfactory; if it is not, then the join can be broken and remade. This technique results in a typical loss per splice of 0.2 dB, and so it is particularly attractive for use in long-haul routes.

Although fusion splicing results in very low loss connections, it does produce a permanent connection. In medium and short-haul routes, where it may be desired to change the network configuration at some time, this is a positive disadvantage. In these systems, demountable connectors are used. There are many different types currently available, but nearly all use a precision made ferrule to accurately align the fibre cores, and so reduce losses. This method is shown in figure 2.21.

Prior to insertion into the connector, the protective fibre coating is first

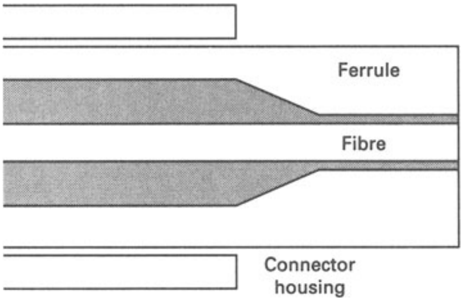


Figure 2.21 Basic construction of a ferrule type connector

stripped off using a solvent. A taper at the end of the connector ferrule grips the inserted fibre, which usually protrudes slightly from the end. The fibre is then *cleaved* to produce a plane surface. (Cleaving involves scoring the fibre surface with a diamond, and then gently bending away from the scratch until the fibre snaps. The result should be a plane end.) Any irregularities on the surface of the fibre end will scatter the light, resulting in a loss of power, thus polishing of the fibre end with successively finer abrasives is often used. The main body of the connector is then crimped on to the fibre, resulting in a mechanically strong connection. Most manufacturers will supply sources and detectors in packages which are compatible with the fibre connectors, and so installation costs can be kept low.

2.7.2 Optical fibre couplers

In order to distribute or combine optical signals, we must use some form of coupler. Again there are many types, but probably the most common one for use in MM systems is the Y coupler. These can be made by butting together the chamfered ends of two output fibres (figure 2.22a), and then fusing them with the input fibre (figure 2.22b). The amount of optical power sent down each arm can be controlled by altering the input fibre core area seen by each output arm.

An alternative design, which allows for multi-way splitting, is the *fused biconical taper coupler* shown in figure 2.22c. In this design, the fibres are first ground, or etched, to reduce the cladding thickness, twisted together, and then fused, by heating to 1500°C, to produce an interaction region. Using this basic technique, any number of fibres can be coupled together to form a *star coupler*. Couplers are commonly supplied with bare fibre ends, for fusion splicing, or in a package with bulkhead connectors.

Single-mode couplers rely on the coupling of the evanescent field we examined in section 2.2.1. The amount of power in this field is highly dependent on the normalised frequency variable, V – a low value of V leads to a high evanescent field. The most common type of coupler is the fused biconical taper we have just discussed. As the amount of coupling is dependent upon the contact length and cladding thickness, the fibres are stretched while being heated. This stretching reduces the core diameter, and so the value of V falls. This has the effect of increasing the power in the evanescent field, so increasing the coupling. It should be noted that power can be coupled from, and to, either fibre.

An alternative coupler can be made by implanting a dielectric waveguide into a substrate (figure 2.23). The most commonly used substrate material is *lithium niobate, $LiNbO_3$*. The guides are made by diffusing titanium into the substrate. In common with the SM fibre coupler, power is transferred between the waveguides through the evanescent fields. Couplers of this type form the basis of a large family of components known as *Integrated Optics*

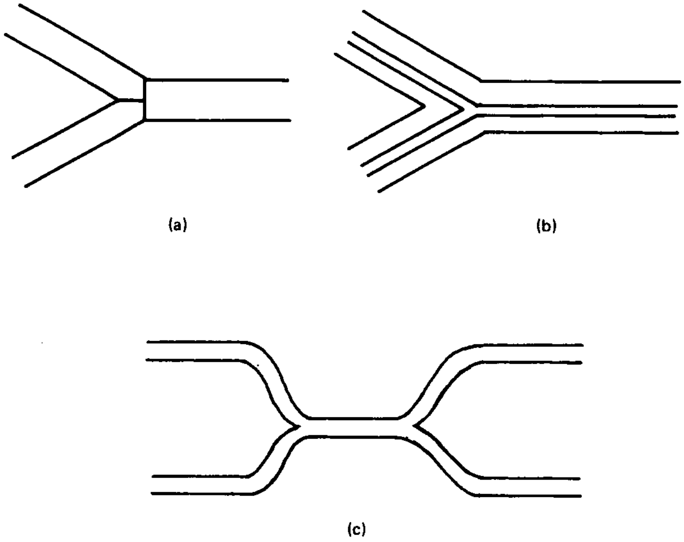


Figure 2.22 (a) Chamfered ends of input/output fibres for (b) a fused Y-coupler, (c) a fused biconical taper coupler

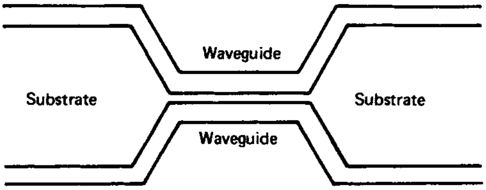


Figure 2.23 Schematic of an SM, evanescent field coupler/power splitter

which we will encounter in section 3.7.3.

An important parameter to be considered when specifying couplers is the insertion loss, or *excess loss*. This is the ratio of the total output power to the input power. Typical MM couplers have an excess loss of 3 dB, while that of SM couplers can be less than 0.5 dB. The major source of loss in couplers is the attenuation introduced by the connections, and so it is important to use low-loss connectors or fusion splices.

3 Optical Transmitters

To be useful in an optical link, a light source needs the following characteristics:

(1) it must be possible to operate the device continuously at a variety of temperatures for many years;
(2) it must be possible to modulate the light output over a wide range of modulating frequencies;
(3) for fibre links, the wavelength of the output should coincide with one of the transmission windows for the fibre type used;
(4) to couple large amounts of power into an optical fibre, the emitting area should be small;
(5) to reduce material dispersion in an optical fibre link, the output spectrum should be narrow.

We shall examine several sources that satisfy these requirements – the light emitting diode, *LED*; the semiconductor laser diode, *SLD*; solid-state and gas lasers; and fibre lasers. Before we examine these sources in greater detail, it will be useful to discuss light emission in semiconductors, and semiconductor physics in general. (Three very comprehensive references are Kressel and Butler, *Semiconductor Lasers and Heterojunction LEDs* [1]; Kressel *et al.*, chapter 2 in *Topics in Applied Physics, Vol. 39* [2]; and Casey and Panish, *Heterostructure Lasers, Parts A and B* [3].)

3.1 Semiconductor diodes

All *semiconductor* light sources use a forward biased p–n junction diode to generate light. As LEDs and SLDs are commonly used in optical fibre links, it will be useful for us to examine the physics of a semiconductor p–n junction in some detail. Let us begin our study by examining intrinsic semiconductor material, that is, material that has not been doped to either p- or n-type.

3.1.1 Intrinsic semiconductor material

Under thermal equilibrium, there will be a certain number of electrons available for conduction in the conduction band, *CB*, and a corresponding number of holes in the valence band *VB*. The density of free electrons in the CB is generally quoted in terms of the *Fermi–Dirac probability function.* This function describes the most likely distribution of electron energies as

$$F(E) = \frac{1}{1 + \exp[(E - E_f)/kT]} \tag{3.1}$$

where $F(E)$ is the probability that an electron has an energy E, and E_f is the energy level at which $F(E)$ is exactly 0.5 – the *Fermi level.* We can simplify (3.1) by noting that, in the conduction band, $E - E_f$ is generally greater than kT and so

$$F(E) \approx \exp[-(E - E_f)/kT] \tag{3.2}$$

We can check (3.2) by noting that as temperature increases, $F(E)$ tends to unity and so all the electrons are thermally excited to the CB. Conversely, if the temperature is absolute zero then $F(E)$ tends to zero and there are no electrons available for conduction.

In order to find the density of thermally excited electrons in the CB, n, we simply multiply the density of available levels, $S(E)$, by the probability of finding an electron in a particular level, and integrate over all available levels. So

$$n = \int_{E_c}^{E_t} S(E)F(E)\,\mathrm{d}E \tag{3.3}$$

where E_c is the energy level at the bottom of the CB, and E_t is the level at the top of the CB. As (3.3) tends to zero very quickly as E tends to E_t, we can take the upper level in (3.3) to be ∞ to a good approximation. Thus the integral in (3.3) results in

$$n \approx N_c \exp[-(E_c - E_f)/kT] \tag{3.4}$$

where $N_c = \dfrac{2(2\pi m_e kT)^{\frac{3}{2}}}{h^2}$

and m_e is the effective mass of an electron.

We can find the density of holes by noting that every thermally generated free electron leaves behind it a hole. Thus the probability that a level is *not filled* is $1 - F(E)$ and the density of holes is given by

$$p \approx N_v \exp[-(E_f - E_v)/kT] \tag{3.5}$$

where $N_v = \dfrac{2(2\pi m_h kT)^{\frac{3}{2}}}{h^2}$

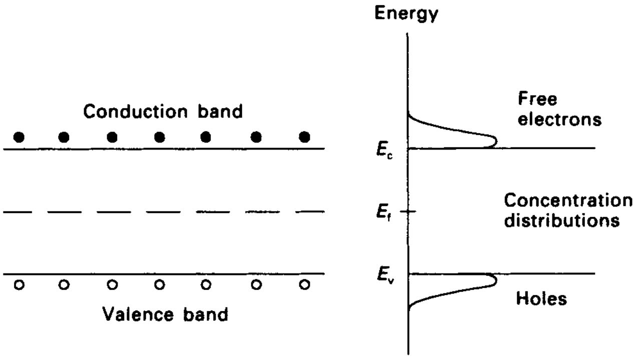


Figure 3.1 Fermi level and carrier concentration distribution in intrinsic semiconductor material

and m_h is the effective mass of a hole. Now, as we have already noted, each thermally generated electron leaves behind a hole, and so the density of electrons in the CB must be the same as the density of holes in the VB, that is, $n = p$. Thus

$$N_c \exp[-(E_c - E_f)/kT] = N_v \exp[-(E_f - E_v)/kT]$$

which, after some rearranging, becomes

$$E_f = \frac{E_g}{2} + \frac{kT}{2} \ln(N_v/N_c) \tag{3.6}$$

where E_g is the band-gap of the semiconductor material. As m_h and m_e are of the same order of magnitude, we can approximate $\ln(N_v/N_c)$ to zero. Thus we can see that the Fermi level in intrinsic semiconductor lies mid-way between the VB and CB as shown in figure 3.1. We can also find the density of carriers in intrinsic material, n_i, by noting

$$\begin{aligned} n_i^2 &= np \\ &= N_c \exp[-(E_c - E_f)/kT]\, N_v \exp[-(E_f - E_v)/kT] \\ &= N_c N_v \exp[-(E_c - E_v)/kT] \\ &= N_c N_v \exp(-E_g/kT) \end{aligned}$$

or

$$n_i = \sqrt{N_c N_v} \exp(-E_g/2kT) \tag{3.7}$$

We should note that (3.7) is independent of E_f and so it will apply equally well to intrinsic and extrinsic semiconductors. Thus we can say that the *electron–hole product in intrinsic and extrinsic semiconductor is a constant.*

Example

The effective mass of electrons in intrinsic silicon at a temperature of 300 Kelvin is 8.8 × 10^{-31} kg, and that of holes is 4.6 × 10^{-31} kg. Estimate the density of electrons in the conduction band, and of holes in the valence band. Also find the Fermi level and the total carrier density. The band-gap energy for silicon is 1.12 eV.

We can find N_c and N_v by using

$$N_c = 2\frac{(2\pi m_e kT)^{\frac{3}{2}}}{h^2} \quad \text{and} \quad N_v = 2\frac{(2\pi m_h kT)^{\frac{3}{2}}}{h^2}$$

Thus $N_c = 2.38 \times 10^{25}\ \mathrm{m}^{-3}$ and $N_v = 0.9 \times 10^{25}\ \mathrm{m}^{-3}$.

The Fermi level is given by equation (3.6) as

$$\begin{aligned} E_f &= \frac{E_g}{2} + \frac{kT}{2q}\ln(N_v/N_c) \\ &= \frac{E_g}{2} - 0.013 \\ &\approx \frac{E_g}{2} \end{aligned}$$

Thus we can see that the intrinsic carrier densities, although quite high, do not significantly affect the Fermi level. (We include the electronic charge in this derivation to give the Fermi level in electron-volts.)

We can find the total carrier density from equation (3.7). Thus

$$\begin{aligned} n_i &= \sqrt{N_c N_v}\exp(-E_g/2kT) \\ &= 1.47 \times 10^{25} \exp\left(\frac{-1.6 \times 10^{-19} \times 1.12}{2 \times 1.38 \times 10^{-23} \times 300}\right) \\ &= 7.1 \times 10^{15}\ \mathrm{m}^{-3} \end{aligned}$$

3.1.2 Extrinsic semiconductor material

Let us now turn our attention to extrinsic semiconductor material. When we dope intrinsic material with donor atoms, the donor atom density, N_d, is usually sufficient to mask the effects of thermally generated electrons. Thus $n = N_d$ and (3.4) becomes

$$N_d \approx N_c \exp[-(E_c - E_{fn})/kT] \tag{3.8}$$

from which the Fermi level is given by

$$E_{fn} = E_c + kT \ln(N_d/N_c) \tag{3.9}$$

If we dope the material with acceptor atoms, we get

$$N_a \approx N_v \exp[-(E_f - E_v)/kT] \tag{3.10}$$

and

$$E_{fp} = E_v - kT \ln(N_a/N_v) \tag{3.11}$$

Thus we can see that when we dope intrinsic material with donor atoms, the Fermi level moves towards the CB, and so there is a greater probability of finding electrons in the CB – so-called *n-type* material; whereas when we dope with acceptor atoms, the Fermi level moves towards the VB – so-called *p-type* material. This situation is shown in figure 3.2.

If we apply a voltage to an extrinsic semiconductor so that majority carriers (electrons in n-type and holes in p-type) are injected into the material, a current will flow under the influence of the applied bias. So, for the n-type material, the current density due to the drift of injected carriers (electrons) will be

$$\begin{aligned} J_{\text{ndrift}} &= nqv_d \\ &= nq\mu_n E \end{aligned} \tag{3.12}$$

where v_d is the *drift velocity* of the electrons given by $v_d = \mu_n E$, μ_n is the *electron mobility*, and E is the *electric field strength*. We can write a similar expression for the drift current due to holes being injected into the p-type material as,

$$J_{\text{pdrift}} = pq\mu_p E \tag{3.13}$$

As well as the drift of carriers throughout the material, there can also be *diffusion* of *minority* carriers down a concentration gradient. Let us consider

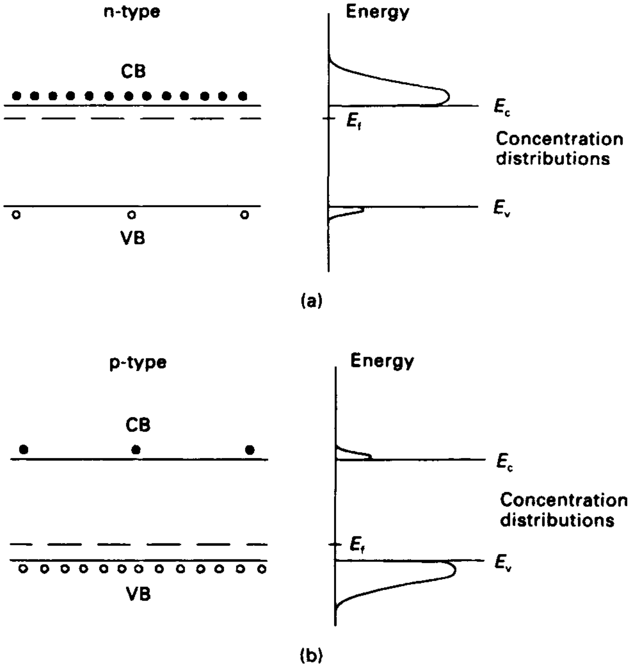


Figure 3.2 Fermi level and carrier concentration distribution in (a) n-type and (b) p-type semiconductor material

a sample of n-type material, with no applied bias. Next, let us introduce, by some means, a large number of minority carriers (holes) at the left-hand side of the sample. (The holes could appear as a result of the generation of electron–hole pairs caused by the absorption of light, see chapter 4.) These holes will tend to diffuse down a concentration gradient, away from the area where they were produced. (A similar situation occurs with the diffusion of gases.) This gives rise to a *diffusion current* given by, for holes in n-type material

$$J_{\text{diff}} = -D_p \frac{q \mathrm{d} p_n}{\mathrm{d}x} \tag{3.14}$$

and, for electrons in p-type material

$$J_{\text{diff}} = -D_n \frac{q \mathrm{d} n_p}{\mathrm{d}x} \tag{3.15}$$

where D_n and D_p are the diffusion coefficients for electrons and holes respectively. (The negative sign in these equations arises from the fact that the minority carrier concentration gradient *reduces* as distance x increases.) As these minority carriers are in a region of high majority carrier density, they will tend to recombine as they diffuse through the sample. In particular, the minority carrier density will reach the background level after one *diffusion length* – symbol L_n for electrons in p-type, and L_p for holes in n-type. (We will return to the diffusion length shortly.)

So, we have seen that the current in a block of semiconductor material can consist of drift current, due to the movement of *majority* carriers under the influence of an electric field, and diffusion current due to the diffusion of *minority* carriers down a concentration gradient. As we will see in the next section, both types of current are present in a p–n junction diode.

Example

P-type GaAs is formed by doping intrinsic material with Zn atoms at a density of 10^{24} m^{-3}. Determine the Fermi level if GaAs has E_g = 1.424 eV, N_c = 4.7 × 10^{23} m^{-3}, and N_v = 7 × 10^{24} m^{-3}.

If the sample is 1 cm long, and a voltage of 5 volts is placed across it, determine the total current density in the sample. (The mobility of electrons and holes in GaAs is 0.85 and 0.04 m^2/V s respectively.)

The intrinsic carrier density due to thermal excitation is given by

$$\begin{aligned} n_i &= \sqrt{N_c N_v}\exp(-E_g/2kT) \\ &= 2 \times 10^{12}\ \mathrm{m}^{-3} \end{aligned}$$

and so we can see that a doping density of 1 × 10^{25} m^{-3} will mask the effects of thermally generated carriers. Thus the Fermi level will be given by (equation 3.11)

$$\begin{aligned} E_{fp} &= E_v - kT\ln(N_a/N_v) \\ &= E_v - \frac{kT}{q}\ln(N_a/N_v) \\ &= E_v + 0.05 \end{aligned}$$

Again, because the band-gap is quoted in electron-volts, we must include the electronic charge in the calculations.

The drift current of the electrons will be given by (3.12). So

$$J_{\mathrm{ndrift}} = \frac{n_i^2}{N_a}\, q\mu_n E$$

$$= \left(\frac{2 \times 10^{12}}{1 \times 10^{24}}\right)^2 \times 1.6 \times 10^{-19} \times 0.85 \times \frac{5}{1 \times 10^{-2}}$$

$$= 2.72 \times 10^{-16}\ \text{A/m}^2$$

Similarly, the drift current of the holes will be, equation (3.13)

$$J_{\text{pdrift}} = N_a q \mu_h E$$

$$= 3.2 \times 10^6\ \text{A/m}^2$$

Thus we can see that the majority carrier drift current is significantly higher than the minority carrier current.

3.1.3 The p–n junction diode under zero bias

A p–n junction diode is formed by joining p- and n-type extrinsic semiconductors together. A *heterojunction* is formed if the p- and n-type materials are different, whereas we get a *homojunction* if the materials are identical. Heterojunction diodes are dealt with later: here we will examine homojunction diodes.

Figure 3.3 relates to a p–n junction diode under zero bias. As shown in figure 3.3b, the hole concentration in the p-type, where they are in a majority, is far greater than the concentration in the n-type, where they are in a minority. This gives rise to a concentration gradient down which the holes will diffuse. When these holes reach the n-type material, they will recombine with the free electrons, so consuming some majority carriers. A similar situation occurs with electrons diffusing into the p-type region. So, there will be an area either side of the junction that is depleted of carriers – the so-called *depletion region*.

Now, as holes migrate across the junction, they leave behind them negative acceptor atoms. (The acceptor atoms in the p-type are tightly bound in the crystal lattice, and so they cannot follow the holes.) This has the effect of making the depletion region in the p-type negatively charged. A similar situation causes a positively charged region in the n-type. Thus there is a charge distribution as shown in figure 3.3c. This distribution gives rise to an electric field as shown in figure 3.3e.

The direction of the electric field is such as to oppose the diffusion of carriers from both sides of the junction. So, in effect we have both diffusion and drift currents in the depletion region. However, as there is no bias across the diode, there can be no net flow of current and so the sum of drift and diffusion currents must equal zero.

Let us initially consider the flow of electrons across the junction. As we can reason from figure 3.3, the electron drift and diffusion currents act in the same direction. Thus, from (3.12) and (3.15) we get

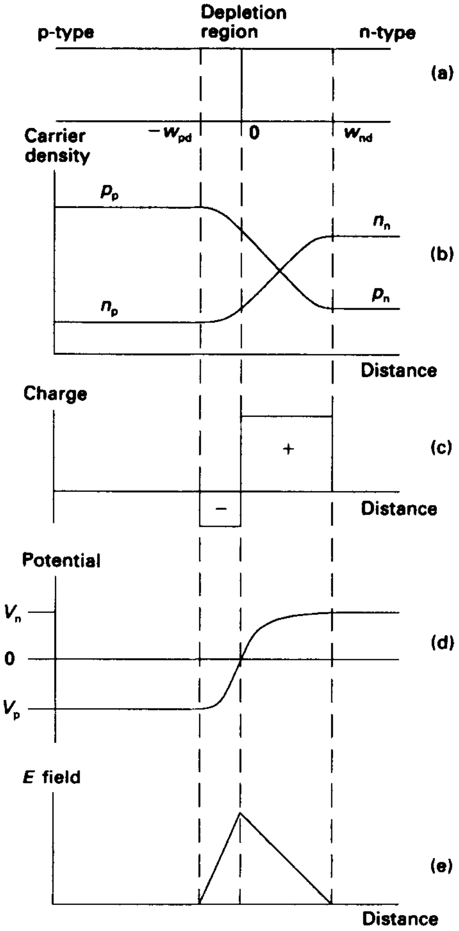


Figure 3.3 The p–n junction under zero bias: (a) schematic, (b) carrier distribution, (c) charge distribution, (d) variation in potential and (e) electric field distribution

$$J_n = -nq\mu_n E - D_n \frac{q\mathrm{d}n_p}{\mathrm{d}x} = 0 \tag{3.16}$$

and, as $E = -\mathrm{d}V/\mathrm{d}x$

$$J_n = +nq\mu_n \frac{\mathrm{d}V}{\mathrm{d}x} - D_n \frac{q\mathrm{d}n_p}{\mathrm{d}x} = 0$$

and so

$$\mathrm{d}V = \frac{D_\mathrm{n}}{\mu n}\frac{\mathrm{d}n_\mathrm{p}}{n} \tag{3.17}$$

We can integrate (3.17) to give the *barrier potential* that must be overcome before the diode will conduct. The limits of the integral will be $V_\mathrm{p} < V < V_\mathrm{n}$ and $n_\mathrm{p} < n < n_\mathrm{n}$ where n_n is the density of electrons in the n-type, and n_p the density of electrons in the p-type. So, if we integrate (3.17) we get

$$V_\mathrm{b} = V_\mathrm{n} - V_\mathrm{p} = \frac{D_\mathrm{n}}{\mu_\mathrm{n}}\ln\left(\frac{n_\mathrm{n}}{n_\mathrm{p}}\right) \tag{3.18}$$

Now, $n_\mathrm{n} = N_\mathrm{d}$ and $n_\mathrm{i}^2 = np = n_\mathrm{p}N_\mathrm{a}$, and so we can write

$$V_\mathrm{b} = \frac{D_\mathrm{n}}{\mu_\mathrm{n}}\ln\left(\frac{N_\mathrm{d}N_\mathrm{a}}{n_\mathrm{i}^2}\right)$$

Einstein's relationship states that

$$\frac{D_\mathrm{n}}{\mu_\mathrm{n}} = \frac{kT}{q}$$

and so, V_b becomes

$$V_\mathrm{b} = \frac{kT}{\mathrm{q}}\ln\left(\frac{N_\mathrm{d}N_\mathrm{a}}{n_\mathrm{i}^2}\right) \tag{3.19}$$

Thus we can see that the barrier potential is such that the n-type material is positive with respect to the p-type. The term kT/q is approximately 25 mV at room temperature, and so the barrier voltage is dependent on the doping level in the p- and n-type material.

We can find the width of the depletion region, and hence the capacitance of the diode, by noting that the total charge in the depletion region under zero bias must equal zero. Thus we can write

$$w_\mathrm{pd}N_\mathrm{a} = w_\mathrm{nd}N_\mathrm{d} \tag{3.20}$$

where w_pd and w_nd are the widths of the depletion regions in the p-type and n-type respectively. We can manipulate this equation to write the total width of the depletion region as

$$w = w_\mathrm{dn}(1 + N_\mathrm{d}/N_\mathrm{a}) \tag{3.21}$$

We should note that the depletion layer will extend unequally into the n-side if the p-type is heavily doped – a p$^+$–n diode. Such diodes are often used as avalanche photodiodes, dealt with in chapter 4.

In order to find w_{pd} and w_{nd} we need to apply Poisson's equation:

$$\frac{\mathrm{d}^2V}{\mathrm{d}x^2} = -\frac{Pt}{\epsilon} \tag{3.22}$$

where Pt is the charge density in the material, and ϵ is the permittivity of the semiconductor. Now, in the n-type material we can write

$$\frac{\mathrm{d}^2V}{\mathrm{d}x^2} = -\frac{qN_{\text{d}}}{\epsilon}$$

and so

$$\frac{\mathrm{d}V}{\mathrm{d}x} = -\frac{qN_{\text{d}}x}{\epsilon} + \text{constant} \tag{3.23}$$

We can find the value of the constant by noting that the E field is zero at the edge of the n-type depletion region. Thus

$$\frac{\mathrm{d}V}{\mathrm{d}x} = 0 \text{ at } x = -w_{\text{nd}}$$

and so the constant in (3.23) is

$$-\frac{qN_{\text{d}}}{\epsilon}w_{\text{nd}}$$

Equation (3.23) now becomes

$$\frac{\mathrm{d}V}{\mathrm{d}x} = -\frac{qN_{\text{d}}}{\epsilon}(w_{\text{nd}} + x)$$

One further integration yields

$$V_{\text{n}} = -\frac{qN_{\text{d}}}{\epsilon}\left(w_{\text{nd}}x + \frac{x^2}{2}\right) + \text{constant} \tag{3.24}$$

As we are taking the potential at the junction between the two materials to be zero, the constant in (3.24) is zero. Thus

$$V_{\text{n}} = -\frac{qN_{\text{d}}}{\epsilon}\left(w_{\text{nd}}x + \frac{x^2}{2}\right) \tag{3.25}$$

in the n-type material. By following a similar procedure with the p-type material, we get

$$V_p = \frac{qNa}{\epsilon}\left(\frac{x^2}{2} - w_{pd}x\right) \tag{3.26}$$

Thus the barrier potential is also given by

$$\begin{aligned} V_b &= V_n\big|_{x=-w_{nd}} - V_p\big|_{x=w_{pd}} \\ &= \frac{q}{2\epsilon}(N_d w_{nd}^2 + N_a w_{pd}^2) \end{aligned} \tag{3.27}$$

In order to find the width of the depletion region, we can find V_b from (3.19), and equate it to V_b from (3.27). This is done in the example at the end of this section.

Let us now turn our attention to the variation in carrier density across the depletion region. If we consider the edge of the depletion layer in the p-type, that is $x = -w_{dp}$, we can write, from (3.19)

$$\frac{qV_b}{kT} = \ln\left(\frac{N_d N_a}{n_i^2}\right)$$

and so

$$\frac{N_d N_a}{n_i^2} = \exp\left(\frac{qV_b}{kT}\right)$$

Now, $n_i^2 = n_p N_a$ at $x = -w_{dp}$, and so

$$\frac{N_d N_a}{n_p N_a} = \exp\left(\frac{qV_b}{kT}\right)$$

from which the density of minority carriers in the p-type is

$$n_p = N_d \exp(-qV_b/kT) \tag{3.28}$$

at the edge of the depletion region. By a similar procedure we can write the density of holes in the n-type as

$$p_n = N_a \exp(-qV_b/kT) \tag{3.29}$$

Thus we can see that, in crossing the depletion region, the carrier densities fall from their maximum values of N_d and N_a to n_p and p_n.

Example

A silicon p$^+$–n junction diode is formed from p-type material with N_a = 10^{24} m^{-3}, and n-type material with N_d = 10^{21} m^{-3}. Determine the barrier potential, the width of the depletion region, the maximum field strength, and the concentration of minority carriers at the depletion region boundaries. (The density of thermally generated carriers in silicon is 1.4 × 10^{16} m^{-3}, and ϵ_r = 11.8.)

From (3.19) the barrier potential is

$$V_b = \frac{kT}{q}\ln\left(\frac{N_d N_a}{n_i^2}\right)$$
$$= 25 \times 10^{-3}\ln(5.1 \times 10^{12})$$
$$= 0.73 \text{ volt}$$

This potential is the familiar voltage drop produced by a silicon diode.

In order to find the width of the depletion region, we must apply (3.27). Thus,

$$V_b = \frac{q}{2\epsilon}(N_d w_{nd}^2 + N_a w_{pd}^2)$$

and so

$$0.73 = 7.7 \times 10^{11} w_{nd}^2 + 7.7 \times 10^{14} w_{pd}^2$$

Now, as the total charge in the depletion region equals zero, we have

$$w_{pd} N_a = w_{nd} N_d$$

or

$$w_{pd} = w_{nd} \times 10^{-3}$$

Therefore

$$0.73 = 7.7 \times 10^{11} w_{nd}^2 + 7.7 \times 10^{8} w_{nd}^2$$

and so

$$w_{nd} = 1\ \mu\text{m}$$

and

$$w_{\mathrm{pd}} = 1 \text{ nm}$$

Thus we can see that the width of the depletion layer is approximately 1 μm, and it is mainly in the lightly doped n-type material.

The maximum field strength occurs at the junction between the two materials. Thus

$$\begin{aligned} E_{\max} &= -\left.\frac{\mathrm{d}V}{\mathrm{d}x}\right|_{x=0} \\ &= \frac{qN_{\mathrm{d}}}{\epsilon}w_{\mathrm{nd}} \\ &= 1.5 \text{ MV/m} \\ &= 1.5 \text{ V/μm} \end{aligned}$$

We can find the minority carrier densities by using (3.28) and (3.29) to give

$$\begin{aligned} n_{\mathrm{p}} &= 1 \times 10^{21} \exp(-qV_{\mathrm{b}}/kT) \\ &= 4.6 \times 10^{8} \text{ m}^{-3} \end{aligned}$$

and

$$\begin{aligned} p_{\mathrm{n}} &= 1 \times 10^{24} \exp(-qV_{\mathrm{b}}/kT) \\ &= 4.6 \times 10^{11} \text{ m}^{-3} \end{aligned}$$

This example has shown that if the p- and n-type doping levels are different by two orders of magnitude, the depletion region is mainly in the lightly doped part of the semiconductor. We will return to this point when we consider photodiodes in the next chapter.

3.1.4 The p–n junction diode under forward bias

In the previous section we saw that a p–n junction diode has a depletion region across the junction. Under conditions of zero bias, the drift and diffusion currents balance each other out. However, if we bias the diode by connecting the n-type to a source of electrons, the barrier potential is reduced as shown in figure 3.4. This has the effect of upsetting the balance between drift and diffusion currents, and a current will flow through the diode. All semiconductor light sources generate light under forward bias, and so an

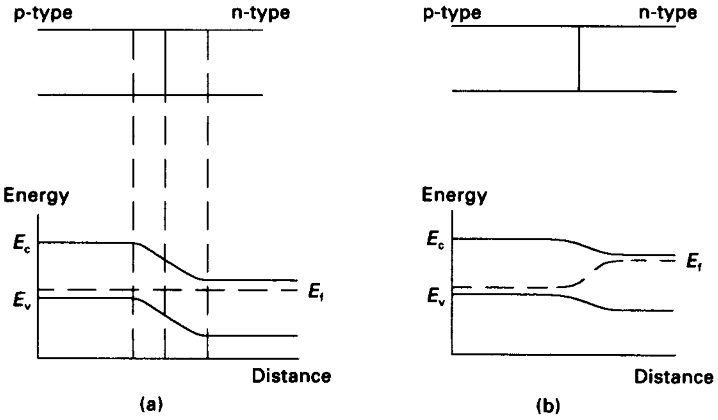


Figure 3.4 Schematic and energy diagram for a p–n junction diode under (a) zero bias and (b) forward bias

understanding of this area will help us in our later analyses. (Photodiodes operate under reverse bias, and so we will consider this in the next chapter.)

Under a forward bias of voltage V, the voltage across the depletion region will be $V_d = V_b - V$. Thus the minority carrier densities at the edges of the depletion region will become

$$n'_p = N_d \exp(-qV_b/kT)\exp(qV/kT)$$

and

$$p'_n = N_a \exp(-qV_b/kT)\exp(qV/kT)$$

By substituting from (3.28) and (3.29) we can write

$$n'_p = n_p \exp(qV/kT) \tag{3.30}$$

and

$$p'_n = p_n \exp(qV/kT) \tag{3.31}$$

These equations show that, under forward bias, the minority carrier density on either side of the depletion region increases exponentially with the bias voltage. This excess of carriers causes a large diffusion current to flow through the diode.

Let us initially consider the component of diode current caused by the diffusion of minority carriers, and ignore recombination. The minority carrier concentration in the p-type varies from a maximum of n'_p, at the edge of the depletion layer, to a minimum of n_p at the p-type contact. Similarly, in the n-type the minority carrier concentration goes from a maximum of p'_n to a minimum of p_n at the n-type contact. So, the diffusion current density will be given by

$$\begin{aligned} J &= D_n q \frac{dn}{dx} - D_p q \frac{dp}{dx} \\ &= D_n q \frac{(n'_p - n_p)}{(x_p - w_{pd})} + D_p q \frac{(p'_n - p_n)}{(x_n - w_{nd})} \end{aligned}$$

where x_p and x_n are the widths of the p- and n-type regions respectively. As the diode is forward biased, the depletion region will be very small and so w_{pd} and w_{nd} will be almost zero. So, if we make this assumption, and substitute for n'_p and p'_n, we get

$$J = J_o\{\exp(qV/kT) - 1\} \tag{3.32}$$

where

$$J_o = \frac{D_n q n_p}{x_p} + \frac{D_p q p_n}{x_n}$$

As J is the current density, the diode current will also follow the same form as (3.32).

Let us now take account of carrier recombination. Electrons injected into the p-type will recombine with holes to try to maintain thermal equilibrium. The holes that are lost through recombination are replaced by the injection of carriers from the external contact, and so there will be an electric field across the diode. Thus the current density at any point will be made up of the diffusion of minority carriers towards the external contact, J_{ndiff}, and the drift of majority carriers, J_{hr}, from the external contact. At all points in the p-type region, the total current will be constant and given by

$$J_t = J_{ndiff} + J_{hr}$$

Figure 3.5 shows the situation at a certain point in the p-type material. Carrier recombination can be described in terms of the *carrier lifetime*, τ_n. This is defined as the time taken for an injected carrier concentration to fall to 1/e times its original value. Since the total current is constant across the p-type, we can write

$$\frac{d}{dx} J_t = 0$$

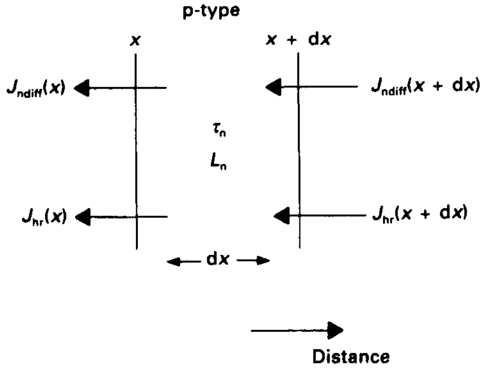


Figure 3.5 Current flow at a certain point in the p-type region of a forward biased p–n junction diode

that is

$$\frac{d}{dx}J_{ndiff} + \frac{d}{dx}J_{hr} = 0 \tag{3.33}$$

Now, the charge lost per second due to recombination in the element dx is

$$Q = \frac{q(n - n_p)}{\tau_n} A dx$$

The p-type contact must supply this charge and so

$$Q = dI_{hr}$$

that is

$$dI_{hr} = q\frac{(n - n_p)}{\tau_n} A dx$$

and so

$$\frac{d}{dx}J_{hr} = \frac{q(n - n_p)}{\tau_n} \tag{3.34}$$

As regards the diffusion current, we can write

$$\frac{\mathrm{d}}{\mathrm{d}x}J_{\mathrm{ndiff}} = -\frac{\mathrm{d}}{\mathrm{d}x}\left(qD_{\mathrm{n}}\frac{\mathrm{d}n}{\mathrm{d}x}\right)$$

$$= -qD_{\mathrm{n}}\frac{\mathrm{d}^2 n}{\mathrm{d}x^2} \qquad (3.35)$$

By substituting equations (3.34) and (3.35) into (3.33) we get

$$D_{\mathrm{n}}\frac{\mathrm{d}^2 n}{\mathrm{d}x^2} - \frac{(n - n_{\mathrm{p}})}{\tau_{\mathrm{n}}} = 0 \qquad (3.36)$$

In order to solve this equation we need to apply the boundary conditions that $n = n'_{\mathrm{p}}$ at $x = 0$ and $n = n_{\mathrm{p}}$ at $x = -x_{\mathrm{p}}$. (Here we are assuming that the width of the depletion region is negligible when compared with the diode dimensions.) So, the solution to (3.36) is

$$n(x) - n_{\mathrm{p}} = (n'_{\mathrm{p}} - n_{\mathrm{p}})\exp\left(\frac{-x}{\sqrt{D_{\mathrm{n}}\tau_{\mathrm{n}}}}\right) \qquad (3.37)$$

From equation (3.37) we can see that the excess electron density falls to 1/e times its original value in distance $\sqrt{D_{\mathrm{n}}\tau_{\mathrm{n}}}$. This distance is known as the *diffusion length*, L_{n}. By following a similar analysis with holes, we get

$$p(x) - p_{\mathrm{n}} = (p'_{\mathrm{n}} - p_{\mathrm{n}})\exp\left(\frac{-x}{\sqrt{D_{\mathrm{p}}\tau_{\mathrm{p}}}}\right) \qquad (3.38)$$

and so the diffusion length for holes is $L_{\mathrm{p}} = \sqrt{D_{\mathrm{p}}\tau_{\mathrm{p}}}$.

Example

A GaAs p$^+$n junction diode is formed from p-type material with $N_{\mathrm{a}} = 10^{24}$ m^{-3}, and n-type material with $N_{\mathrm{d}} = 10^{21}$ m^{-3}. Determine the concentration of minority carriers as a function of distance from the junction, assuming an external forward bias of 1.3 volt. What is the current density if the p- and n-type are both ten diffusion lengths long? (In GaAs, $\epsilon_{\mathrm{r}} = 13.1$, $D_{\mathrm{n}} = 22 \times 10^{-3}$ m^2/s, $D_{\mathrm{p}} = 1 \times 10^{-3}$ m^2/s, $\tau_{\mathrm{n}} = \tau_{\mathrm{p}} = 50$ ns.)

By following a similar procedure to that used in the previous example, we find that the barrier potential of this diode is 1.2 volt. Thus the equilibrium carrier densities are

$$n_{\mathrm{p}} = 5\ \mathrm{m}^{-3} \qquad \text{and} \qquad p_{\mathrm{n}} = 5 \times 10^3\ \mathrm{m}^{-3}$$

Now

$$n'_p = n_p \exp(qV/kT)$$

and

$$p'_n = p_n \exp(qV/kT)$$

and so

$$n'_p = 4.62 \times 10^{22}\ \mathrm{m^{-3}}$$

and

$$p'_n = 4.62 \times 10^{25}\ \mathrm{m^{-3}}$$

Thus the excess carrier densities are

$$\begin{aligned} n(x) - n_p &= (n'_p - n_p)\exp\left(\frac{-x}{\sqrt{D_n\tau_n}}\right) \\ &= 4.62 \times 10^{22}\exp(-30 \times 10^3 x) \end{aligned}$$

and

$$p(x) - p_n = 4.62 \times 10^{25}\exp(-140 \times 10^3 x)$$

As regards the current density, we can use (3.32) to give

$$J = \left(\frac{D_n q n_p}{x_p} + \frac{D_p q p_n}{x_n}\right)\{\exp(qV/kT) - 1\}$$

Now, x_p and x_n are both ten diffusion lengths. Thus we can write

$$x_p = 10\sqrt{D_p\tau_p} \qquad \text{and} \qquad x_n = 10\sqrt{D_n\tau_n}$$

giving

$$x_p = 70\ \mu\mathrm{m} \qquad \text{and} \qquad x_n = 330\ \mu\mathrm{m}$$

So

$$\begin{aligned} J &= 2.7 \times 10^{-15} \times 9.2 \times 10^{21} \\ &= 2.5 \times 10^{7}\ \mathrm{A/m^2} \end{aligned}$$

This example has shown that, although the diode is only just biased above the barrier potential, a large number of carriers are available for conduction, and this causes a high current density.

3.2 Light emission in semiconductors

In this section we will examine light generation in semiconductor diodes. In particular we will discuss the rate at which the semiconductor can generate light, and comment on the efficiency of certain semiconductor materials. Before we derive the rate equations, it will be useful for us to examine direct and indirect band-gap materials.

3.2.1 Direct and indirect band-gap materials

As we have already seen, when we apply forward bias to a semiconductor diode, the barrier voltage of a p–n semiconductor junction diode reduces, so allowing electrons and holes to cross the depletion region (figure 3.6). The minority carriers, electrons in the p-type and holes in the n-type, recombine by electrons dropping down from the conduction band to the valence band. This recombination results in the electrons losing a certain amount of energy equal to the band-gap energy difference, E_g.

Recombination can occur by two different processes: *indirect transitions* (also known as *non-radiative* recombinations) which produce lattice vibrations, or *phonons*; and *direct transitions* (or *radiative* recombinations) which produce *photons* of light. We can see the difference between these by examining the energy/wave number, *E–k*, diagrams of two different semiconductors. (The *E–k* diagrams are plots of electron energy against wave number which

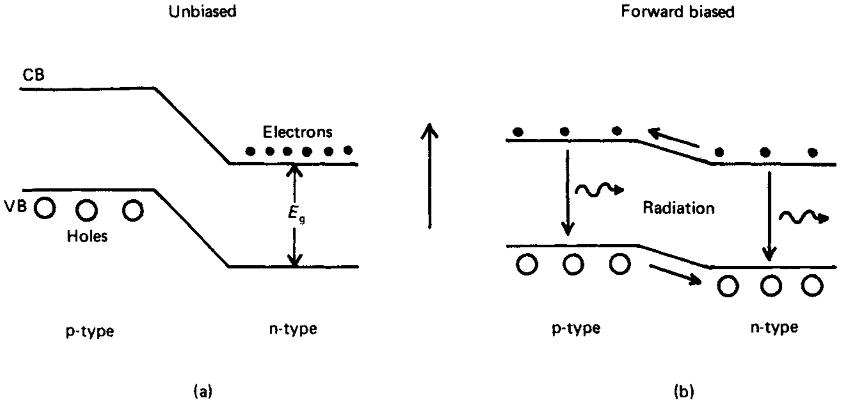


Figure 3.6 Carrier recombination in a forward biased p–n junction diode

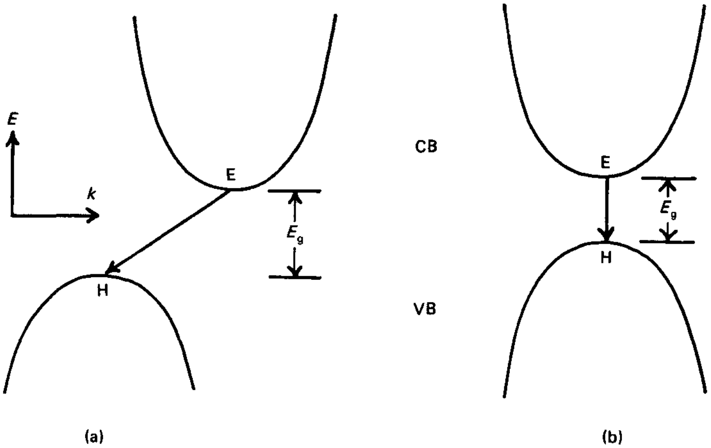


Figure 3.7 Energy/wave-number diagrams for (a) an indirect and (b) a direct band-gap semiconductor

we can regard as being proportional to the electron momentum.)

Figure 3.7a shows a simplified *E–k* diagram for an indirect band-gap material, for example silicon, *Si*, or germanium, *Ge*. As can be seen, the electron and hole momenta are different. So, if an electron drops down from region *E* in the CB to region *H* in the VB, then a change of momentum has to take place, and this results in the emission of a phonon.

Direct band-gap semiconductors can be made from compounds of elements from groups III and V of the periodic table (so-called *III–V semiconductors*). Gallium arsenide, *GaAs*, is a direct band-gap material, with an *E–k* diagram similar to that shown in figure 3.7b. As can be seen, the electron and hole momenta are the same – the regions *E* and *H* are coincident. Thus an electron dropping down from the CB to the VB does so *directly*. Under these circumstances, the energy lost is given up as a photon of light whose free-space wavelength is given by

$$\lambda_0 = \frac{hc}{qE_g} = \frac{1244}{E_g} \text{ (nm)} \qquad (3.39)$$

where h is Planck's constant, 6.624×10^{-34} J s, E_g is the band-gap in electron-volts, eV, and q is the electronic charge, 1.6×10^{-19} C. Thus, to be an efficient semiconductor light source, the LED or laser should be made of a direct band-gap material.

Table 3.1 lists the band-gap energy, and transition type, of a range of semiconductors. We can see from this that all the common single element materials have an indirect band-gap and so are never used as light sources. However, the III–V semiconductors have a direct band-gap, and so are most

Table 3.1 Characteristics of various semiconductor materials (D – direct, I – indirect band-gap)

Semiconductor material	*Transition type*	*Band-gap energy (eV)*	*Wavelength of emission* (μm)
InAs	D	0.36	3.44
PbS	I	0.41	3.02
Ge	I	0.67	1.85
GaSb	D	0.72	1.72
Si	I	1.12	1.11
InP	D	1.35	0.92
GaAs	D	1.42	0.87
CdTe	D	1.56	0.79
GaP	I	2.26	0.55
SiC	I	3.00	0.41

often used. For example, a compound of gallium, aluminium and arsenic has a band-gap of between 1.38 and 1.55 eV, resulting in light of wavelength in the region 900–800 nm. The 100 nm spread in wavelength occurs because E_g depends on the ratio of Ga to Al. (We will return to this point presently.)

3.2.2 Rate equations

The previous section introduced the basic principle of photon generation in semiconductor diodes. However, we have not yet examined the rate of light generation, and this is the subject that will now be considered. To simplify the following analysis we will initially assume that the electron and hole densities either side of the junction are constant.

Photon generation occurs by electrons in the CB recombining with holes in the VB. As we saw when we considered the zero biased diode, recombination of carriers can occur on both sides of the junction. If the material is a direct band-gap one, each carrier recombination will, ideally, produce a photon. Thus we can see that the rate of photon generation will depend on several factors: the density of electrons in the CB, n; the density of holes in the VB, p; and absorption by the material. Taking account of these factors, we can write the rate of photon generation, $\mathrm{d}\phi/\mathrm{d}t$, as

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = anp - b\phi \tag{3.40}$$

where a and b are constants relating to photon generation and absorption respectively, and ϕ has units of photons/m^3.

Now, each photon consumes one electron–hole pair, *EHP*, and so we can write the rate of decrease in electron density as

$$\frac{dn}{dt} = -anp + b\phi \tag{3.41}$$

Equations (3.40) and (3.41) are non-linear in form. However, we can generate an approximate solution by considering the steady-state solution, and then introducing a transient.

In equilibrium, the rate of photon generation equals the rate of change of electron density. So,

$$an_e p_e = b\phi_e \tag{3.42}$$

If we disturb the equilibrium with an injection of electrons, (3.41) becomes

$$\frac{d}{dt}(n_e + \delta n) = -a(n_e + \delta n)\,(p_e + \delta p) + b(\phi_e + \delta\phi) \tag{3.43}$$

If we use (3.42), and note that δn equals δp, we get

$$\frac{d}{dt}\delta n = -a\delta n\,(n_e + p_e) + b\delta\phi \tag{3.44}$$

where we have assumed that $\delta n \delta p \approx 0$. The first term in (3.44) is the rate of change of electron density due to photon generation, which we can write as

$$\frac{d}{dt}\delta n = -\frac{\delta n}{\tau_r} \tag{3.45}$$

where τ_r is the *radiative recombination time* defined by

$$\tau_r = \frac{1}{a\,(n_e + p_e)} \tag{3.46}$$

Unfortunately, not all EHPs contribute to the light output; some electrons may fall into traps in the crystalline structure so generating phonons rather than photons. Thus we need to introduce *non-radiative recombination.* These two recombination rates will add together, and so we can modify (3.45) to give

$$\frac{d}{dt}\delta n = -\frac{\delta n}{\tau_r} - \frac{\delta n}{\tau nr}$$

or

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\delta n = -\,\frac{\delta n}{\tau_{\mathrm{n}}} \tag{3.47}$$

where τ_{n} is the *electron* recombination time given by

$$\frac{1}{\tau_{\mathrm{n}}} = \frac{1}{\tau_{\mathrm{r}}} + \frac{1}{\tau_{\mathrm{nr}}} \tag{3.48}$$

(This τ_{n} is identical to the recombination time for electrons that we introduced in section 3.1.4.)

The solution to (3.47) is an exponential given by

$$\delta n(t) = \delta n(0)\exp(-t/\tau_{\mathrm{n}}) \tag{3.49}$$

We can now write the electron rate equation, (3.44), as

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\delta n = -\frac{\delta n}{\tau_{\mathrm{n}}} + b\delta\phi \tag{3.50}$$

By following a similar argument with the photon rate equation, (3.40) can be written as

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{\delta n}{\tau_{\mathrm{r}}} - b\delta\phi \tag{3.51}$$

From these two equations we can see that an increase in photon density will increase the electron density (equation 3.50) and this will increase the photon density (equation 3.51). Thus (3.50) and (3.51) are intimately linked.

Let us now consider *forward bias operation.* As we saw in section 3.1.4, a current will flow due to EHP recombination when we forward bias the diode. Thus we can modify the electron rate equation to

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\delta n = \frac{1}{q}\,\frac{\mathrm{d}J}{\mathrm{d}x} - \frac{\delta n}{\tau_{\mathrm{n}}} + b\delta\phi \tag{3.52}$$

This equation describes the variation in electron density due to carrier injection, EHP recombination and photon absorption. If we consider photon generation due to high injection currents, we can effectively neglect photon absorption. The rate equations then become

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{\delta n}{\tau_{\mathrm{r}}} \tag{3.53}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\delta n = \frac{1}{q}\,\frac{\mathrm{d}J}{\mathrm{d}x} - \frac{\delta n}{\tau_\mathrm{n}} \tag{3.54}$$

Thus we can see that an increase in current causes an increase in photon density, so producing light. This is the principle behind light emitting diodes, *LEDs*, which are dealt with in detail later.

We should also note that the injected carriers will modify the radiative recombination time, (3.46), to

$$\tau_\mathrm{r} = \frac{1}{a(n_\mathrm{e} + p_\mathrm{e} + \delta n)} \tag{3.55}$$

We now have two regions of interest: under high current injection, $\delta n >> (n_\mathrm{e} + p_\mathrm{e})$ and τ_r depends on the injected carrier concentration; under low current injection, $\delta n << (n_\mathrm{e} + p_\mathrm{e})$ and τ_r is independent of the injected carriers.

We can determine the efficiency of any particular light source by finding the ratio of radiative recombination rate to the total recombination rate. So, the efficiency, η, of a light source is

$$\begin{aligned}\eta &= \frac{\delta n/\delta\tau_\mathrm{r}}{\delta n/\delta\tau_\mathrm{n}} \\ &= \frac{\tau_\mathrm{n}}{\tau_\mathrm{r}} \\ &= \frac{\tau_\mathrm{nr}}{\tau_\mathrm{r} + \tau_\mathrm{nr}}\end{aligned} \tag{3.56}$$

Thus we can see that, in order to be an efficient light source, the non-radiative recombination lifetime must be far higher than the radiative recombination lifetime. We will return to this point when we consider LEDs and lasers.

Example

A semiconductor light source is made from the p$^+$n GaAs diode described in the previous example. The voltage across the diode is pulsed from 0.9 volt to 1.2 volts. Determine the optical power generated by the device. (Assume a cross-sectional area of 4 × 10^{-9} m^{-2}.)

By following a similar analysis to that used in the previous example, we find

$$n(x) - n_{\mathrm{p}} = 8.06 \times 10^{15} \exp(-30 \times 10^{3}x)\ \mathrm{m}^{-3}$$

$$p(x) - p_{\mathrm{n}} = 8.06 \times 10^{18} \exp(-140 \times 10^{3}x)\ \mathrm{m}^{-3}$$

for a diode bias of 0.9 V, and

$$n(x) - n_{\mathrm{p}} = 9.5 \times 10^{20} \exp(-30 \times 10^{3}x)\ \mathrm{m}^{-3}$$

$$p(x) - p_{\mathrm{n}} = 9.5 \times 10^{23} \exp(-140 \times 10^{3}x)\ \mathrm{m}^{-3}$$

for a bias of 1.2 V.

It should be evident from these calculations that the injected carrier density is far larger than the equilibrium carrier density.

Now, the carrier density is a function of distance from the junction. So, we need to carry this variation with x throughout our calculations. As we are operating in the high injection regime, the radiative recombination time is

$$\tau_{\mathrm{r}} = \frac{1}{a\delta n}$$

The constant a is of the order of 10^{-16} m^3/s, and so

$$\tau_{\mathrm{r}} = 10.5 \times 10^{-6} \exp(30 \times 10^{3}x) \quad \text{for electrons, and}$$

$$\tau_{\mathrm{r}} = 10.5 \times 10^{-9} \exp(140 \times 10^{3}x) \quad \text{for holes.}$$

Now, by using (3.53), we find

$$\begin{aligned}\frac{\mathrm{d}\phi}{\mathrm{d}t} &= \frac{\delta n}{\tau_{\mathrm{r}}} \\ &= \frac{9.5 \times 10^{20} \exp(-30 \times 10^{3}x)}{10.5 \times 10^{-6} \exp(30 \times 10^{3}x)} \\ &= 9.05 \times 10^{25} \exp(-60 \times 10^{3}x)\ \mathrm{m}^{3}/\mathrm{s} \text{ for electrons,}\end{aligned}$$

and

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = 9.05 \times 10^{31} \exp(-280 \times 10^{3}x)\ \mathrm{m}^{3}/\mathrm{s} \text{ for holes.}$$

In order to find the rate of photon generation per unit area, we need to integrate these flux densities with respect to x. So, for electrons we have

$$\frac{d\phi}{dt} = \frac{-9.05 \times 10^{25}}{60 \times 10^{3}} \left| \exp(-60 \times 10^{3}) \right|_{0}^{70\ \mu m}$$

$$= 1.5 \times 10^{21}\ m^{2}/s$$

while for the holes we have

$$\frac{d\phi}{dt} = \frac{-9.05 \times 10^{31}}{280 \times 10^{3}} \left| \exp(-280 \times 10^{3}) \right|_{0}^{330\ \mu m}$$

$$= 3.2 \times 10^{26}\ m^{2}/s$$

We also need to take account of the efficiency of the device. As we have already seen

$$\eta = \frac{\delta n/\delta \tau_{r}}{\delta n/\delta \tau_{n}}$$

$$= \frac{\tau_{n}}{\tau_{r}}$$

and so the efficiency of electron recombination is

$$\eta(x) = \frac{50 \times 10^{-9}}{10.5 \times 10^{-6} \exp(30 \times 10^{3} x)}$$

On integrating this efficiency across the p-type, we get

$$\eta = 1.6 \times 10^{-7} \text{ for the electrons.}$$

By following a similar procedure with hole recombination in the n-type, we get

$$\eta = 3.4 \times 10^{-5} \text{ for the holes.}$$

Hence the total photon generation rate per unit area is

$$\frac{d\phi}{dt} = (1.5 \times 10^{21} \times 1.6 \times 10^{-7}) + (3.2 \times 10^{26} \times 3.4 \times 10^{-5})$$

$$= 2.4 \times 10^{14} + 1.1 \times 10^{22}$$

$$= 1.1 \times 10^{22}\ m^{2}/s$$

As the area of the device is 4×10^{-9} m^{2}, we get a photon generation rate of 4.4 × 10^{13}. Each photon carries energy of *hf* Joules, and so the power generated by the diode is 10 μW.

This example has shown that, for a p^+n diode, it is the hole recombination in the n-type that is the dominant light generating mechanism. We have also seen that the efficiency of the device is very low.

In the next section we will consider heterojunction diodes – the most widely used type of semiconductor diode for light generation.

3.3 Heterojunction semiconductor light sources

As we have seen in the previous section, light emission can occur on both sides of the p–n junction. We also saw that the efficiency of the diode was very low. However, if we concentrate the recombining carriers to a small active area, the light output will increase, and we can launch more power into a fibre. We can achieve such confinement by forming a junction between two dissimilar band-gap material – a *heterojunction* – which results in certain carriers experiencing a potential step, so inhibiting them from travelling farther through the lattice. In order to confine both holes and electrons, we must use two heterojunctions, the so-called *double-heterojunction*, or *DH* structure. Although most LEDs and lasers use this structure, we will initially examine a single heterojunction, or *SH*, diode.

Figure 3.8 shows the energy diagram of an SH diode. This particular diode is made of wide band-gap Ga$_{0.8}$Al$_{0.2}$As, and narrow band-gap GaAs. (The numerical subscripts refer to the proportions of the various elements that make up the alloy.) Such diodes are normally called P–n, or N–p, where the capital letter denotes the material with the higher band-gap. (The most

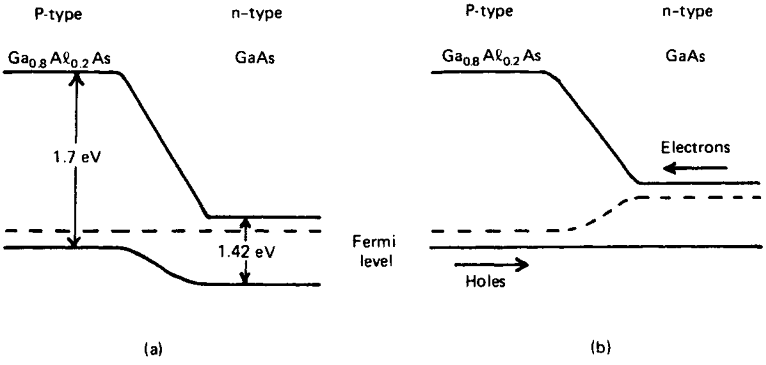


Figure 3.8 Energy diagram of a heterojunction under (a) zero bias and (b) forward bias

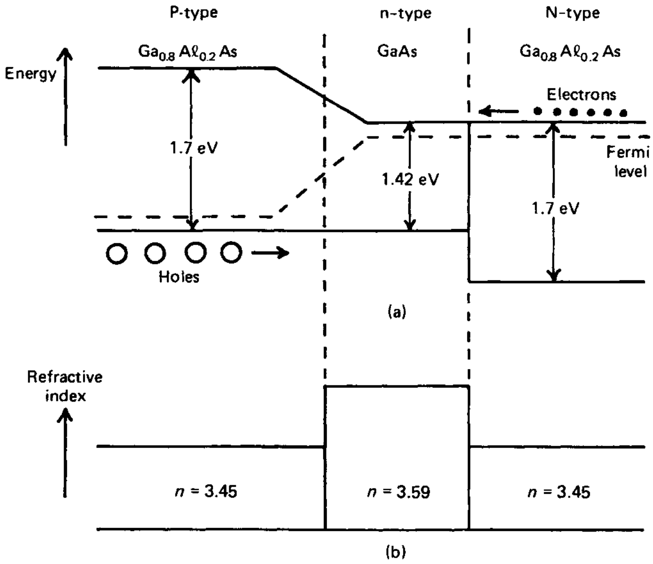


Figure 3.9 (a) Energy diagram and (b) refractive index profile for a forward biased P–n–N, double heterojunction diode

widely used dopants are sulphur, *S*, for n-type and zinc, *Zn*, for p-type.) We can see from the diagram that the potential step for holes, δE_v, is lower than the potential step for electrons, δE_c. This is more obvious when the diode is under forward bias, figure 3.8b. So, under forward bias, injected holes travel into the n-type region, but electrons cannot cross into the P-type. Hence there are a great number of holes in the GaAs n-type, and these recombine within a diffusion length of the junction. This area is known as the *active region* and, as the recombination occurs in GaAs, it generates 870 nm wavelength light.

A double heterojunction, *DH*, structure will confine both holes and electrons to a narrow active layer. As figure 3.9 shows, the potential steps either side of the active region, the GaAs, inhibit carrier movement. Thus, under forward bias, there will be a large number of carriers injected into the active region where they are effectively confined. Carrier recombination occurs in this small active layer so leading to an efficient device. An additional advantage of the DH structure is that the refractive index of the active region is greater than that of the surrounding material. Hence light emission occurs in an optical waveguide, which serves to narrow the output beam.

GaAs emits light at 870 nm; however the first optical window occurs at 850 nm. The addition of aluminium to the GaAs layer causes the band-gap, and hence the emission wavelength, to change. Hence diodes for the first window are commonly made of an $Al_xGa_{1-x}As$ active layer, surrounded by

$Al_yGa_{1-y}As$, with $y > x$. This alloy is a direct band-gap semiconductor for $x < 0.37$. If $0 < x < 0.45$, we can find E_g from the following empirical relationship.

$$E_g = 1.42 + 1.25x + 0.27x^2 \tag{3.57}$$

As it is the active layer that emits the light, the surrounding material can be an indirect band-gap semiconductor. As an example, a diode with $x = 0.03$ and $y = 0.2$ will emit light of wavelength 852 nm. We can find the refractive index of the material from

$$n = 3.59 - 0.71x \qquad \text{for } 0 < x < 0.45 \tag{3.58}$$

For operation in the second and third transmission windows, 1.3 and 1.55 μm, the diode is usually made of an indium–gallium–arsenide–phosphide alloy, $In_{1-x}Ga_xAs_yP_{1-y}$, surrounded by indium phosphide, *InP*. To ensure that the active region is a direct band-gap material, x should be lower than 0.47 and, in order to match the active layer alloy to the InP crystal lattice, $y \approx 2.2x$. With these values of x and y, we can estimate the active region band-gap from another empirical relationship

$$E_g = 1.35 - 1.89x + 1.48x^2 - 0.56x^3 \tag{3.59}$$

with the refractive index being given by

$$n^2 = 9.6 + 4.52x - 37.62x^2 \tag{3.60}$$

As an example, $In_{0.74}Ga_{0.26}As_{0.56}P_{0.44}$ has a band-gap energy of 0.95 eV, which results in an emission wavelength of 1.3 μm.

3.4 Light emitting diodes (LEDs)

At present there are two main types of LED used in optical fibre links: the *surface emitting* LED; and the *edge emitting* LED, or *ELED*. Both devices use a DH structure to constrain the carriers and the light to an active layer. Table 3.2 compares some typical characteristics of the two LED types. From this table we can see that ELEDs are superior to surface emitters in terms of coupled power, and maximum modulation frequency. For these reasons, surface emitters are generally used in short-haul, low data-rate links, whereas ELEDs are normally found in medium-haul routes. (Lasers are normally used in long-haul routes.) We should note that LEDs emit light over a wide area. Thus these devices can usually only couple useful amounts of power into large numerical aperture, MM fibres.

Table 3.2 Comparison of surface and edge emitting LED characteristics

LED type	*Maximum modulation frequency (MHz)*	*Output power (mW)*	*Fibre coupled power (mW)*
Surface emitting	60	<4	<0.2
Edge emitting	200	<7	<1.0

3.4.1 Surface emitting LEDs

Figure 3.10 shows the structure of a typical surface emitting LED. The DH diode is grown on an N-type substrate, at the top of the diode, which has a circular well etched into it. In this particular design, the light produced by the active region travels through the substrate and into a large core optical fibre held in place by epoxy resin. Some designs dispense with the fibre entirely, preferring to rely on the LED package to guide the light.

At the back of the device is a gold heatsink which, apart from a small circular contact, is insulated from the diode. This heatsink forms one of the contacts, and so all the current flows through the hole in the insulating layer. The current flows through the P-type material and forms a small, circular active region, with a typical current density of 2000 A/cm^2. This results in the production of an intense beam of light.

The refractive index change across the heterojunctions serves to constrain some of the emitted light to the active region. This light is either absorbed, or finally emitted in an area greater than the fibre core. Hence the actual amount of light coupled into the fibre is considerably less than that emitted

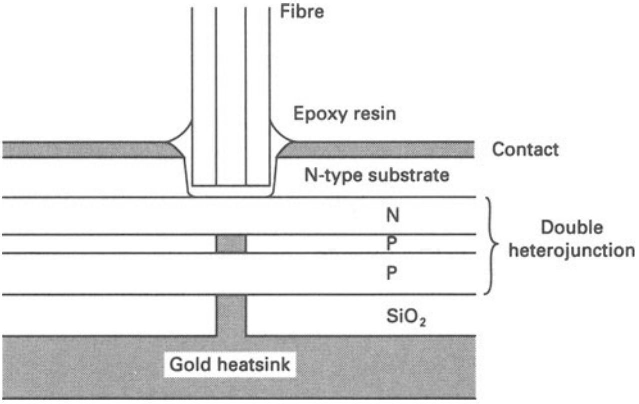


Figure 3.10 Cross-section through a typical surface emitting LED

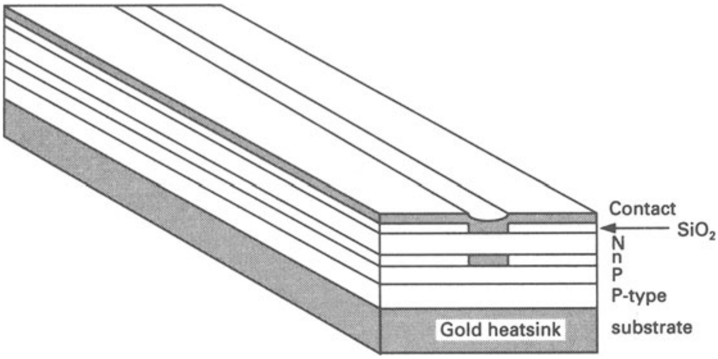


Figure 3.11 Structure of an edge-emitting, N–n–P, double heterojunction, stripe-contact LED

by the LED. Although a micro-lens placed in the well at the top of the device will increase the coupled power, the efficiency of this arrangement is dependent on the correct truncation of the lens and the fibre alignment. In practice the launched power is two to three times that achieved by an equivalent butt-coupled LED.

3.4.2 Edge emitting LEDs (ELEDs)

In order to reduce the losses caused by absorption in the active layer, and make the beam more directional, we can take the light from the edge of the LED. Such a device is known as an *edge emitting LED*, or *ELED*, and a typical structure is shown in figure 3.11.

As can be seen, the narrow stripe on the upper contact defines the shape of the active region. As the heterojunctions act to confine the light to this region, the output is more directional than from a surface emitting device, and this leads to a greater launch power. A further increase in output power results from the use of a reflective coating on the far end of the diode.

3.4.3 Spectral characteristics

As we saw in section 3.2, light emission is due to electrons randomly crossing the band-gap, so-called *spontaneous emission* of light. In practice, the conduction and valence bands consist of many different energy levels (figure 3.12). It is therefore possible for recombinations to occur across a wide range of energy differences. The distribution in electron densities peaks at an energy of approximately $E_g + kT/2$, and that of the hole densities at an energy of approximately $E_g - kT/2$. Thus the energy difference has a mean of $E_g + kT$, and a deviation of δE_g which is typically between kT and $2kT$.

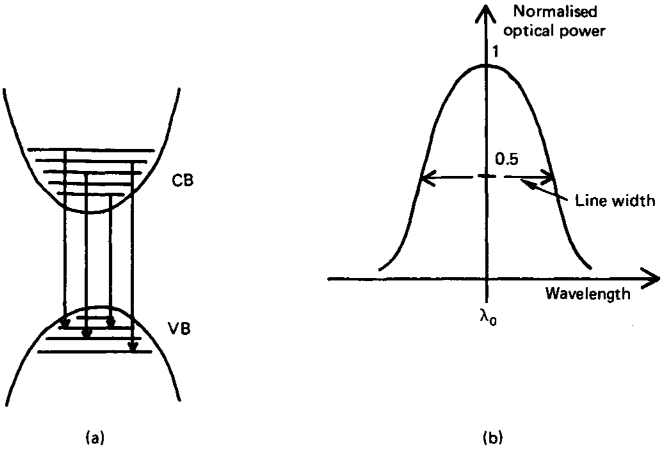


Figure 3.12 (a) Photon emission from conduction band energy levels and (b) resultant spectral characteristic

Although the actual deviation is dependent on the amount of impurity doping, the approximation will suit our purposes.

The spread in recombination energy results in a spread of emitted wavelengths about a nominal peak, as shown in figure 3.12b. The half power wavelength spread is known as the *source linewidth* and, as we saw in the previous chapter, a large linewidth will result in considerable material dispersion. However, LEDs can launch a large number of modes into the fibre, and so modal dispersion is usually dominant. In most LEDs, the linewidth is typically 30 nm which translates to a frequency spread of approximately 1.3×10^{13} Hz! Clearly LEDs are not the optical equivalent of an r.f. oscillator; they are, however, useful for simple intensity modulation such as that used in analogue and PCM links.

3.4.4 Modulation capabilities and conversion efficiency

The output power/drive current characteristic of an LED is approximately linear. If we superimpose an a.c. signal on to a d.c. bias level, we can write the output optical power, $p(\omega)$, as

$$p(\omega) = \frac{p(0)}{(1 + (\omega\tau)^2)^{\frac{1}{2}}} \tag{3.61}$$

where $p(0)$ is the unmodulated power output, and τ is the time constant of the LED and drive circuit. When we considered optical fibre bandwidth, we saw that a 3 dB drop in optical power corresponds to a 6 dB drop in electrical power. Therefore the 3 dB *electrical* bandwidth of the LED is 1/2 $\pi\tau$ Hz.

With careful design of the drive circuit, the dominant time constant will be that of the LED. This is governed by the recombination time of the carriers in the active region, τ_n. As we have already seen, when both radiative and non-radiative recombinations are present, τ_n is given by

$$\frac{1}{\tau_n} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}} \tag{3.62}$$

where τ_r and τ_{nr} are the radiative and non-radiative recombination times respectively. These time constants also give us a measure of the diode conversion efficiency, which we briefly examined in section 3.2.2. The *internal quantum efficiency*, η_{int}, is given by

$$\eta_{int} = \frac{\tau_{nr}}{\tau_{nr} + \tau_r} \tag{3.63}$$

So in order to produce a fast device, both τ_r and τ_{nr} should be kept low, with the proviso that $\tau_{nr} >> \tau_r$ in order to keep the efficiency high.

Let us now return to the electron rate equation, previously given as (3.54)

$$\frac{d}{dt}\,\delta n = \frac{1}{q}\,\frac{dJ}{dx} - \frac{\delta n}{\tau_r}$$

If we assume that the semiconductor is lightly doped, and we consider the steady state condition, that is we apply a current of J which injects an electron density of δn, we get

$$0 = \frac{1}{q}\,\frac{J}{d} - \frac{\delta n}{\tau_r}$$

and so

$$J \approx \frac{q\delta n d}{\tau_r} \tag{3.64}$$

where d is the distance between the heterojunctions, and we have temporarily ignored non-radiative recombination. Now, under high levels of injection, the radiative recombination time, equation (3.55), becomes

$$\tau_r = \frac{1}{a\delta n}$$

and so (3.64) becomes

$$J = \frac{qd}{a\tau_r^2}$$

Hence

$$\tau_r \propto \left[\frac{d}{J}\right]^{\frac{1}{2}} \tag{3.65}$$

From (3.65) we can see that in order to reduce the radiative recombination time, and so produce a more efficient device, we should operate with high current densities.

Let us now take account of non-radiative recombinations. When a heterojunction diode is formed, there is a slight mismatch between the heterojunction crystal lattices. This introduces traps at the interface between the two materials, characterised by the *surface recombination velocity, S*. Thus

$$\tau_{nr} \propto \frac{d}{S} \tag{3.66}$$

As τ_r and τ_{nr} are dependent on d, a smaller d will result in lower time constants. Unfortunately, a reduction in d causes τ_{nr} to fall faster than τ_r, and so the modulation speed increases at the expense of the efficiency. However, τ_r is inversely proportional to $\sqrt{J}$, and so we could reduce τ_r by increasing the current density. The problem with this is that a high current density causes difficulties with heatsinking, which tends to impair the device lifetime.

From (3.55) we can see that for high doping levels, $>10^{24}$ m^{-3}, τ_r is inversely proportional to the doping level. So we could reduce τ_r by increasing the doping. Unfortunately, this tends to increase the number of non-radiative recombination centres, and so τ_{nr} will also reduce. Therefore there is a trade-off between the modulation bandwidth and the LED efficiency. Most LEDs operate with high doping levels, and current state-of-the-art devices have a typical internal quantum efficiency of 50 per cent. In spite of this, the external efficiency (a measure of the launch power into a fibre) is typically less than 10 per cent, and so LEDs are generally low power devices.

3.5 Semiconductor laser diodes (SLDs)

Unlike LEDs, which emit light spontaneously, lasers produce light by *stimulated emission*. Stimulated emission occurs when a photon of light impinges on an already excited atom and, instead of being absorbed, the incident photon

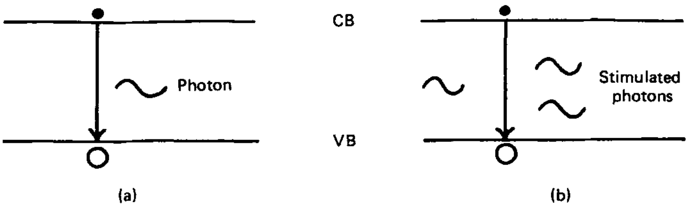


Figure 3.13 Light generation by (a) spontaneous emission and (b) stimulated emission

causes an electron to cross the band-gap, so generating another photon (figure 3.13). The stimulated photon has the same frequency and phase as the original, and these two generate more photons as they travel through the lattice. In effect, the lattice amplifies the original photon; indeed, the acronym laser stands for Light Amplification by the Stimulated Emission of Radiation. As the generated photons are all in phase, the light output is coherent and has a narrow linewidth.

Before stimulated emission can occur, the CB must contain a large number of electrons, and the VB a large number of holes. This is a *quasi-stable* state known as a *population inversion.* It results from the injection of a large number of carriers into a heavily doped, ELED active layer. If a population inversion is present then, by virtue of the light confinement from the heterojunctions, some stimulated emission occurs. However, in order to ensure that it is the dominant light generating process, we must provide some additional optical confinement.

In a laser diode, the extra confinement results from cleaving the end faces so that they form partial reflectors, or *facets.* The resulting structure, known as a *Fabry–Perot etalon,* is shown in schematic form in figure 3.14. The

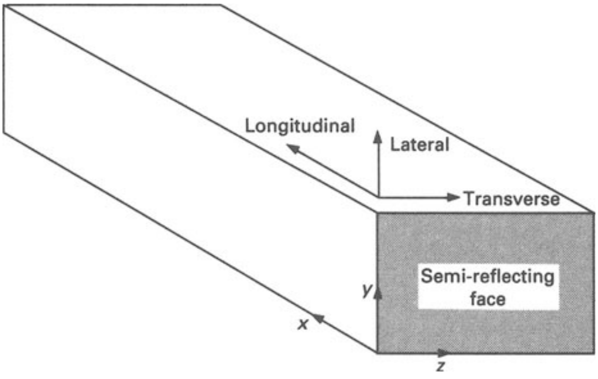


Figure 3.14 A basic Fabry–Perot cavity

facets reflect some of the spontaneously emitted light back into the active region, where it causes stimulated emission and hence gain. So, provided the optical gain in the cavity exceeds the losses, stimulated emission will be dominant.

3.5.1 Stimulated emission

Laser diodes and LEDs differ in several ways: a laser diode requires the application of a constant current to maintain stimulated emission; the output beam is more directional; and the response time is faster. In this section we will examine the simple stripe contact laser, which is similar in construction to a stripe contact ELED. We will deal with other SLD structures later. We begin our study by examining the optical gain that stimulated emission produces. We will then go on to study the spectral characteristics of SLDs.

It should be evident that, because light emission occurs in a rectangular cavity, propagation can occur along all three axes; *longitudinal*, *transverse* and *lateral* propagation. Let us initially consider a longitudinal TE wave given as $E(x, t)$. If we neglect the effects of the cavity side-walls, and assume that the cavity confines all of the E field, then we can write $E(x, t)$ as

$$E(x, t) = |E| \exp(-\alpha x/2)\exp j(\omega t - \beta_1 x) \tag{3.67}$$

where α is the attenuation of the optical *power* per unit length (hence the factor 1/2) and β_1 is the phase constant in the active region. So the field just to the right of the mirror at $x = 0$, is

$$E(0, t) = |E| \exp(0)\exp j\omega t \tag{3.68}$$

Now, when the field undergoes a round-trip of distance $2L$, it is reflected off both mirrors, and amplified by stimulated emission. Thus after one round-trip, the travelling field, E_r, is

$$E_r(0, t) = \sqrt{R_1 R_2}|E| \exp[(g-\alpha)L] \exp j(\omega t - 2\beta_1 L)] \tag{3.69}$$

where R_1 and R_2 are the *reflectivity* of the mirrors at $x = 0$ and $x = L$ respectively, and g is the *power* gain per unit length. (The reflectivity is defined as the ratio of the reflected to the incident power, hence the presence of the square root.) For amplification to occur, the magnitude of the reflected wave must be greater than that of the original wave, that is

$$\sqrt{R_1 R_2}|E| \exp[(g-\alpha)L] \geq |E|$$

Therefore the optical gain for lasing is given by

$$g \geq \alpha + \frac{1}{2L} \times \ln\left(\frac{1}{R_1 R_2}\right) \qquad (3.70)$$

As we have already noted, it is the current density in the active region, J, that produces the population inversion, and hence the cavity gain. In order to determine the relationship between the gain and current density, we must consider the rate equations for a SLD. A full analysis of stimulated emission in laser diodes would involve us in the black-body radiation law, and quantum mechanics in general. This is beyond the scope of this text; instead we will quote necessary results where needed.

As we saw when we considered light emission in semiconductors, electrons dropping down from the CB to the VB generate photons. There are two ways in which this can occur: spontaneous emission and stimulated emission. If we have a VB electron density of n_2, we can express the rate of electron density decay due to *spontaneous* recombination as

$$\left.\frac{\mathrm{d}n}{\mathrm{d}t}\right|_{\mathrm{spon}} = \frac{n_2}{\tau_{\mathrm{sp}}} = A_{21} n_2 \qquad (3.71)$$

where A_{21} is a constant with units of s^{-1}.

Now, the semiconductor will absorb some of the light generated by spontaneous emission. As we have seen when we considered light emission in an LED, spontaneous emission generates light with a spread of wavelength. Thus we have light of frequency f_0 and half-power spread δf propagating through the SLD. One way of finding δf is to excite the SLD material with light of varying wavelength, and plot the variation of absorption with frequency. The result is a Lorentzian curve, $g(f)$, given by

$$g(f) = \frac{\delta f}{2\pi[(f - f_0)^2 + (\delta f/2)^2]} \qquad (3.72)$$

where we have used the following normalisation

$$\int_{-\infty}^{\infty} \mathrm{g}(f)\mathrm{d}f = 1 \qquad (3.73)$$

Thus the units of g(f) are 1/Hz or seconds.

Some of the light generated by spontaneous emission will be absorbed by the SLD material. This will have the effect of increasing the electron density in the VB at a rate

$$\left.\frac{\mathrm{d}n}{\mathrm{d}t}\right|_{\mathrm{abs}} = B_{12}\, \phi\, hf\, n_1\, g(f) \qquad (3.74)$$

where B_{12} is another constant with units of m^3/J s^2, ϕ is the optical flux density (photons/m^3), and n_1 is the electron density in the CB. (We have included the factor hf because we are considering photon density.)

Electrons will also be lost due to stimulated recombination. So

$$\left.\frac{\mathrm{d}n}{\mathrm{d}t}\right|_{\text{stim}} = B_{21}\ \phi\ hf\ n_2\ g(f) \tag{3.75}$$

The constants A_{21}, B_{12} and B_{21} are known as the Einstein A and B coefficients. (Albert Einstein was the first person to suggest the possibility of stimulated emission.) If the semiconductor is in equilibrium, we can write

$$\left.\frac{\mathrm{d}n}{\mathrm{d}t}\right|_{\text{abs}} = \left.\frac{\mathrm{d}n}{\mathrm{d}t}\right|_{\text{spon}} + \left.\frac{\mathrm{d}n}{\mathrm{d}t}\right|_{\text{stim}}$$

or

$$B_{12}\phi hfn_1g(f) = A_{21}n_2 + B_{21}\phi hfn_2g(f)$$

Quantum mechanics predicts $B_{12} = B_{21} = B$, and that

$$\frac{A}{B} = \frac{8\pi hf^3}{v_g^{\ 3}} \tag{3.76}$$

where f is the frequency of the photons, and v_g is the group velocity of the light in the semiconductor material. Now, we can express the net decrease in electron density due to stimulated emission, from (3.74) and (3.75), as

$$\begin{aligned}\left.\frac{\mathrm{d}n}{\mathrm{d}t}\right|_{\text{loss}} &= (B_{21}n_2 - B_{12}n_1)\phi\ hf\ g(f) \\ &= (n_2 - n_1)B\ \phi\ hf\ g(f)\end{aligned} \tag{3.77}$$

This decrease in VB electron density causes a corresponding increase in the stimulated photon density, that is

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = (n_2 - n_1)B\ \phi\ hf\ g(f)$$

These photons are emitted at a velocity of v_g along the x-axis and so the rate of photon emission *per unit area* is

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = (n_2 - n_1)B\ \phi\ hf\ g(f)\ \mathrm{d}x$$

Thus the power emitted per unit area is

$$\mathrm{d}P = (n_2 - n_1)B\ \phi\ hf\ g(f)\ \mathrm{d}x\ hf \tag{3.78}$$

Emitted photons pass through this unit area at a velocity of v_g, and so we can write the power per unit area as

$$P = \phi\ hf\ v_g$$

or

$$\phi = \frac{P}{hfv_g} \tag{3.79}$$

We can substitute this into (3.78) to give

$$\begin{aligned} \mathrm{d}P &= (n_2 - n_1)B\frac{P}{hfv_g}\ hf\ g(f)\ \mathrm{d}x\ hf \\ &= (n_2 - n_1)\frac{BP}{v_g}\ g(f)\ hf\ \mathrm{d}x \end{aligned}$$

Therefore

$$\frac{\mathrm{d}P}{P} = (n_2 - n_1)\frac{B}{v_g}\ g(f)\ hf\ \mathrm{d}x \tag{3.80}$$

The solution to (3.80) is an exponential given by

$$P(x) = P(0)\exp(gx) \tag{3.81}$$

where

$$g = (n_2 - n_1)\frac{B}{v_g}\ g(f)\ hf$$

We can eliminate B from this equation by using (3.76) and by noting that $A = 1/\tau_{sp}$, where τ_{sp} is the carrier lifetime for spontaneous emission. Thus

$$g = (n_2 - n_1)\frac{{v_g}^2}{8\pi f^2\tau_{sp}}\ g(f)$$

or

$$g = (n_2 - n_1)\frac{\lambda_0^2}{8\pi\epsilon_r\tau_{sp}}\,g(f) \qquad (3.82)$$

where ϵ_r is the relative permittivity of the active region. From (3.82) we can see that n_2 must be greater than n_1 for stimulated emission to occur. This is the quasi-stable state known as a *population inversion.* We can produce a population inversion in SLDs by injecting a large number of electrons into the active region of a double heterojunction diode.

Example

A semiconductor laser diode has a GaAs active region and a population inversion of 2.5 × 10^{24} m^{-3}. Determine the optical gain under these conditions. (Take ϵ_r = 13.1, τ_{sp} = 4 ns and assume a half-power linewidth of 10 nm for the gain function.)

We can use equation (3.82) to express the gain as

$$g = (n_2 - n_1)\frac{\lambda_0^2}{8\pi\epsilon_r\tau_{sp}}\,g(f)$$

Now, $g(f)$ is given by equation (3.72)

$$g(f) = \frac{\delta f}{2\pi[(f - f_0)^2 + (\delta f/2)^2]}$$

A linewidth of 10 nm about a centre wavelength of 870 nm gives a frequency spread of 1.1 × 10^{12} Hz. Thus the gain function *at the centre frequency* is

$$g(f_0) = \frac{\delta f}{2\pi[0 + (\delta f/2)^2]}$$
$$= 5.8 \times 10^{-13}\ \text{s}$$

Thus, the optical gain is

$$g = 2.5 \times 10^{24}\frac{(870 \times 10^{-9})^2}{8\pi \times 13 \times 4 \times 10^{-9}} \times 5.8 \times 10^{-13}$$
$$= 8.4 \times 10^5\ \text{m}^{-1}$$
$$= 8.4 \times 10^3\ \text{cm}^{-1}$$

We should note at this stage, that the gain is effectively clamped at a value given by (3.70). Thus, even if we increase the population inversion, the gain cannot increase past the limit given by (3.70).

It should be evident that we must bias the SLD at a certain current to maintain a population inversion. Below this *threshold current*, the SLD will emit light spontaneously as there will not be enough current to generate a population inversion. In order to find the threshold current density, J_{th}, we must study the rate equations for a SLD.

In section 3.2.2 we derived the rate equations for an LED, equations (3.40) and (3.41). As we are considering stimulated emission, we can write the SLD rate equations as

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \frac{1}{q}\frac{J}{d} - \frac{n_2}{\tau_{\text{sp}}} - C(n_2 - n_1)\phi \tag{3.83}$$

and

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = C(n_2 - n_1)\phi + \frac{Dn_2}{\tau_{\text{r}}} - \frac{\phi}{\tau_{\text{ph}}} \tag{3.84}$$

where C is a constant of proportionality for stimulated emission, and τ_{ph} is the stimulated photon lifetime in the active region. We can justify these equations by simple book-keeping. The first term in (3.83) is the injected carrier density; the second term is the number of carriers lost due to recombination; and the third term is the total loss due to stimulated emission and absorption. As regards (3.84) the first term is the total increase in light due to stimulated emission and absorption; the second term is the fraction of spontaneous emission coupled into a laser mode; and the third term is the loss due to photons being emitted by the cavity. (Although the constant D in (3.84) is typically very low ($\approx 10^{-3}$), its presence helps to explain operation below threshold.)

Rather than examine the dynamic behaviour of a SLD, we will initially study the steady-state rate equations. Thus, dn/dt and dϕ/dt are zero, and we can write

$$0 = \frac{1}{q}\,\frac{J}{d} - \frac{n_2}{\tau_{\text{sp}}} - C(n_2 - n_1)\phi \tag{3.85}$$

and

$$0 = C(n_2 - n_1)\phi + \frac{Dn_2}{\tau_{\text{r}}} - \frac{\phi}{\tau_{\text{ph}}} \tag{3.86}$$

We can combine these equations to give

$$\frac{\phi}{\tau_{\mathrm{ph}}} = \frac{Dn_2}{\tau_{\mathrm{r}}} + \left[\frac{1}{q}\frac{J}{d} - \frac{n_2}{\tau_{\mathrm{sp}}}\right] \tag{3.87}$$

The first term in (3.87) is the spontaneous emission term, while the term in the brackets relates to stimulated emission.

Now, with a laser diode there are three regions of interest: operation *below* threshold, operation *at* threshold, and operation *above* threshold. If we consider operation *below threshold*, the stimulated emission term is zero and so

$$\frac{1}{q}\frac{J}{d} = \frac{n_2}{\tau_{\mathrm{sp}}}$$

which implies

$$n_2 = \frac{\tau_{\mathrm{sp}}}{q}\frac{J}{d}$$

We should also note that zero stimulated emission implies, from (3.87)

$$\begin{aligned}\frac{\phi}{\tau_{\mathrm{ph}}} &= \frac{Dn_2}{\tau_{\mathrm{r}}}\\ &= \frac{D}{\tau_{\mathrm{r}}}\frac{\tau_{\mathrm{sp}}}{q}\frac{J}{d}\\ &= \frac{D}{\tau_{\mathrm{r}}}\frac{\tau_{\mathrm{sp}}}{q}\frac{I}{\text{volume}}\end{aligned}$$

Thus we can write the output optical power as

$$P = \frac{D}{\tau_{\mathrm{r}}}\frac{\tau_{\mathrm{sp}}}{q} I\,hf \tag{3.88}$$

Equation (3.88) shows that if we operate a SLD below threshold, the power output is directly proportional to the applied current, that is, *it is operating as an LED.*

Let us now turn our attention to operation *at threshold.* If we increase the drive current, the light output will increase until there is sufficient spontaneously emitted light to cause stimulated emission. At this stage, we can neglect spontaneous emission, and so (3.86) becomes

$$0 = C(n_{\text{th}} - n_1)\phi - \frac{\phi}{\tau_{\text{ph}}}$$

and so

$$n_{\text{th}} = \frac{1}{C\tau_{\text{ph}}} + n_1 \tag{3.89}$$

where n_{th} is the threshold electron density. So, provided we know the photon lifetime, we can find n_{th}. To find τ_{ph}, we note that C is given by

$$C = B\ hf\ g(f)$$

and, from our previous analysis, the optical gain at threshold is (equation 3.81)

$$g = (n_{\text{th}} - n_1)\frac{Bhf}{v_g}g(f)$$

and so

$$\begin{aligned} g &= \left(n_{\text{th}} - n_1\right)\frac{C}{v_g} \\ &= \left(\frac{1}{C\tau_{\text{ph}}} + n_1 - n_1\right)\frac{C}{v_g} \\ &= \frac{1}{\tau_{\text{ph}}v_g} \end{aligned} \tag{3.90}$$

Now, the optical gain at threshold is also given by (3.70) as

$$g \geq \alpha + \frac{1}{2L} \times \ln\left(\frac{1}{R_1R_2}\right)$$

and so we can find τ_{ph} from

$$\frac{1}{\tau_{\text{ph}}} = v_g\left[\alpha + \frac{1}{2L} \times \ln\left(\frac{1}{R_1R_2}\right)\right] \tag{3.91}$$

Thus we can see that the photon lifetime only depends on the physical parameters of the SLD and not on the level of injection. As we will see in the next section, the photon lifetime also sets a limit on the maximum rate of modulation.

We can now find the threshold current density. By substituting (3.91) into (3.89) we get

$$n_{\text{th}} = \frac{v_g}{Bhfg(f)}\left[\alpha + \frac{1}{2L} \times \ln\left(\frac{1}{R_1R_2}\right)\right] + n_1 \qquad (3.92)$$

Now, the bias current supplies these carriers, and so

$$J_{\text{th}} = \frac{qdn_{\text{th}}}{\tau_{\text{sp}}} \qquad (3.93)$$

If we assume that $n_{\text{th}} >> n_1$, we can write

$$J_{\text{th}} = \frac{qd}{g(f)}\frac{8\pi\epsilon_r}{\lambda_0^2}\left[\alpha + \frac{1}{2L} \times \ln\left(\frac{1}{R_1R_2}\right)\right]$$

and so J_{th} at the nominal wavelength of emission is

$$J_{\text{th}} = qd\,\frac{\pi\delta f}{2}\,\frac{8\pi\epsilon_r}{\lambda_0^2}\left[\alpha + \frac{1}{2L} \times \ln\left(\frac{1}{R_1R_2}\right)\right] \qquad (3.94)$$

We can see from this equation that J_{th} is directly proportional to the width of the active region, and the linewidth of the ELED that makes up the SLD. Although (3.94) is reasonably accurate at low temperatures, under normal operating conditions we can approximate J_{th} by

$$J_{\text{th}}(T) = 2.5\ J_{\text{th}}\exp(T/120) \qquad (3.95)$$

Example

A 300 μm long, GaAs SLD has a loss per cm of 10 and facet reflectivity of $R_1 = R_2 = 0.3$. Determine the gain required before lasing occurs, and the threshold current at this point. Take the width and thickness of the active layer to be 20 μm and 2 μm respectively. Assume an operating wavelength of 870 nm, a spontaneous emission linewidth of 30 nm, and $\epsilon_r = 13.1$.

We can find the gain required for lasing by using (3.70). Thus

$$g = \alpha + \frac{1}{2L} \times \ln\left(\frac{1}{R_1R_2}\right)$$

$$= 50\ \text{cm}^{-1}$$

We can now use (3.94) to give the threshold current density as

$$\begin{aligned} J_{\text{th}} &= qd\,\frac{\pi\delta f}{2}\,\frac{8\pi\epsilon_r}{\lambda_0^2}\,g \\ &= 3.58 \times 10^6 \text{ A/m}^2 \\ &= 358 \text{ A/cm}^2 \text{ at low temperatures.} \end{aligned}$$

If we operate the diode at 300 K, we can use (3.95) to give

$$\begin{aligned} J_{\text{th}}(T) &= 2.5\ J_{\text{th}} \exp(T/120) \\ &= 2.5 \times 358 \times \exp\,(300/120) \\ &= 11 \text{ kA/cm}^2 \text{ at 300 Kelvin} \end{aligned}$$

Thus

$$I_{\text{th}} = 660 \text{ mA}$$

Although these values are typical for gain guided stripe contact lasers, the actual threshold current is likely to be slightly higher than predicted. This is because we have been assuming that the active region confines all of the generated light. In practice, some light will escape the active region so increasing the threshold requirement.

If we now operate the laser *above threshold*, there is negligible spontaneous emission, and so (3.87) becomes

$$\frac{\phi}{\tau_{\text{ph}}} = \left[\frac{1}{q}\,\frac{J}{d} - \frac{n_2}{\tau_{\text{sp}}}\right] \tag{3.96}$$

Now, as the laser is operating above threshold, the gain is effectively clamped at the value given by (3.70) and so any increase in carrier density will not increase the gain – it will, however, increase the light output. Thus we can see that n_2/τ_{sp} is held at its threshold value. Under these conditions we can write (3.96) as

$$\begin{aligned} \frac{\phi}{\tau_{\text{ph}}} &= \left[\frac{1}{q}\,\frac{J}{d} - \frac{n_{\text{th}}}{\tau_{\text{sp}}}\right] \\ &= \frac{1}{q}\left[\frac{J - J_{\text{th}}}{d}\right] \end{aligned} \tag{3.97}$$

and so the output optical power is

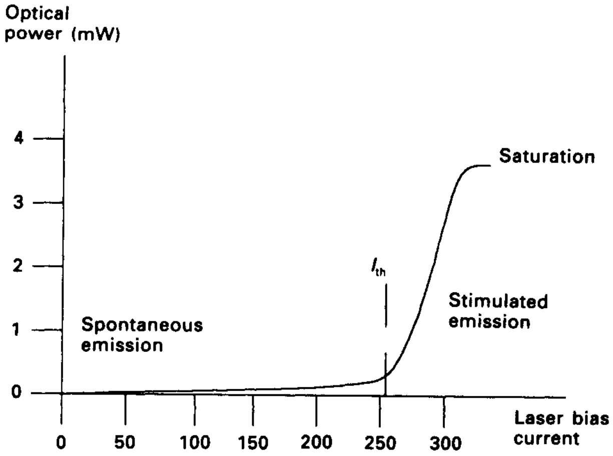


Figure 3.15 Variation of light output with drive current for a SLD

$$P = \frac{I - I_{\text{th}}}{q} hf \tag{3.98}$$

Thus we can see that, for the SLD operating above threshold, the output power is directly proportional to the amount of bias current above threshold. In practice, SLDs operating above threshold do not exhibit a strictly linear relationship between light output and bias current. This is because of *mode-hopping* which we will deal with in the next section.

Figure 3.15 shows the measured variation of output power with diode current for a typical 850 nm SLD. This figure clearly shows the threshold point above which the SLD operates as a laser rather than an ELED. We should also note that the output power saturates at high currents. This is because high currents cause heating of the diode, and this reduces the conversion efficiency.

3.5.2 Spectral characteristics

In common with the planar optical waveguide, only light waves of certain wavelengths can propagate in the cavity. The condition for successful propagation is that the reflected and original waves must be in phase. At the start of the last section we found that the field just to the right of the mirror at $x = 0$ is given by, equation (3.68)

$$E(0, t) = |E|\exp(0)\exp j\omega t$$

We also found that the field travelling back down the cavity, equation (3.69), is

$$E_r(0, t) = \sqrt{R_1R_2}\,|E| \exp[(g - \alpha) L]\text{expj } (\omega t - 2\beta_1 L)$$

In order for the wave to propagate successfully, the phase of the two waves must be the same at $x = 0$, that is

$$\text{expj } (-2\beta_1 L) = 1 \tag{3.99}$$

Therefore

$$2\beta_1 L = 2\pi N \tag{3.100}$$

where N is an integer. Since $\beta_1 = 2\pi n_1/\lambda_0$, (3.100) becomes

$$\lambda_0 = \frac{2n_1 \ L}{N} \tag{3.101}$$

Thus we can see that the laser will only amplify wavelengths that satisfy (3.101). Each wavelength is known as a *longitudinal mode*, or simply a mode (not to be confused with the modes in an optical fibre.) The modes cause a line spectrum, and solution of (3.101) will yield the mode spacing.

Example

A 600 μm long, $Al_{0.03}Ga_{0.97}As$ SLD has a linewidth of 5 nm. Determine the number of laser modes.

As the active region of the laser is a compound semiconductor, we can use equations (3.57) and (3.58) to give

$$E_g = 1.46 \text{ eV resulting in } \lambda_0 = 853 \text{ nm}$$

and

$$n = 3.57$$

With these figures, the nominal mode number (from equation 3.101) is 5022. The next mode corresponds to $N = 5023$, resulting in a mode spacing of 0.17 nm. Thus, with a linewidth of 5 nm, we have approximately 30 different laser modes of varying wavelength.

The spectral emission of a laser is highly dependent on the bias current. Below threshold, spontaneous emission predominates and so the linewidth is similar to that of an LED. However, if we operate above threshold, we find that the linewidth reduces. This reduction occurs because the cavity exponentially amplifies the first mode to reach threshold, at the expense of all other modes. To see this, let us return to the steady-state solution for the photon density, equation (3.86):

$$0 = C(n_2 - n_1)\phi + \frac{Dn_2}{\tau_r} - \frac{\phi}{\tau_{ph}}$$

We can rearrange this equation to give the photon concentration as

$$\phi = \frac{Dn_2}{\tau_r}\left[\frac{1}{\tau_{ph}} - C(n_2 - n_1)\right]^{-1}$$

Now, the term outside the brackets is the amount of spontaneous emission coupled into a laser mode. Thus we can interpret this equation as an amplification factor, G, given by

$$G = \left[\frac{1}{\tau_{ph}} - C(n_2 - n_1)\right]^{-1} \tag{3.102}$$

acting on the spontaneous emission of an ELED. As the gain function has a Lorenztian distribution, we can express (3.102) as

$$G(\omega) = \left[\frac{1}{\tau_{ph}} - \{C(n_2 - n_1) + b\,(\omega - \omega_0)^2\}\right]^{-1} \tag{3.103}$$

When we operate a SLD below threshold, the term in { } is small, and so the cavity amplifies all the propagating modes to the same extent. As we increase the diode current, the amplification increases, but the mode whose wavelength is closest to the nominal operating wavelength is amplified the most. This effect is shown in figure 3.16. Thus we can see that when we operate a SLD above threshold, the linewidth is considerably less than that of an ELED.

In practice, modes close to the fundamental also undergo significant amplification, and so the output consists of a range of modes following a *gain profile*. We can approximate this profile to the Gaussian distribution

$$g(\omega) = g(\omega_0)\exp\left(\frac{-(\omega - \omega_0)^2}{2\sigma^2}\right) \tag{3.104}$$

where σ is the linewidth of the laser output. This result, together with the line spectrum, causes the emission spectrum shown in figure 3.16d. The linewidth of typical stripe contact SLDs can vary from 2 to 5 nm.

If we operate the laser at currents significantly higher than threshold, the gain profile may shift slightly so that one of the modes close to the nominal

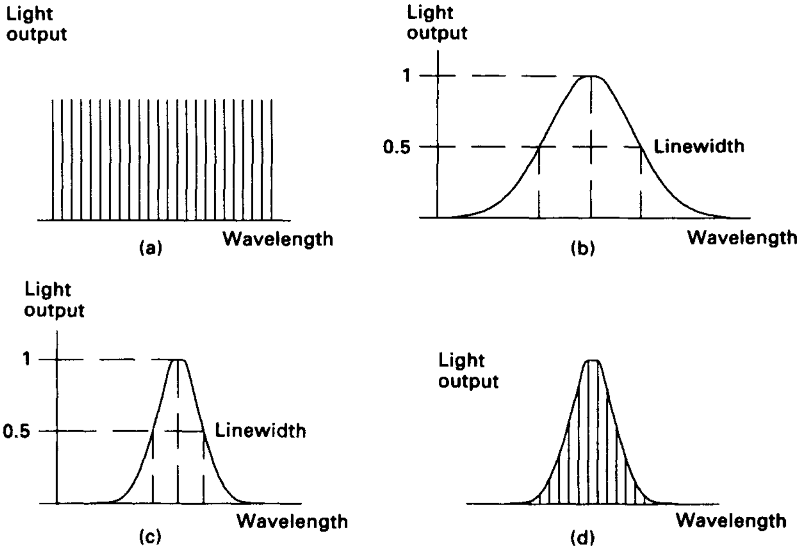


Figure 3.16 (a) Allowable modes in a SLD; (b) gain profile of a SLD operating below threshold; (c) gain profile of a SLD operating above threshold; and (d) resultant emission spectrum

wavelength becomes dominant. This effect is known as *mode-hopping* and it is responsible for kinks in the power/current characteristic. If we modulate the laser by varying the drive current, mode-hopping can alter the operating frequency, and so dynamic mode-hopping is also known as *chirp*. Mode-hopping can cause problems in high-data-rate optical fibre links. If the link is operating at a zero dispersion wavelength, any chirp on the optical pulse will change the operating wavelength so causing pulse dispersion. Thus stripe contact lasers are not commonly found in high-data-rate optical fibre links. Instead we must use alternative laser structures, such as those considered in section 3.5.4.

As well as longitudinal modes, there are also *transverse* and *lateral modes*. These tend to produce an output beam which is highly divergent, resulting in inefficient launching into an optical fibre. The ideal situation is one in which only the fundamental transverse and lateral modes are present. (This would give a parallel beam of light of very small cross-sectional area.) The condition for a single lateral mode is identical to that for a planar dielectric waveguide, and so

$$d < \frac{\lambda_0}{2(n_1{}^2 - n_2{}^2)^{\frac{1}{2}}} \tag{3.105}$$

where n_1 and n_2 are the refractive indices of the active region and the surrounding material respectively. In most laser diodes, the active region is typically less than 1 μm thick and (3.105) is usually satisfied. Unfortunately single transverse mode operation is more difficult to achieve. This is because the width of the active region is set by the current density profile in the active layer, which can be difficult to control in the stripe contact lasers we are considering.

3.5.3 Modulation capabilities

As we have seen, stimulated emission only occurs if a population inversion is present in the active region. It takes a significant length of time for the SLD current to set up a population inversion, and so SLDs are generally biased above threshold using a constant current source.

Let us initially consider digital modulation of the SLD. If we bias the SLD at threshold, and we increase the drive current by a small fraction, δI, we can write

$$n = n_{\text{th}} + \delta n \qquad \text{and} \qquad \phi = \phi_{\text{s}} + \delta\phi$$

where ϕ_{s} is the steady-state photon density. We can substitute these densities into the rate equations, (3.83) and (3.84) to give

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\delta n = -C\delta n\phi_{\text{s}} - \frac{\delta n}{\tau_{\text{sp}}} - Cn_{\text{th}}\delta\phi \tag{3.106}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\delta\phi = C\delta n\phi_{\text{s}} \tag{3.107}$$

In deriving (3.106) and (3.107) we have assumed that spontaneous emission is negligible; we have neglected the $\delta n\delta\phi$ term; and we have made use of equations (3.89), (3.93) and (3.97). For reasons of clarity, these equations are reproduced here:

$$n_{\text{th}} \approx \frac{1}{C\tau_{\text{ph}}}$$

$$J_{\text{th}} = \frac{qdn_{\text{th}}}{\tau_{\text{sp}}}$$

$$\frac{\phi_{\text{s}}}{\tau_{\text{ph}}} = \frac{1}{q}\left[\frac{J - J_{\text{th}}}{d}\right]$$

We can combine (3.106) and (3.107) by differentiating (3.106) with respect to time, and substituting for $\mathrm{d}\delta\phi/\mathrm{d}t$ from equation (3.107). Thus we can write

$$\frac{\mathrm{d}^2\delta n}{\mathrm{d}t^2} + \left[C\phi_s + \frac{1}{\tau_{sp}}\right]\frac{\mathrm{d}}{\mathrm{d}t}\,\delta n + C^2 n_{th}\phi_s\,\delta n = 0$$

We can re-write this equation as

$$\frac{\mathrm{d}^2\delta n}{\mathrm{d}t^2} + 2\sigma\,\frac{\mathrm{d}\delta n}{\mathrm{d}t} + \omega_0^{\,2}\,\delta n = 0 \tag{3.108}$$

where

$$\begin{aligned} 2\sigma &= C\phi_s + 1/\tau_{sp} \\ &= \frac{\phi_s}{n_{th}\tau_{ph}} + \frac{1}{\tau_{sp}} \end{aligned} \tag{3.109}$$

and

$$\begin{aligned} \omega_0^{\,2} &= C^2 n_{th}\phi_s \\ &= \frac{\phi_s}{n_{th}\tau_{ph}{}^2} \end{aligned} \tag{3.110}$$

Equation (3.108) is a standard differential equation whose solution describes a damped oscillation. Thus we can write the solution as

$$\delta n = A\sin(\omega t)\exp(-\sigma t)$$

where $\omega^2 = \omega_0^{\,2} - \sigma^2$, and A is a constant of integration. We can find A by using the initial conditions $n = n_{th}$ and $\phi = 0$ at $t = 0$. Thus we find

$$A = \frac{\omega}{C}$$

and so

$$\delta n = \frac{\omega}{C}\sin(\omega t)\exp(-\sigma \mathrm{t}) \tag{3.111}$$

we can substitute (3.111) into (3.107) to give

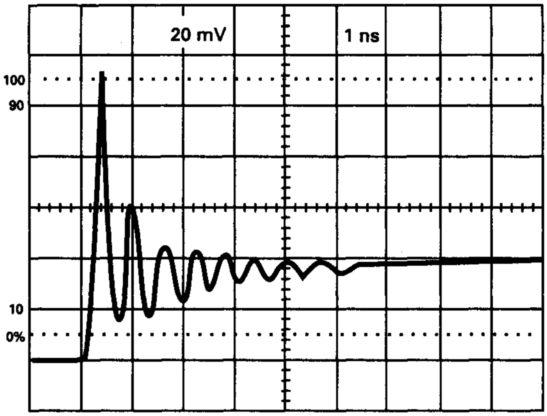


Figure 3.17 Output pulse from an AlGaAs laser diode

$$\delta\phi = -\phi_s\cos(\omega t)\exp(-\sigma \mathrm{t}) \qquad (3.112)$$

Now, the damping coefficient, σ, is usually small with respect to ω_0, and so we get $\omega \approx \omega_0$ and $A = (n_{th}\phi_s)^{\frac{1}{2}}$. With this approximation we can write

$$\delta n = (n_{th}\phi_s)^{\frac{1}{2}}\sin(\omega_0 t)\exp(-\sigma t) \qquad (3.113)$$

and

$$\delta\phi = -\phi_s\cos(\omega_0 t)\exp(-\sigma t) \qquad (3.114)$$

So, under digital modulation, the optical pulse suffers from ringing at an angular frequency given by (3.110). Figure 3.17 shows the actual output pulse of an AlGaAs laser diode operating at 850 nm. As can be seen, the figure clearly shows the damped response predicted by (3.114).

Let us now turn our attention to analogue modulation of the laser diode. Once again we assume that the diode is biased sufficiently above threshold so that the current never falls below I_{th}. Now, if we modulate the diode current at an angular frequency of ω_m, we can write

$$J = J_0 + J'\exp(\mathrm{j}\omega_m t)$$

We can also express the carrier and photon densities as

$$n = n_0 + n'\exp(\mathrm{j}\omega_m t) \qquad \text{and} \qquad \phi = \phi_0 + \phi'\exp(\mathrm{j}\omega_m t)$$

We can now substitute these values into the rate equations, (3.83) and (3.84), to give

$$\mathrm{j}\omega_{\mathrm{m}}n' = \frac{1}{q}\frac{J'}{d} - \frac{n'}{\tau_{\mathrm{sp}}} - \frac{\phi'}{\tau_{\mathrm{ph}}} - Cn'\phi_0 \tag{3.115}$$

and

$$\mathrm{j}\omega_{\mathrm{m}}\phi' = Cn'\phi_0 \tag{3.116}$$

If we compare (3.115) to the rate equation we derived for a step function, equation (3.106), we can see that we have an additional term due to the sinusoidal change in diode current. We can combine these two equations by substituting for n' from (3.116) into (3.115) to give

$$-\omega_{\mathrm{m}}{}^2\phi' = C\phi_0\frac{1}{q}\frac{J'}{d} - \frac{\mathrm{j}\omega_{\mathrm{m}}\phi'}{\tau_{\mathrm{sp}}} - \frac{C\phi_0\phi'}{\tau_{\mathrm{ph}}} - C\phi_0\mathrm{j}\omega_{\mathrm{m}}\phi'$$

By collecting terms, we can write

$$\phi'\left[-\omega_{\mathrm{m}}{}^2 + \mathrm{j}\omega_{\mathrm{m}}\left(C\phi_0 + \frac{1}{\tau_{\mathrm{sp}}}\right) + \frac{C\phi_0}{\tau_{\mathrm{ph}}}\right] = C\phi_0\,\frac{1}{q}\,\frac{J'}{d} \tag{3.117}$$

We can simplify (3.117) by using the definitions of damping factor, equation (3.109), and natural frequency, equation (3.110), to give

$$\phi'\left(-\omega_{\mathrm{m}}{}^2 + \mathrm{j}\omega_{\mathrm{m}}2\sigma + \omega_0{}^2\right) = C\phi_0\frac{1}{q}\frac{J'}{d} \tag{3.118}$$

where we have also made use of $n_{\mathrm{th}} = 1/(C\tau_{\mathrm{ph}})$. The laser emits photons at a rate of $1/\tau_{\mathrm{ph}}$, and so we can write the modulated optical power, P', as

$$P' = \frac{hcI'}{q\lambda_0}\,\frac{\omega_0{}^2}{(-\omega_{\mathrm{m}}{}^2 + \mathrm{j}\omega_{\mathrm{m}}2\sigma + \omega_0{}^2)} \tag{3.119}$$

where we have again made use of $n_{\mathrm{th}} = 1/(C\tau_{\mathrm{ph}})$. We can define the change in optical power under d.c. modulation as

$$P'(0) = \frac{hcI'}{q\lambda_0}$$

and so (3.119) becomes

$$P' = P'(0)\frac{\omega_0{}^2}{(-\omega_{\mathrm{m}}{}^2 + \mathrm{j}\omega_{\mathrm{m}}2\sigma + \omega_0{}^2)}$$

Thus we can write the modulation depth, $m(\omega)$, as

$$m(\omega) = \frac{P'}{P'(0)} = \frac{\omega_0^2}{(-\omega_m^2 + j\omega_m 2\sigma + \omega_0^2)} \tag{3.120}$$

Example

A 300 μm long, GaAs SLD has a loss per cm of 10 and facet reflectivity of $R_1 = R_2 = 0.3$. The width of the active region is 2 μm, ϵ_r for GaAs is 13.1, and the lifetime of spontaneously emitted photons is 4 ns. If the laser is biased with a current density 800 A/cm^2 above the threshold of 11 kA/cm^2, plot (3.120) as a function of the frequency of the modulating signal.

In order to plot (3.120) we need to calculate the lifetime of the stimulated photons. Equation (3.91) gives τ_{sp} as

$$\frac{1}{\tau_{ph}} = v_g \left[\alpha + \frac{1}{2L} \times \ln\left(\frac{1}{R_1 R_2}\right)\right]$$

and so

$$\frac{1}{\tau_{ph}} = 4.16 \times 10^{11}$$

hence

$$\tau_{ph} = 2.4 \text{ ps}$$

The natural frequency of the laser is given by (3.110) as

$$\omega_0^2 = \frac{\phi_s}{n_{th}\tau_{ph}^2}$$

and we can find the density of carriers at threshold from (3.93):

$$J_{th} = \frac{qdn_{th}}{\tau_{sp}}$$

Thus

$$n_{th} = 1.4 \times 10^{24} \text{ m}^{-3}$$

and so

$$\omega_0^2 = \frac{\phi_s}{2.2 \times 10^7}$$

Now, from (3.97), ϕ_s is given by

$$\frac{\phi_s}{\tau_{ph}} = \frac{\tau_{ph}}{q}\left[\frac{J - J_{th}}{d}\right]$$

hence

$$\phi_s = 6 \times 10^{19}\ \mathrm{m}^{-3}$$

and so

$$\omega_0^2 = 7.6 \times 10^{18} \qquad \text{or} \qquad \omega_0 = 2.8 \times 10^9\ \text{rad/s.}$$

The damping factor is given by (3.109)

$$2\sigma = \frac{\phi_s}{n_{th}\tau_{ph}} + \frac{1}{\tau_{sp}}$$

and so

$$2\sigma = 2.7 \times 10^8$$

Thus (3.120) becomes

$$m(\omega) = \frac{7.6 \times 10^{18}}{(-\omega_m^2 + j\omega_m \times 2.7 \times 10^8 + 7.6 \times 10^{18})}$$

A plot of this equation is shown in figure 3.18.

We can see from this figure that the frequency response of the SLD has a peak at a frequency close to the natural frequency of the cavity, ω_0. This is known as a *relaxation resonance*, and it is caused by a resonance between the photon and electron populations in the laser. Under digital modulation, this gives rise to the ringing shown in figure 3.17. We should also note that the response falls off at a rate ω_0^2/ω_m^2 at frequencies above ω_0. Thus the bandwidth of the SLD is set by the frequency of the relaxation resonance.

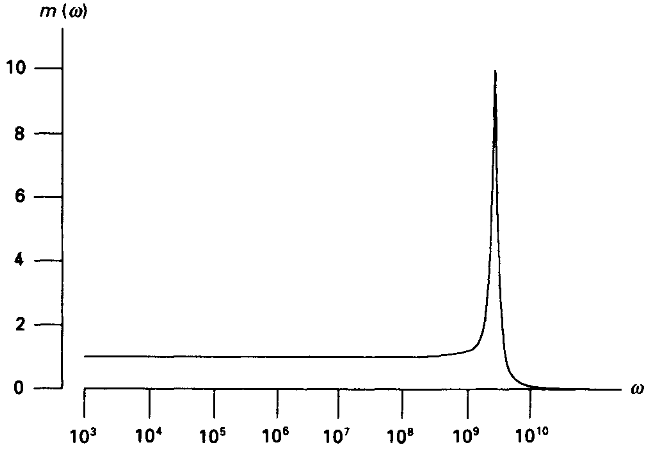


Figure 3.18 Frequency response of a SLD biased above threshold

3.5.4 SLD structures

As SLDs are normally used in long-haul, high-data-rate routes, which use SM fibres, it is generally desirable to minimise the linewidth and operate with a single lateral mode. It is also important to reduce the threshold current, as this will produce a more efficient device.

At present, the most common SLD structure for general use is the stripe contact design we have been considering. The most obvious way of reducing I_{th} is to reduce the active region cross-sectional area. As this is set by the area of the stripe contact, we could reduce I_{th} by reducing the cavity length. Unfortunately this causes the gain required for threshold to increase (see equation 3.94) so causing J_{th} to increase. As a high current density causes heatsinking problems, the cavity length is usually limited to typically 150 μm, and so we must reduce the contact width to reduce I_{th}.

To a certain extent the width of the active region is set by the width of the contact. In practice, I_{th} fails to fall in proportion to the contact stripe width, if it is less than about 6 μm. This is because the injected current tends to diffuse outwards as it travels through the laser. Ultimately, we get an active region that is independent of the contact width. So the threshold current of stripe contact lasers is usually no less than 120 mA.

In order to reduce I_{th} further, and operate with a single lateral mode, we must use a different structure. In a *buried heterostructure, BH,* laser, the diode current is constrained to flow in a well-defined active region, as shown in figure 3.19. The heterojunctions either side of the active region provide carrier confinement, and so the width of this region can be made very small, typically 2 μm or less. The heterojunctions will also produce a narrow op-

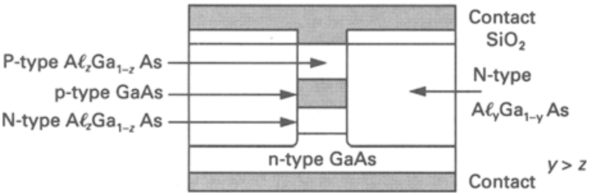


Figure 3.19 Cross-section through a buried-heterojunction, semiconductor laser diode

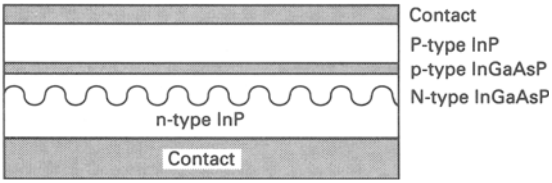


Figure 3.20 Cross-section through a distributed feedback semiconductor laser diode

tical waveguide, and so single lateral mode operation is often achievable. The threshold current of these devices is typically 30 mA.

A further advantage of the BH structure is that, by using a small active region, the gain profile is considerably narrowed. Thus the emission spectrum of some BH lasers can consist of a single line – a considerable advantage in long-haul routes operating at a zero dispersion wavelength. Unfortunately the gain profile is dependent on the junction temperature, and so the wavelength of emission can change during operation so introducing dispersion.

A truly single-mode source results from distributing the feedback throughout the laser, the so-called *distributed feedback*, or *DFB*, laser. In these devices, a grating replaces the Fabry–Perot cavity resonator (figure 3.20).

The effect of this grating is to select just one propagating mode. This happens because each perturbation reflects some of the light and, in order to propagate successfully, the phase of the twice reflected light must match that of the incident light. We can write this condition as

$$2n_2\ddot{A} = m\lambda_0 \tag{3.121}$$

where $\ddot{A}$ is the period of the grating, n_2 is the refractive index of the material above the grating, and m is an integer. (The factor of 2 appears in the left-hand side of (3.121) because the light must be reflected twice in order to be in phase with the incident wave.) If (3.121) is not satisfied, the scattered light from the grating will interfere destructively, and the wave will not propagate.

Equation (3.121) is a special case of *Bragg's law* and if m equals unity, the wave is said to be incident at the first Bragg condition. It is also possible for light to be reflected using the second Bragg condition. In fact we can see that if $m = 2$, the grating period will increase, so making it easier to manufacture. We should note that the grating is not part of the active layer. This is because a grating in the active region will cause surface dislocations, and this will increase the non-radiative recombination rate. Instead, the grating is usually placed in a waveguide layer where it interacts with the evanescent field.

A modification of the DFB laser is the *distributed Bragg reflector, DBR*, laser. In this device, short lengths of grating, which act as frequency selective reflectors, replace the Fabry–Perot resonator. Hence many modes propagate in the active region, but only a single wavelength is reflected back and undergoes amplification.

The threshold current of both these devices is typically 20 mA, and their linewidth is quite narrow <0.5 nm. Thus high-data-rate/long-haul routes often use these devices. As we have seen, these lasers rely on the grating period to select a particular wavelength. However, changes in temperature will cause the grating to expand or contract, and so the wavelength will change. We can control the laser temperature by mounting the semiconductor on a *Peltier cooler*. If we then place a thermistor close to the device, we can use a simple control loop to maintain the laser temperature.

3.5.5 Packaging and reliability

SLDs for use in the laboratory are usually mounted in brass studs, similar to the one shown in figure 3.21. With this package, the body of the stud forms the anode of the SLD, and so we must connect the lead at the rear of the package to a negative voltage, current source. A thread on the back of the stud enables us to bolt the diode on to an efficient heatsink.

For commercial applications, the SLD is commonly mounted on a Peltier cooler in a dual-in-line package, also shown in figure 3.21. A photodiode placed on the non-emitting end of the laser provides the power monitoring facility. A fibre pig-tail, with a lens grown on the laser end of the fibre, provides the output. For launching into MM fibre, a hemispherical lens is often used, whereas for launching into SM fibres, a tapered lens is more common. The lenses can be made by dipping the fibre end into low-melting-point glass. With this technique, up to 66 per cent of the output power can be coupled into the fibre.

As a SLD ages, the threshold current requirement tends to increase (this is caused by the carrier lifetime decreasing with age.) Thus a feedback loop must be used that monitors the laser output, and increases the drive current

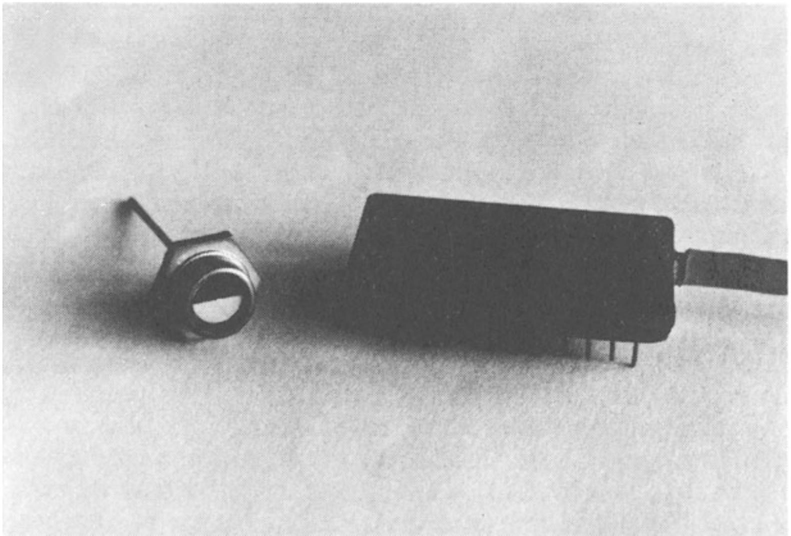


Figure 3.21 Stud mounted and dual-in-line laser diode packages

accordingly. Unfortunately, the threshold point tends to become less well-defined as time passes, and so the control loop must include some means of raising an alarm if the threshold requirement becomes too great. Accelerated life testing suggests that this condition occurs after, typically, 20 to 25 years.

3.6 Solid-state and gas lasers

So far we have confined our discussion to semiconductor laser diodes. Although we find these devices in optical fibre links, we seldom find them in free-space optical links because of their low output power. Instead, we can use high-power solid-state or gas lasers. Such lasers are usually physically large, and modulation of the light output can prove difficult. In spite of this, such lasers are often used in the laboratory, and their application to free-space links is developing.

We begin our study by examining the Nd^{3+}: YAG laser which is extensively used in manufacturing industry as a cutting tool. We will then go on to study the HeNe gas laser, which is often used as a teaching aid in laboratory experiments. We will leave the problem of modulating the light output until we consider lightwave modulation later.

3.6.1 Nd^{3+}:YAG lasers

Neodymium lasers are solid-state lasers in which Nd^{3+} is used as a dopant in the host material. This host material can be certain types of glass, or single crystal rods of yttrium–aluminium–garnet, YAG – $Y_3Al_5O_{12}$, where the Nd^{3+} ions displace some of the yttrium atoms. As Nd^{3+}:YAG is an insulating material, we have to generate a population inversion by external means. The most common way of producing a population inversion is to pump the laser rod with the output of a tungsten–halogen lamp. Under these conditions, we find that several watts of output power are produced for several hundred watts of pump power. Thus we can see that Nd^{3+}:YAG lasers are very inefficient devices (typically 1 per cent).

Figure 3.22 shows the schematic of a typical Nd^{3+}:YAG laser The crystal rod is placed inside a Fabry–Perot cavity, to provide optical feedback, and one of the mirrors is made slightly reflecting in order to couple light out of the cavity. We should note that the ends of the crystal rod are cut at an angle to the axis known as the ***Brewster*** angle. The use of Brewster windows means that only transverse magnetic, *TM*, light is coupled out of the laser rod. This light is reflected off the Fabry–Perot mirrors, and so the TM light will have a lower threshold than TE light. Linearly polarised light is of great importance if we wish to use the external modulators which we consider later.

As we should expect, the emission wavelength of a Nd^{3+}:YAG laser is dependent on electron transitions between levels in the crystal material, as shown in figure 3.23. Electrons appear in the upper lasing level, the $^4F_{3/2}$ level, by dropping down from the main pump bands in the higher *F*, *G* and *H* levels. The spontaneous lifetime of electrons in the upper lasing level is typically 500 μs. In decaying to the lower lasing level at $^4I_{11/2}$ the electrons lose 1.17 eV, and so the wavelength of the laser is 1.06 μm. By incorporating a highly selective Fabry–Perot etalon in the laser cavity, we can operate

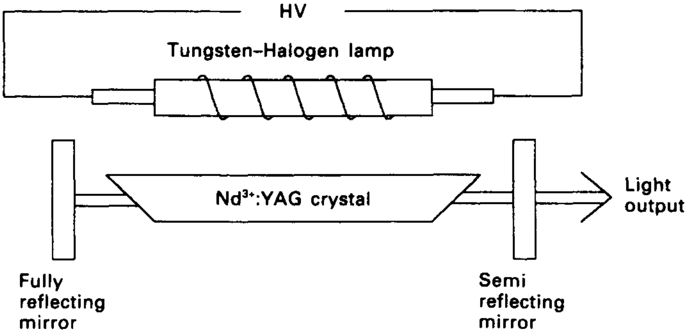


Figure 3.22 Schematic of a typical Nd^{3+}:YAG laser

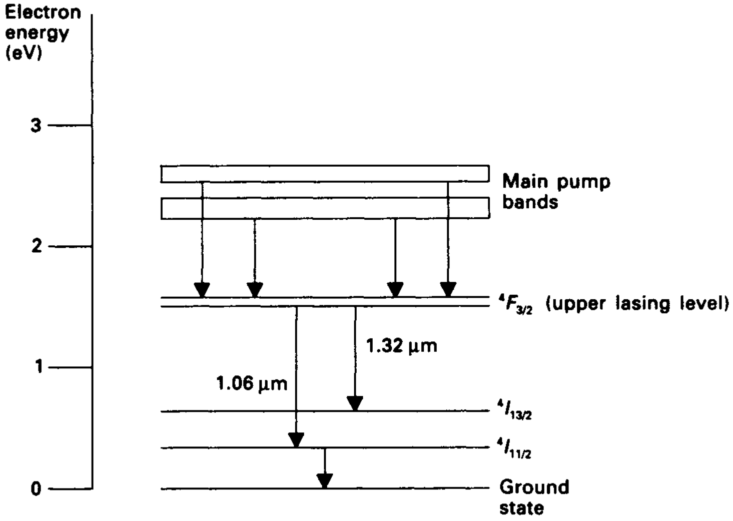


Figure 3.23 Energy levels in a Nd^{3+}:YAG laser

the laser with a single longitudinal mode, resulting in operation at a single frequency.

Example

A Nd^{3+}:YAG laser has a cavity length of 20 cm, and a refractive index of 1.5. Determine the number of laser modes if the gain linewidth of the laser is 0.674 nm.

By using equation (3.101) we find that the mode spacing is 500 MHz. The gain linewidth is 18×10^{10} Hz, and so 360 modes can propagate in the laser cavity. This is why a highly selective etalon is often used.

Before we leave the Nd^{3+}:YAG laser, we should note that there is a possible laser transition from the upper lasing level to the $^4I_{3/12}$ level. Electrons dropping to this level lose 0.941 eV of energy, resulting in emission at a wavelength of 1.32 μm – one of the low attenuation windows in optical fibre links. We can achieve operation at this wavelength by using coated cavity mirrors that only reflect 1.32 μm light. Although these lasers could be used in optical fibre links, the 1.55 μm transmission window offers lower attenuation, and so 1.32 μm Nd^{3+}:YAG lasers are usually limited to laboratory experiments.

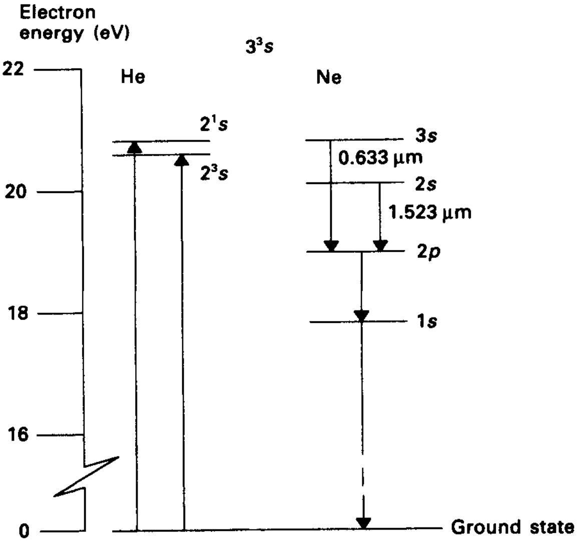


Figure 3.24 Energy levels in a HeNe laser

3.6.2 HeNe lasers

The HeNe laser was the first laser to be operated continuously rather than pulsed. The schematic diagram of a HeNe laser is similar to that of the Nd^{3+}:YAG laser (figure 3.22), except that there is no need for the tungsten–halogen lamp. Instead, we can excite the HeNe by ionising the gas using a high voltage d.c. supply, typically 1–2 kV. The gas mixture usually contains 1 mm Hg of He, and 0.1 mm Hg of Ne. Excitation by the HV supply causes a plasma to be formed in the laser tube, so exciting the helium atoms.

Figure 3.24 shows the energy levels in a HeNe laser. When a plasma is struck across the laser tube, He atoms are excited to the 2^1s and 2^3s levels. As the He atoms are excited to this higher level, they collide with the Ne atoms, so exciting them to the 2*s* and 3*s* levels. The Ne electrons then ultimately decay to the 1*s* level, from which they relax to the ground state by collision with the tube walls. The transition to the 1*s* level can occur through several different routes, each one generating light of a specific wavelength. So, although the He atoms are excited by the electrical discharge, it is the Ne atoms that cause the laser output.

The most commonly used transition in HeNe lasers is from the 3*s* level to the 2*p* level. In doing this, the electrons lose an energy of 1.963 eV which results in light of wavelength 0.633 μm – red light. It is also possible for decay to occur from the 2*s* level to the 2*p* level. In this case, the energy difference is 0.816 eV, and so we can see that light is emitted at 1.523 μm.

This coincides with the lowest attenuation window in optical fibre, and so we should expect to find 1.523 μm lasers in optical fibre communications links. However, the strength of the 1.523 μm line is very low, typically <200 μW, and so HeNe lasers are not very useful in optical links. Instead, they are often used as a cheap source of 1.523 μm light for use in the laboratory.

3.7 Light-wave modulation

In this section we will consider various techniques for modulating the output of a light source. With a semiconductor light source, such as the SLD or LED, we can modulate the light output by varying the drive current. However, solid-state and gas lasers are usually CW devices, and so we must use some form of external modulator.

We begin by studying LED and SLD drive circuits. We will then go on to examine external modulators, which are used in high-data-rate systems.

3.7.1 LED drive circuits

For analogue modulation of an LED, we can use the simple class A amplifier, shown in figure 3.25. Provided the modulation depth, *m*, is less than 100 per cent, no signal distortion will occur. The modulation depth is defined as

$$m = \frac{\delta I}{I_B} \tag{3.122}$$

where I_B is the LED bias current. With careful selection of the drive transistor,

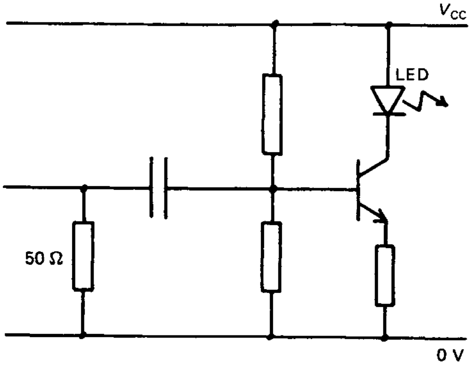


Figure 3.25 A simple analogue driver for LED sources

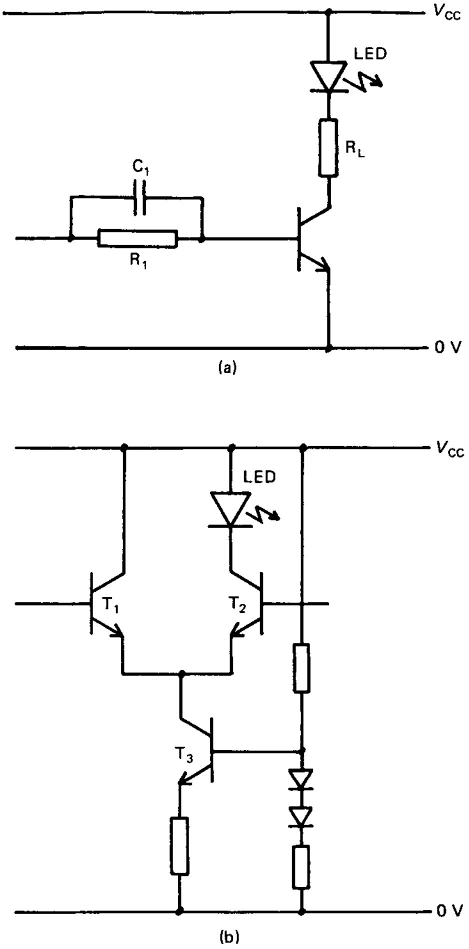


Figure 3.26 (a) A simple digital driver with speed-up capacitor and (b) a basic emitter-coupled switch for LED light sources

the time constant of the LED will limit the maximum frequency of operation.

For digital modulation, we can use the simple transistor switch shown in figure 3.26a. In this circuit, R_L limits the LED current while R_1 limits the transistor current. The purpose of the capacitor, C_1, is to provide a speed-up transient to charge and discharge the LED capacitance. This circuit is suitable for data-rates less than 100 Mbit/s.

For operation at data-rates greater than 100 Mbit/s, an emitter-coupled driver will often suffice (figure 3.26b). When the input is high, T_1 turns on, so diverting current away from the LED, which then turns off. When the input is low, T_1 turns off, and the LED turns on.

For commercial applications, most manufacturers supply a package containing the LED and all the drive circuitry. The light output is taken from a short length of fibre, a *fibre pig-tail*, or through a connector housing. Hence the only connections that need to be made to the unit are the power supply and the signal.

3.7.2 SLD drive circuits

The requirement to bias the laser at, or above, threshold means that SLD drive circuits can be complex. In addition, because I_{th} increases with temperature, a feedback loop regulates the diode current. So, a typical SLD drive circuit consists of a constant current source, incorporated in a feedback loop. Such a circuit is shown in schematic form in figure 3.27, in which a monitor photodiode attached to the non-emitting laser facet provides the feedback signal. In order to alleviate heatsinking problems, most commercial laser packages incorporate semiconductor Peltier coolers, which also help to keep the threshold current low.

As we have already seen, the light output of SLDs is due to stimulated emission. As this process is faster than spontaneous emission, the emitter-coupled circuit of figure 3.26b, shown previously, can be used. However, for high-speed operation, we must specify microwave bipolar transistors, or GaAs MESFETs. Although the rise-time of the laser optical pulse can be very fast $\approx$ 500 ps, the fall-time is usually longer, >1 ns. Charge storage in the active region causes this effect, and so the fall-time often limits the maximum speed of modulation.

In an optical fibre link, additional dispersion due to laser chirp can also place a limit on the maximum data-rate. One way of avoiding both charge storage and laser chirp is to operate the SLD continuously, and use an external modulator to modulate the light output.

3.7.3 External modulators

External modulators fall into two broad areas: waveguide devices for use in optical fibre links; and bulk modulators for use in high power free-space links. Bulk modulators come in a variety of forms, and we will consider some of these at the end of this section. Nearly all waveguide modulators are made using lithium niobate, and it is these that we will consider first. (For further information on external modulators, the interested reader is referred to Yariv [4].)

When we examined optical couplers in the previous chapter, we briefly considered a single-mode, waveguide coupler fabricated on a lithium niobate, $LiNbO_3$, substrate. One property of this material is that the refractive index

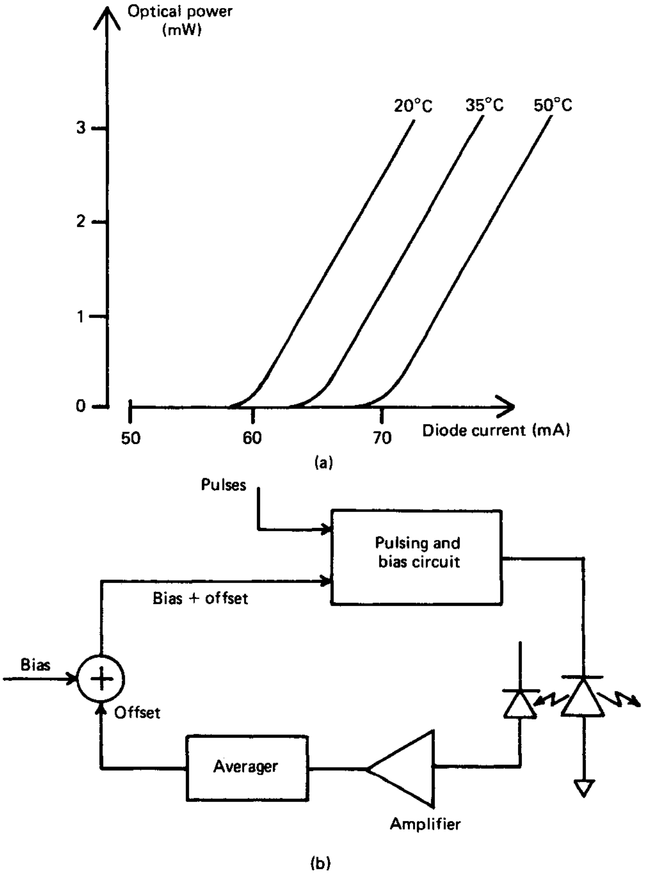


Figure 3.27 (a) Variation of I_{th} with temperature and (b) a simple laser bias stabiliser

varies according to the strength of an externally applied electric field, the so-called *electro-optic* effect. We can exploit this effect to produce phase and intensity modulators. Although we are more usually concerned with varying the intensity of a light source rather than the phase, we will initially consider phase modulators. This will help us when we come to examine the intensity modulator.

Figure 3.28a shows the basic structure of a typical phase modulator. In this device, the electrodes either side of the waveguide set up an electric field, E, across the guide. This has the effect of increasing the refractive index, and hence the propagation time and phase. We can write the change in phase experienced by the optical signal, $\delta\phi$, as

$$\delta\phi = \frac{2\pi}{\lambda_0} \times \delta n \times L \tag{3.123}$$

where L is the length of the guide. The change in refractive index, δn, is related to the electric field by

$$\delta n = n^3 \times \frac{r}{2} \times E \tag{3.124}$$

and so the phase change is directly proportional to the applied voltage. The parameter r is called the *electro-optic coefficient.* In $LiNbO_3$, $r = 30 \times 10^{-12}$ m/V and $n = 2.2$, and so a typical field of 10^7 V/m results in $\delta n = 1.6 \times 10^{-3}$. If the substrate is GaAs, then δn, for the same field, is 2.57×10^{-4}.

The electrodes in the phase modulator form a capacitor, C, which must be charged and discharged by the voltage source supplying the modulating signal. The output resistance of this source is usually fixed at 50 Ω, and so the maximum operating speed is set by the time constant $50C$. We could reduce the capacitance by decreasing the length of the electrodes and increasing their separation. Unfortunately increasing electrode length reduces the phase shift, while increasing the separation reduces the electric field strength. One way round this is to use travelling-wave electrodes as shown in figure 3.28a. In order to eliminate electrical reflections, we must make sure that the whole system is matched to the characteristic impedance of the transmission line. With this modification, the maximum operating speed is typically 20 GHz.

In the Mach–Zehnder interferometer [5] shown in figure 3.28b, a Y-junction waveguide splits the input power equally between the two arms of the device. Phase modulators placed in the two arms alter the relative phases of the fields prior to recombination in another Y junction. If the phase difference between the two paths is $2N\pi$ radians, where N is an integer, the fields will add and light will appear at the output. However, if the phase difference is $(2N + 1)\pi$ radians, the waves will cancel each other out and the output will be zero. It is a simple matter to show that the output power is given by

$$P_{\text{out}} = P_{\text{in}}\cos^2\left(\frac{\Delta\phi}{2}\right) \tag{3.125}$$

where $\Delta\phi$ is the phase difference between the two branches. The measured transmission/drive voltage characteristic of a typical Mach–Zehnder modulator is shown in figure 3.28c.

Mach–Zehnder modulators are of great use when a laser has to be modu-

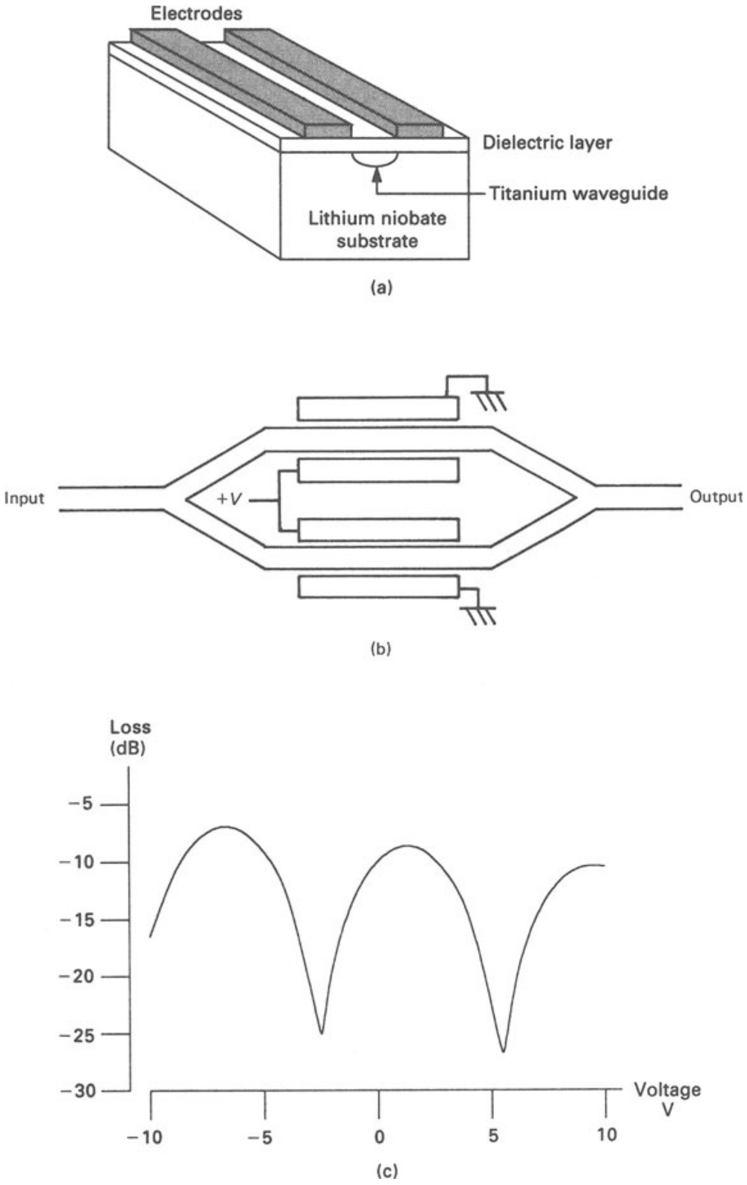


Figure 3.28 (a) A simple $LiNbO_3$ phase modulator; (b) a Mach–Zehnder interferometer; and (c) measured transfer function of a Mach–Zehnder interferometer

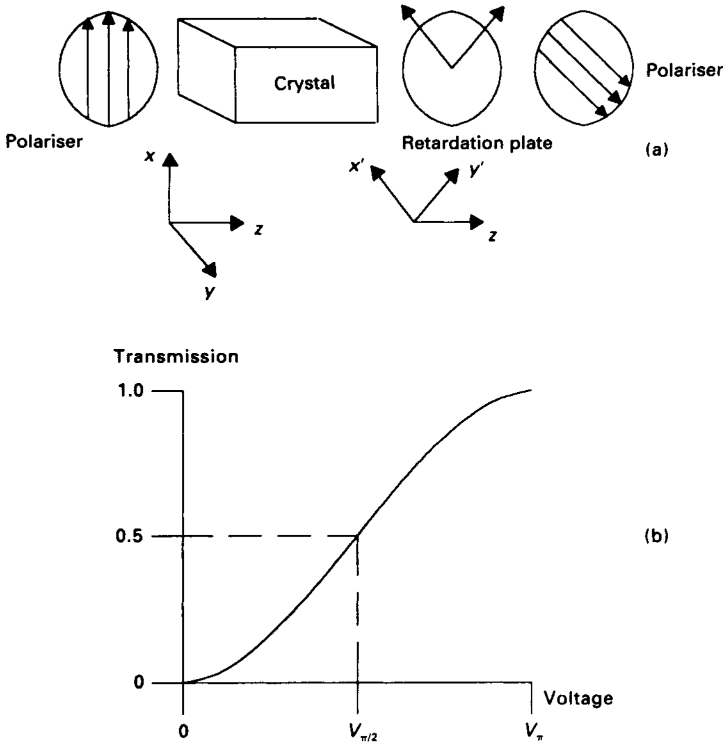


Figure 3.29 (a) Schematic diagram of an electro-optic bulk modulator and (b) transfer function of the modulator

lated at high speed. As we have already noted, the wavelength of emission varies slightly when the drive current to a laser is pulsed on and off. If the laser is a single-mode device designed to operate in a low dispersion link, this change in wavelength could result in considerable dispersion. Thus for high-speed operation, we can operate the laser continuously and use a Mach–Zehnder modulator to modulate the output. Such a technique is widely used at present.

If we want to modulate the light output of a high power laser, we have to use bulk modulators. This is because the optical power density in a single-mode waveguide modulator would be extremely high, and could cause damage to the device. However, if we have a wide output beam, the power density will be lower, and so modulator damage is less likely to occur. In the main, we find that bulk modulators fall into two groups: those that use the electro-optic effect, and those that use the acousto-optic effect.

Bulk modulators using the electro-optic effect rely on the polarisation changes that the modulator crystal produces when excited. Figure 3.29a shows the schematic of a typical electro-optic amplitude modulator. Let us initially neglect the effect of the quarter-wave retardation plate. As we can see, a polariser (which is aligned to the crystal lattice) is used at the input to the device. As the light passes through the energised crystal, two things occur. Firstly the polarisation is altered by a maximum of π/2, so that the light travels along the x'- and y'-axes. The second effect is that the light travelling along the x'-axis experiences a maximum phase delay of π/4 radians with respect to the y'-axis. This effect is known as *birefringence*, which we can interpret as the x' wave travelling in a higher refractive index material than the y' wave. (We should note that both the polarisation shift and the phase shift depend on the voltage across the crystal.)

At the output of the crystal, the light passes through an output polariser which is aligned at right angles to the input polariser. (We are temporarily neglecting the effects of the retardation plate.) Thus we can see that if the crystal is energised, the total change in polarisation is π/2 radians, and light passes through the modulator. However, if the applied voltage is zero, the polarisation state is maintained, and no light appears at the modulator output.

If we want to modulate the light with a sinusoidal waveform, we need to bias the crystal so that it introduces a π/4 polarisation shift. (This enables us to vary the polarisation shift between 0 and the maximum π/2 radians.) Rather than supply a fixed bias to the crystal, a quarter-wave retardation plate is often used to introduce a π/4 shift regardless of the bias signal. Thus we can see that the energised crystal only needs to introduce a maximum shift of π/4 radians.

To see the effect of the applied voltage on the light, let us consider vertically polarised light, propagating along the z-axis. In passing through the crystal, the polarisation changes by a maximum of π/4 radians. Thus the light in the crystal is equally split between the x'- and y'-axes. We can therefore write the input E field components as

$$E_{x'}(0) = A$$

and

$$E_{y'}(0) = A$$

where we have taken the phasor representation of the E fields. Now, in passing through the modulator, the x' wave experiences a total phase shift of Γ radians. Thus we can write

$$E_{x'}(l) = A\exp(-\mathrm{j}\Gamma)$$

and

$$E_{y'}(l) = A$$

where Γ includes the phase shift introduced by the retardation plate.

As these waves pass through the retardation plate, they are shifted by $\pi/4$ radians so that the light appears along the y-axis. We can find the total E field at the output of the modulator by taking the vector sum of these components. Thus we can write

$$\begin{aligned} E_{yo} &= -\frac{A}{\sqrt{2}}\exp(-j\Gamma) + \frac{A}{\sqrt{2}} \\ &= \frac{A}{\sqrt{2}}[1 - \exp(-j\Gamma)] \end{aligned}$$

Now, the intensity of the output light is given by

$$\begin{aligned} I_o &= E_{yo}E^*_{yo} \\ &= \frac{A^2}{2}\left|[1 - \exp(-j\Gamma)][1 - \exp(j\Gamma)]\right| \\ &= A^2(1 - \cos\Gamma) \\ &= 2A^2\sin^2(\Gamma/2) \end{aligned}$$

and the intensity of the input light is

$$\begin{aligned} I_i &= |E_{x'}(0)|^2 + |E_{y'}(0)|^2 \\ &= 2A^2 \end{aligned}$$

Thus we can write the modulator transfer function as

$$\frac{I_o}{I_i} = \sin^2(\Gamma/2)$$

with Γ given by

$$\Gamma = \frac{2\pi l}{\lambda_0}\delta n$$

where δn is defined by (3.124). As we are considering a longitudinal field, we can write

$$\Gamma = \frac{\pi}{\lambda_0}n^3 rV$$

If we also define V_π as

$$V_\pi = \frac{\lambda_0}{n^3 r}$$

where V_π is the voltage needed to produce a phase shift of π radians, we get

$$\frac{I_o}{I_i} = \sin^2\left|\frac{\pi}{2}\;\frac{V}{V_\pi}\right| \tag{3.126}$$

From (3.126) we can see that the transfer function follows a $\sin^2$ form as shown in figure 3.29b. From this figure we can see that the region around the 50 per cent transmission is reasonably linear, and this is where the modulator is normally operated. If we use a quarter-wave retardation plate after the crystal, the modulator will operate in the middle of the linear region resulting in minimal distortion of the modulating signal. A modification to the modulator we are considering is to place the electrodes transverse to the crystal. Under these circumstances, the voltage required to produce a π/4 phase shift is considerably reduced. At present, the maximum speed of these devices is typically 2 GHz.

Example

A bulk modulator uses potassium dihydrogen phosphate, KH_2PO_4, also known as *KDP*. The refractive index of this material at 0.633 μm is approximately 1.5, and the electro-optic coefficient, also at the same wavelength, is 11 × 10^{-12} m/V. Determine the voltage required to produce a phase shift of π radians.

As V_π is given by

$$V_\pi = \frac{\lambda_0}{n^3 r}$$

we find that V_π is 17 kV. With retardation plates this means that we must apply 8.5 kV to give a phase shift of π radians. Such high voltages are typical for the type of modulator we are considering. In view of the magnitude of the drive voltage, we can see that the design of the modulator driver is by no means easy!

We can reduce this voltage to a more manageable one by using *transverse electrodes*. If we use a crystal that is 10 cm long, with a width of 1 cm, we get a V_π voltage of 1.7 kV. As we have already seen, lithium niobate also exhibits electro-optic properties. In $LiNbO_3$, the refractive index at 0.633 μm is 2.2, and r is 31 × 10^{-12} m/V. Thus we find that V_π is 192 volt, for a transverse modulator with a crystal of the same dimensions.

The other type of bulk modulator uses the photo-elastic effect. Figure 3.30a shows the schematic of a typical acousto-optic modulator, *AOM*. The transducer at the bottom of the modulator launches a travelling acoustic wave into the bulk material. As the crests of the sound wave propagate through the crystal, they cause the refractive index of the material to increase due to stress. Thus a travelling wave of high and low refractive index is set up in the crystal. These areas of high refractive index act as mirrors, and so any light incident on the crystal is deflected as shown in figure 3.30b.

Figure 3.30c shows the deflection of a light ray off a high refractive index region. In order for the light to propagate successfully after deflection, the path length AB + BC must be an integral number of optical wavelengths in the crystal. So we can write

$$2\lambda_s \sin\theta_r = \frac{m\lambda_0}{n}$$

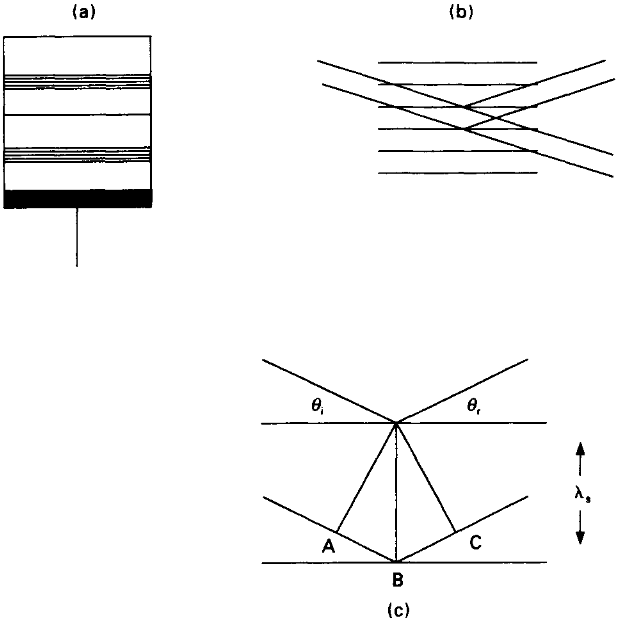


Figure 3.30 (a) Schematic of an acousto-optic modulator; (b) deflection of beam from the crystal; and (c) reflection of beam by area of high refractive index

or

$$\theta_r = \sin^{-1}\left[\frac{m\lambda_0}{n}\frac{1}{2\lambda_s}\right] \tag{3.127}$$

where m is an integer, and n is the refractive index of the material. If $m = 1$, we have *first-order diffraction* of the incident light. With this condition, the angle of incidence equals the angle of reflection, and so

$$\theta_i = \sin^{-1}\left[\frac{\lambda_0}{n}\frac{1}{2\lambda_s}\right] \tag{3.128}$$

Light rays that satisfy (3.128) are said to be incident at the *Bragg angle*. (Indeed, an alternative name for an AOM is a *Bragg cell*.)

We can use the result of (3.127) to deflect the path of any incident light. Let us assume that a carrier wave of frequency f_c is propagating through the crystal. If the incident light hits the crystal at the Bragg angle, equation (3.128), the first-order output beam will be reflected at the same angle. If we now modulate the carrier with a step function, the carrier frequency will change, and the output beam will be deflected. Thus the modulator can act as a beam deflector, or as a simple switch.

One useful feature of AOMs is that the frequency of the output light is either increased or decreased by the acoustic frequency. This is a result of the Doppler shift, which we encounter whenever a sound source moves towards or away from us. Thus we can vary the frequency of a light source by varying the frequency of the carrier wave. We should also note that the light in the higher order diffractions is shifted by different multiples of the carrier. Thus light in the first-order diffraction is shifted by $\pm f_c$, light in the second diffraction is shifted by $\pm 2f_c$, and so on. This feature can be exploited in coherent experiments in the laboratory. Coherent detection uses a local oscillator laser that operates at a different frequency to the source. When the source and local oscillator signals are mixed together, they produce an i.f. equal to the difference between the optical frequencies. In the laboratory, we can generate the local oscillator frequency by passing some of the source light through a Bragg cell. If the frequency of the acoustic signal is that of the i.f., we have no need to use a separate local oscillator laser. Unfortunately the available power in the higher order diffractions is quite low, and so the first-order diffraction is most commonly used.

The maximum operating speed of Bragg cells is reached when the acoustic wavelength in the crystal equals the diameter of the light beam, d. Thus we can write

$$f_{\max} = \frac{v_{ac}}{d} \tag{3.129}$$

where v_{ac} is the velocity of the sound wave in the crystal. Thus we can see that it is important to reduce the spot size of the incident light. Unfortunately this will increase the power density of the light wave, and so crystal damage may occur.

Example

Light of wavelength 0.633 μm, and 200 μm spot size, is incident on a $\mathbf{LiNbO_3}$ crystal. The acoustic velocity of the material is $\mathbf{6.6 \times 10^3}$ m/s. Determine the maximum modulating frequency.

By using (3.129) we get a maximum modulating frequency of 33 MHz. If we reduce the spot size to 50 μm, we get f_{max} = 130 MHz which is typical of such devices.

In view of the low bandwidth of these devices, they are seldom used in optical fibre links.

3.8 Fibre amplifiers and lasers

Considerable work has been carried out [6] in the area of *rare earth doped silica fibre*. If we dope silica fibre with ions of the rare earth erbium, Er^{3+}, we find that the Er^{3+} ions absorb light at 980 nm and 1.49 μm wavelengths, and amplify 1.535 μm wavelength light. (This is similar in operation to the Nd^{3+}:YAG laser we encountered in section 3.6.1.) This is a very important result because it means we can effectively amplify 1.55 μm signals without going through a regenerator. (The regenerator converts the light into an electrical signal for signal processing, before conversion to a light signal again. Thus elimination of this regenerator would be highly advantageous.)

Figure 3.31a shows the schematic of a typical Er^{3+} doped silica fibre amplifier. The pump wavelength is coupled into the Er^{3+} doped fibre, where it produces a population inversion in the Er^{3+} ions. As we have a population inversion, any 1.55 μm light will cause stimulated emission which will amplify the signal. As GaAlAs SLDs emit light at 980 nm, these are often used to provide the pumping light. If the fibre amplifier is used in a single-mode link, the pump wavelength will not propagate very far down the link before it is lost due to radiation.

With a pump power of typically 30 mW, the gain of such an amplifier can be as high as 30 dB. Unfortunately, the amplification process is not ideal. In common with lasers, some spontaneous emission occurs and this is a source of noise for coherent light sources. The minimum theoretical noise factor of an Er^{3+} doped fibre amplifier is 2 (3 dB) and so the signal-to-noise

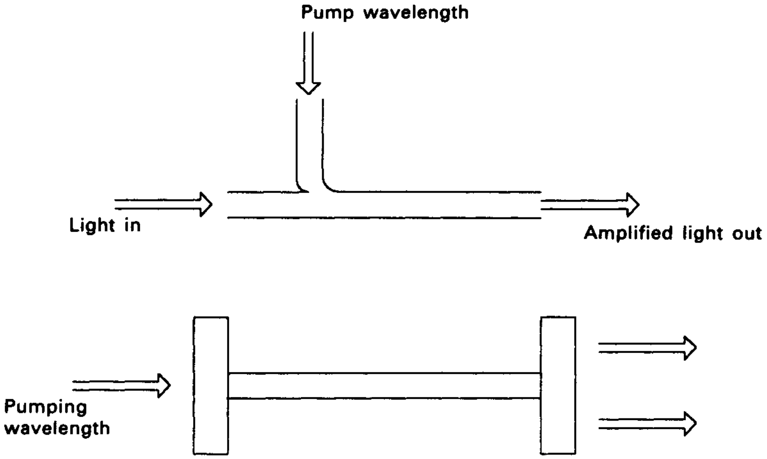


Figure 3.31 (a) Schematic of an Er^{3+} doped silica fibre amplifier.
(b) Schematic of an Er^{3+} doped silica fibre laser

ratio, SNR, is degraded by 2 every time it is amplified. If the distance between the optical amplifiers is such that the gain of each amplifier exactly matches the fibre loss between them, it is found that the SNR falls as the inverse of the transmission distance.

For operation as a laser, we can form a Fabry–Perot cavity resonator by clamping the fibre between two reflecting mirrors, figure 3.31b. In order to pump the laser, we can make the left-hand mirror semi-transparent to the pump wavelength, but totally reflecting to the lasing wavelength. In order to couple light out of the laser, we must make the right-hand mirror semi-transparent to the lasing wavelength, and only slightly reflecting at the pump wavelength. With this arrangement, the pumping wavelength appears through the right-hand mirror, where it can be removed by optical filters or simply lost through radiation in a SM fibre.

4 Photodiodes

In order to convert the modulated light back into an electrical signal, we must use some form of photodetector. As the light at the end of any optical link is usually of very low intensity, the detector has to meet a high performance specification: the conversion efficiency must be high at the operating wavelength; the speed of response must be high enough to ensure that signal distortion does not occur; the detection process should introduce the minimum amount of additional noise; and it must be possible to operate continuously over a wide range of temperatures for many years. A further obvious requirement for optical fibre links is that the detector size must be compatible with the fibre dimensions.

At present, these requirements are met by reverse biased p–n photodiodes. In these devices, the semiconductor material absorbs a photon of light, which excites an electron from the valence band to the conduction band. (This is the exact opposite of photon emission which we examined in the previous chapter.) The photo-generated electron leaves behind it a hole, and so each photon generates two charge carriers. This increases the material conductivity, so-called *photoconductivity*, resulting in an increase in the diode current.

We can modify the familiar diode equation to give

$$I_{\text{diode}} = (I_d + I_s)(\exp[Vq/\eta kT] - 1) \tag{4.1}$$

where I_d is the *dark current*, that is the current that flows when no signal is present, and I_s is the photo-generated current due to the incident optical signal. Figure 4.1 shows a plot of this equation for varying amounts of incident optical power. As we can see, there are three distinct operating regions: forward bias, reverse bias and avalanche breakdown. Under forward bias, region 1, a change in incident power causes a change in terminal voltage, the so-called *photovoltaic mode*. If we operate the diode in this mode, the frequency response of the diode is poor and so photovoltaic operation is rarely used in optical links.

If we reverse bias the diode, region 2, a change in optical power produces a proportional change in diode current. This is the *photoconductive mode* of operation which most detectors use. Under these conditions, the exponential term in (4.1) becomes insignificant, and the reverse bias current is given by

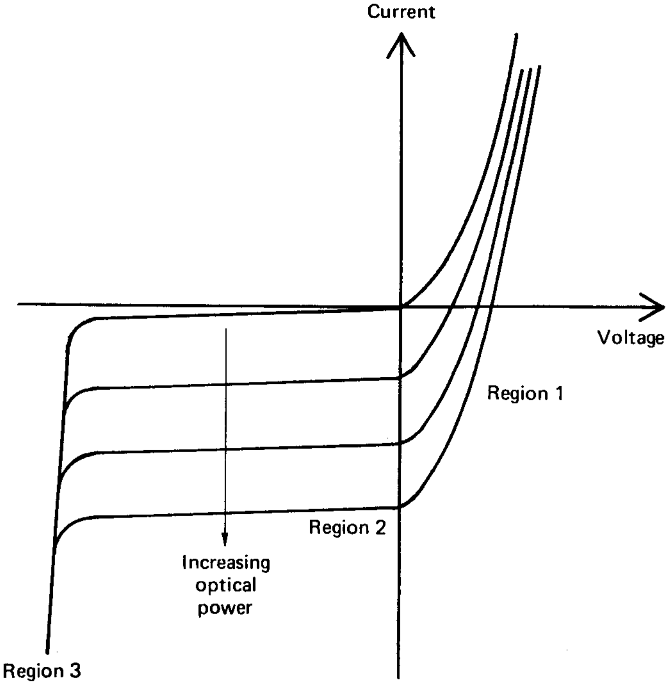


Figure 4.1 *V–I* characteristic of a photodiode, with varying amounts of incident optical power

$$I_{\mathrm{diode}} = I_{\mathrm{d}} + I_{\mathrm{s}} \tag{4.2}$$

We can define the *responsivity* of a photodiode, R_0, as the change in reverse bias current per unit change in optical power, and so efficient detectors need large responsivities.

Avalanche photodiodes, *APDs*, operate in region 3 of the *V–I* characteristic. When biased in this region, a photo-generated electron–hole pair, *EHP*, causes avalanche breakdown, resulting in a large diode current for a single incident photon. Because APDs exhibit carrier multiplication, they are usually very sensitive detectors. Unfortunately the *V–I* characteristic is very steep in this region, and so the bias voltage must be tightly controlled to prevent spontaneous breakdown.

Before we go on to examine the structure and properties of PIN and APD detectors, it will be useful to discuss photoconduction in semiconductor diodes. Although most of our discussion will centre around silicon, we can apply the same basic arguments to other materials.

4.1 Photoconduction in semiconductors

4.1.1 Photon absorption in intrinsic material

As we saw in the introduction to this chapter, when a semiconductor absorbs a photon, an electron is excited from the VB to the CB so causing an increase in conductivity. If the VB electron energy is E_1 and the CB energy level is E_2, then we can relate the change in energy, $E_2 - E_1$, to the wavelength of the incident photon by

$$\lambda_0 = \frac{hc}{E_2 - E_1} \tag{4.3}$$

Now, the lowest possible energy change is the band-gap of the material, and so this results in a cut-off wavelength beyond which the material becomes transparent. These cut-off wavelengths are identical to the emission wavelengths of sources made of the same material, see Table 3.1. Hence, silicon responds to light of wavelengths up to 1.1 μm, whereas germanium photodiodes operate up to 1.85 μm. (It may be recalled from our discussion of photon emission in chapter 3, that sources are made out of direct band-gap materials. However, detectors can be made out of indirect band-gap materials such as Si or Ge.)

The *absorption coefficient*, α, is a measure of how good the material is˙at absorbing light of a certain wavelength. As light travels through a semiconductor lattice, the material absorbs individual photons, so causing the intensity of the light (the number of photons per second) to fall. The reduction is proportional to the distance travelled and so, if the intensity reduces from I to $I - \delta I$ in distance δx, we can write

$$\frac{\delta I}{I} = -\alpha\,\delta x \tag{4.4}$$

We can find the intensity at any point in the lattice by taking the limit of (4.4) and integrating. So, if I_0 is the intensity at the surface, $x = 0$, we can write

$$\ln\frac{I}{I_o} = -\alpha x$$

which gives

$$I = I_o\exp(-\alpha x) \tag{4.5}$$

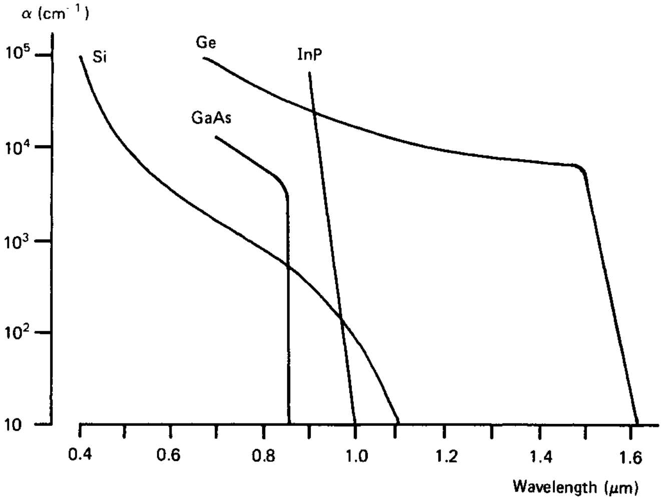


Figure 4.2 Variation of absorption coefficient with wavelength, for a number of semiconductor materials

(We should note that the optical power follows an identical exponential decay.) From our discussion of photo-emission, we can intuitively reason that α will vary with wavelength. Figure 4.2 shows this variation for several semiconductor materials. These plots clearly show an absorption edge which is in close agreement with the cut-off wavelength found from (4.3). The variation of gradient with wavelength is due to the differing density of energy levels in the VB and CB of the material – the same mechanism that causes the spread of wavelength in a semiconductor source.

The absorption coefficient is a very important parameter when considering the design of photodiodes. If the absorbing layer is too thin, a large proportion of the incident light passes straight through, resulting in a low conversion efficiency. If the layer is too thick, the transit time of the carriers is large, limiting the speed of response. Thus there is a trade-off between conversion efficiency and speed of response.

Example

A 40 μm thick silicon detector has an α value of 6.3 × 10^4 m^{-1} when receiving 850 nm wavelength light. Determine the proportion of light that is not absorbed, assuming that no reflection takes place from the surface of the material.

We can use (4.5) to give

$$\frac{I}{I_o} = \exp(-\alpha x)$$

$$= \exp(-6.3 \times 10^4 \times 40 \times 10^{-6})$$

$$= 0.08$$

Thus we can see that only 8 per cent of the light is not absorbed by the detector. This does, however, assume that no light is reflected from the surface of the material. In practice *anti-reflection* coatings are used to minimise reflection from the surface. (Such coatings are similar to the anti-reflection coatings used on spectacles.)

As we shall see later, the carrier transit time of this device is of the order of 100 ps, and so the detector is both efficient and fast.

4.1.2 Photon absorption in reverse-biased p–n diodes

When a photon of light is absorbed by a semiconductor crystal, an electron is excited to the CB where it takes part in the conduction of current. Thus the reverse bias current consists of the normal leakage current that flows even when there is no incident light, the so-called *dark current* – I_d, and the photon generated light, the *signal current* – I_s. Hence we can write the total diode current as

$$I_{\text{diode}} = I_d + I_s \tag{4.6}$$

Let us initially examine the dark current term. In the previous chapter we found that a junction diode under zero bias has a depletion region either side of the junction. Under reverse bias conditions, the external bias voltage has the same polarity as the built-in barrier potential. Thus the depletion region expands and this reduces the current flow. As we saw in the previous chapter, the zero-bias barrier potential is given by, see equation (3.19) given previously

$$V_b = \frac{kT}{q} \ln \left(\frac{N_d N_a}{n_i^2}\right) \tag{4.7}$$

We also found that the barrier potential is given by, see equation (3.27) given previously

$$V_b + V_r = \frac{q}{2\epsilon} (N_d w_{nd}^2 + N_a w_{pd}^2) \tag{4.8}$$

where we have modified (3.27) by the addition of the external bias voltage V_r. Thus we can see that the depletion region width depends on the reverse bias.

By following a similar analysis to that used in section 3.1.4, we find that the dark current is given by

$$I_{\mathrm{d}} = I_{\mathrm{o}}\{\exp(qV/kT) - 1\} \tag{4.9}$$

where V is the total reverse bias voltage, and

$$I_{\mathrm{o}} = \left(\frac{D_{\mathrm{n}}qn_{\mathrm{p}}}{x_{\mathrm{p}}} + \frac{D_{\mathrm{p}}qp_{\mathrm{n}}}{x_{\mathrm{n}}}\right) \times \text{Area}$$

Let us now turn our attention to the illuminated diode. When a semiconductor diode absorbs a photon, an EHP is produced which increases the density of charge carriers. Thus the leakage current, as given by (4.9), is effectively increased. Now, the signal current, I_{s}, is directly dependent on the rate of generation of EHPs which, in turn, is dependent on the number of incident photons per second. With an incident optical power P consisting of photons of energy E_{ph}, the number of photons per second is

$$\begin{aligned} N_{\mathrm{ph}} &= \frac{P}{E_{\mathrm{ph}}} \\ &= \frac{P\lambda_0}{hc} \end{aligned} \tag{4.10}$$

Only some of these photons generate electron–hole pairs. Specifically, the number of carrier pairs generated per second, N, is given by

$$\begin{aligned} N &= \eta N_{\mathrm{ph}} \\ &= \frac{\eta P\lambda_0}{hc} \end{aligned} \tag{4.11}$$

where η is known as the *quantum efficiency*. From our previous discussion, it should be clear that η is highly dependent on α. However, as we shall see presently, η is also dependent on the device structure.

Now, the photo-generated current is equal to the rate of creation of extra charge. Thus I_{s} will be given by

$$\begin{aligned} I_{\mathrm{s}} &= qN \\ &= \frac{q\eta P\lambda_0}{hc} \end{aligned} \tag{4.12}$$

We can rearrange this equation to give the change in current per unit change in optical power, the *responsivity*. Hence

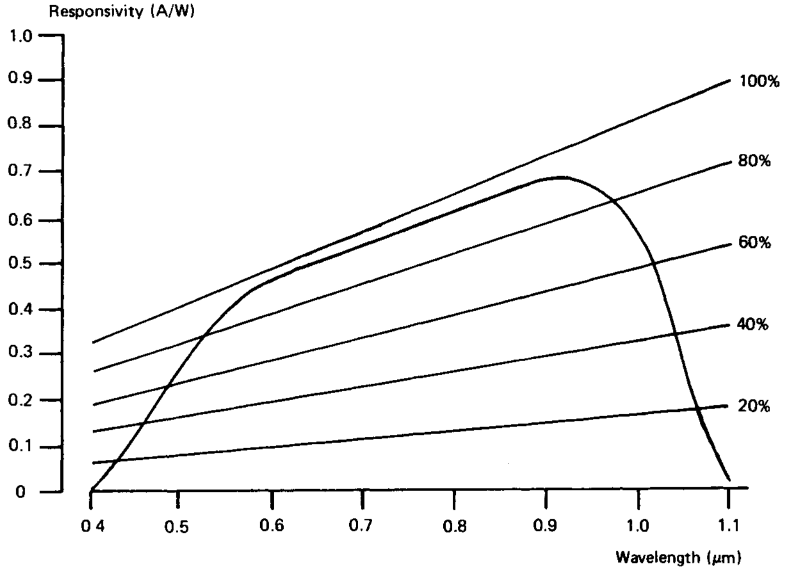


Figure 4.3 Variation of responsivity with wavelength for a typical Si PIN photodiode. Also shown is the theoretical variation of R_0 for a range of quantum efficiencies

$$R_0 = \frac{I_s}{P} = \frac{q\eta\lambda_0}{hc} \tag{4.13}$$

As we can see, R_0 is directly proportional to λ_0, and figure 4.3 shows the theoretical variation of R_0 with λ_0 for various values of η. Also shown is the responsivity characteristic of a typical Si photodiode. At long wavelengths, the curve shows a sharp cut-off coinciding with the cut-off wavelength for silicon. However there is also a lower cut-off region. To explain this we have to examine the structure of a reverse biased p–n photodiode.

Most photodiodes have a planar structure, and so any incident light must first pass through the p-type region before reaching the depletion layer. Because of this, absorption can occur in either the p-type region, the depletion region, or the n-type region (figure 4.4). As the light intensity in the n-type is very low, we can generally ignore absorption in this region. However, absorption in the p-type has a dramatic effect on the quantum efficiency.

Incident light of a low wavelength gives a low penetration depth ($1/\alpha$), resulting in EHP generation in the p-type layer. Unless carrier generation occurs within a diffusion length of the depletion layer boundary, the EHPs recombine and the diode current does not change. However, if photon absorption occurs within a diffusion length of the depletion layer boundary,

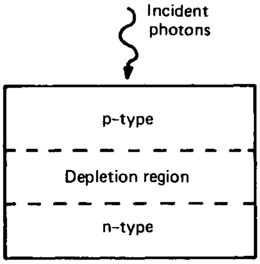


Figure 4.4 Schematic of a reverse biased p–n junction photodiode

the electrons will diffuse into the depletion region. As this is an area of high electric field, they are swept across the diode and so the diode current increases. Thus useful absorption in the p-type only occurs within a diffusion length of the depletion layer. This explains why the quantum efficiency reduces with low wavelength.

Let us now consider absorption in the depletion region. Photon absorption in this region causes the photo-generated EHPs to be swept apart by the electric field – electrons to the n-type, and holes to the p-type. The carriers increase the majority carrier density in these regions, and so the diode current increases. This is obviously more efficient than absorption in the p-type. Hence an efficient photodiode should have a thin p-type layer, less than a diffusion length, and a thick depletion region.

Example

A silicon p$^+$–n junction diode is formed from p-type material with $N_a = 10^{24}$ m^{-3}, and n-type material with $N_d = 10^{21}$ m^{-3}. The diode has a reverse bias voltage of 20 volts across it. Determine the width of the depletion region and the maximum field strength. (The density of thermally generated carriers in silicon is 1.4×10^{16} m^{-3}, and $\epsilon_r = 11.8$.)

From (4.7) the zero-bias barrier potential is

$$\begin{aligned} V_b &= \frac{kT}{q} \ln\left(\frac{N_d N_a}{n_i^2}\right) \\ &= 25 \times 10^{-3} \ln(5.1 \times 10^{12}) \\ &= 0.73 \text{ volt} \end{aligned}$$

Thus the total barrier potential is 20.73 volt.

In order to find the width of the depletion region, we must apply (4.8).

Thus

$$V_b + V_r = \frac{q}{2\epsilon}(N_d w_{nd}^2 + N_a w_{pd}^2)$$

and so

$$20.73 = 7.7 \times 10^{11} w_{nd}^2 + 7.7 \times 10^{14} w_{pd}^2$$

Now, as the total charge in the depletion region equals zero, we have

$$w_{pd} N_a = w_{nd} N_d$$

or

$$w_{pd} = w_{nd} \times 10^{-3}$$

Therefore

$$20.73 = 7.7 \times 10^{11} w_{nd}^2 + 7.7 \times 10^{8} w_{nd}^2$$

giving

$$w_{nd} = 5.2\ \mu\text{m}$$

and

$$w_{pd} = 5.2\ \text{nm}$$

Thus we can see that the width of the depletion layer is approximately 5.2 μm, and it is mainly in the lightly doped n-type material.

The maximum field strength occurs at the junction between the two materials. Thus

$$\begin{aligned} E_{\max} &= -\left.\frac{dV}{dx}\right|_{x=0} \\ &= \frac{qN_d}{\epsilon} w_{nd} \\ &= 8\ \text{MV/m} \\ &= 8\ \text{V/}\mu\text{m} \end{aligned}$$

The diode is now illuminated with 100 nW of 850 nm light. If the p-type is 10 μm thick, and the n-type is 400 μm thick, determine the light intensity in the depletion region. (Assume that the diode is anti-reflection coated, and that the reflectivity of this coating is 20 per cent at 850 nm. Take $\alpha = 6.3 \times 10^4\ \mathrm{m}^{-1}$.)

As the reflectivity of the diode is 20 per cent, only 80 per cent of the incident light will penetrate to the p-type. Thus the optical power at the surface of the p-type is

$$I_p = 80\,\mathrm{nW}$$

This light will be exponentially attenuated as it crosses the diode. Now, the p-type is 10 μm thick, and the depth of the depletion region in the p-type is negligible in comparison. Thus the intensity at the depletion region boundary is

$$\begin{aligned} I_{\mathrm{dep}} &= 80 \times 10^{-9}\exp(-6.3 \times 10^4 \times 10 \times 10^{-6}) \\ &= 43\,\mathrm{nW} \end{aligned}$$

As the depletion region is 5.2 μm thick, the power at the n-type edge of the depletion region is 31 nW. Thus only 12 per cent of the optical power contributes to EHP generation, that is η = 12 per cent.

In practice, the situation is not as bad as this. As we have already discussed, if carriers are generated within a diffusion length of the depletion region, they will contribute to the diode current. If we take $D_p = 13 \times 10^{-4}\ \mathrm{m}^2\,\mathrm{s}^{-1}$, $D_n = 50 \times 10^{-4}\ \mathrm{m}^2\,\mathrm{s}^{-1}$, and $\tau_p = \tau_n = 10\,\mu\mathrm{s}$, we find

$$L_p = 114\,\mu\mathrm{m}, \text{ and}$$

$$L_n = 223\,\mu\mathrm{m}$$

As we can see, the depth of the p-type is less than a diffusion length, and so absorption in this region will produce useful EHPs. However, the thickness of the p-type is greater than a diffusion length. Thus useful absorption takes place over

$$10 + 5.2 + 223 = 238.2\,\mu\mathrm{m}$$

The light power after travelling this distance is

$$80 \times 10^{-9}\exp(-6.3 \times 10^4 \times 238.2 \times 10^{-6}) = 0.024\,\mathrm{pW}$$

and so we can see that almost all the light that penetrates through the anti-reflection coating generates EHPs. Thus the quantum efficiency is approximately 80 per cent.

This example has shown that, although the depletion region can be small, the efficiency of the diode can be quite high due to absorption within a diffusion length of the depletion region. However, a thin depletion region implies a large depletion capacitance which results in a slow detector. In addition, the diffusion of carriers to the depletion region is a slow process, and this reduces the detector speed. The solution is to use a PIN photodiode, and this is the subject of the next section.

4.2 PIN photodiodes

As we have just seen, a simple p^+–n diode can act as an efficient detector. However, the width of the depletion region results in a high diode capacitance. The solution is to use a PIN structure in which, under reverse bias conditions, the thickness of the depletion region is effectively that of the intrinsic material.

Figure 4.5 shows the schematic diagram of a PIN photodiode. Also shown is the variation of the reverse bias electric field intensity, E, across the diode. As can be seen, the field reaches a maximum in the intrinsic layer, the *I-layer*, which is usually high enough to enable the carriers to reach their saturation velocity. As a result, PIN photodiodes are usually very fast detectors.

In the figure, we have labelled the I-layer as n^-. We use this notation to show that the material has been lightly doped with about 10^{19} donor atoms per cubic metre. This is done because it is difficult to produce a totally intrinsic layer, and so the doping controls the diode characteristics. Because

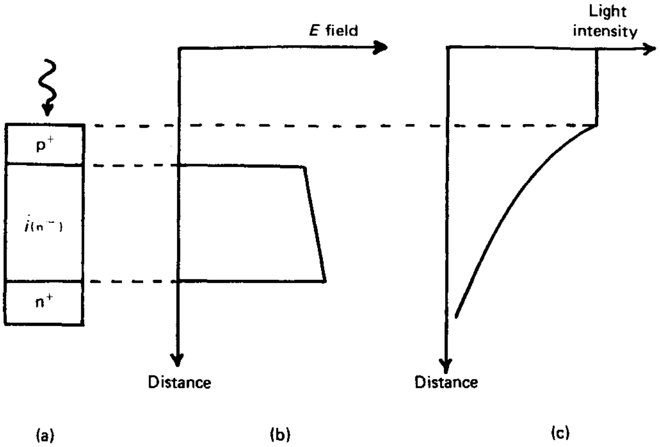


Figure 4.5 (a) PIN photodiode schematic; (b) electric field intensity; and (c) light intensity across the photodiode

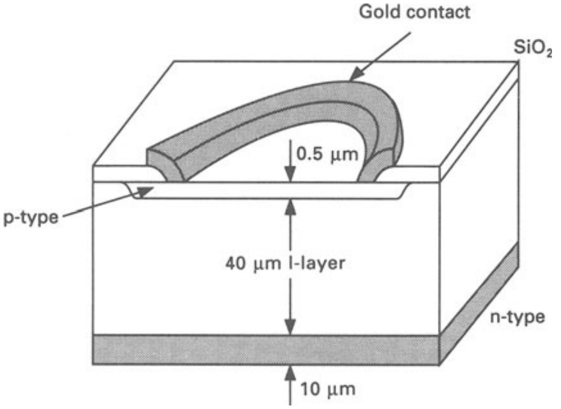


Figure 4.6 Cross-section through a typical silicon PIN photodiode

of this doping, the p- and n-type regions are heavily doped ($p^+ \approx 10^{24}\,m^{-3}$ and $n^+ \approx 10^{22}\,m^{-3}$) in order to approximate to a PIN diode.

4.2.1 Structure

Figure 4.6 shows a typical Si PIN photodiode structure. The diameter of these devices ranges from 50 µm, for high-speed operation, to 200 µm, for low-speed operation. The greater the diode diameter the greater the light collecting capability; however, high-speed operation requires a small detector. In order to avoid this problem, some photodiode packages have a hemispherical lens which collects light from a large area, and focuses it on to a small area detector.

Under reverse bias conditions, all the n^- carriers are swept away, and so the depletion region extends from the p-type right through to the n-type. The bias voltage at which this occurs is known as the *punch-through* voltage. If the bias increases beyond this point, the depletion region will extend beyond the contact rim. Since the SiO_2 layer is transparent to light, the semiconductor can absorb photons that do not pass through the p^+ layer. This absorption mechanism is more efficient than absorption through the p-type layer, and so the overall quantum efficiency can be as much as 85 per cent for infra-red light.

4.2.2 Depletion layer depth and punch-through voltage

In a normal p–n diode, the depletion region extends into both the p- and n-type layers. However, as we have already seen, if the doping in the p-type is higher than in the n-type, giving a p^+–n^- diode, then most of the depletion region will be in the n^- material. As PIN diodes have a $p^+n^-n^+$ structure, nearly all of the depletion region exists in the lightly doped intrinsic layer. If the doping level in the I-layer is N_d then, by applying equation (4.8), we can approximate the width of the depletion layer by

$$d = \left(\frac{2\epsilon_0\epsilon_r(V_b + V_r)}{qN_d}\right)^{\frac{1}{2}} \tag{4.14}$$

where V_b is the zero bias barrier potential ($\approx$ 0.75 V for Si). If the I-layer is completely depleted – the *punch-through condition* – the depletion layer depth will be that of the I-layer. So, we can find the punch-through voltage by rearranging (4.14) to yield V_r.

Example

A Si PIN photodiode has a 40 μm thick I-layer with N_d equal to 10^{19} m^{-3}. Determine the punch-through voltage.

By applying (4.14) we find

$$\begin{aligned} V_b + V_r &= \frac{qd^2}{2\epsilon_0\epsilon_r} N_d \\ &= 12.25 \end{aligned}$$

Thus the punch-through voltage for this diode is 11.5 volt.

As mentioned previously, the electric field intensity is usually made high enough to ensure that the carriers drift at their saturation velocity. In silicon, this occurs at a field strength of 2 V μm^{-1}, which implies a bias voltage of 80 V for a 40 μm thick I-layer. However, it is seldom necessary to operate detectors at such high voltages, because the carrier transit time across the depletion layer is usually insignificant when compared with other speed limiting factors.

4.2.3 Speed limitations

If a photodiode is to detect a digital signal, the sum of the rise- and fall-times of the electrical signal must be less than the interval between optical pulses. If we cannot satisfy this condition, then inter-symbol interference, *ISI*, will occur. Three main factors limit the photodiode response time: carrier diffusion time from the p^+ and n^+ regions; carrier diffusion carrier transit time across the I-layer; and the junction capacitance interacting with any external load resistance.

As we have previously seen, photon absorption can occur in the p^+ region of the photodiode. As the optical power falls exponentially as we move through the diode, this region will produce a considerable number of EHPs. If the EHPs are produced within a diffusion length of the depletion region, the electrons will diffuse into the I-layer where they will ultimately contribute to current flow. In principle, as soon as an EHP is produced in the p^+ region, a current will flow because of the production of a hole. However, the duration of the current pulse will be longer than the optical pulse because the electrons have to cross the I-layer. If the thickness of the p^+ layer is greater than a diffusion length, the maximum transit time of the electrons will be the sum of the electron lifetime in the p^+ region ($\approx$ 10 ns) and the transit time across the I-layer ($\approx$ 0.1 ns). So, in order to produce a fast device, we must ensure that the electrons do not spend a long time in the p^+ layer, that is we must use a p^+ region that is considerably less than a diffusion length.

Let us now consider EHP generation in the I-layer. If photon absorption occurs near the p^+ region, the electrons are accelerated across the I-layer by the electric field. If we consider a 40 μm thick I-layer, then $E = 0.3\ \text{V}\ \mu\text{m}^{-1}$ at a bias of 12 V. The mobility of electrons in intrinsic silicon is $1350\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$ and so the electron velocity is $4 \text{ x } 10^4\ \text{m s}^{-1}$, giving a transit time of approximately 1 ns. If we can operate with a sufficiently high E field, the electrons will reach their saturation velocity of $10^5\ \text{m s}^{-1}$, resulting in a maximum transit time of 0.4 ns (considerably lower than the diffusion time of electrons in the p^+ region).

If EHP generation occurs close to the n^+ region, it is the holes that have to traverse the I-layer. The saturation velocity of holes in silicon is $0.5 \times 10^5\ \text{m s}^{-1}$, and this gives a transit time of 0.8 ns. However, we should remember that the optical power is quite low close to the n^+ region, and so there will not be many EHPs produced. Thus, in general, we can neglect the effects of EHP generation in this region. It is interesting to note that if the diode is illuminated from the n^+ side, it will be the holes that have to traverse the I-layer. As holes are slower than electrons, the falling edge of the current pulse will be much slower than if electrons were the carrier.

So, we have seen that the current pulse produced by an optical signal will have a fast rise-time, with a fall-time governed by diffusion effects and hole transit across the I-layer. Figure 4.7 shows the theoretical impulse response of a silicon PIN photodiode. Also shown are the individual contributions due to electron and hole propagation. As can be seen, it is the diffusion of electrons that causes the poor fall-time.

So far we have ignored the effects of the depletion layer capacitance. If we initially neglect the package capacitance, the diode capacitance will be given by

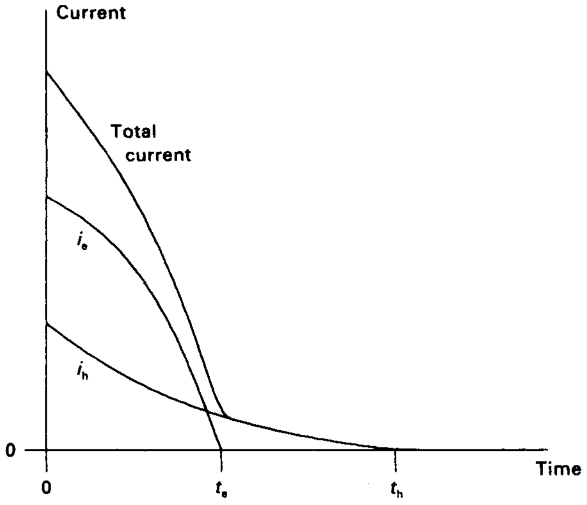


Figure 4.7 Theoretical impulse response of a silicon PIN photodiode

$$C_j = \frac{\epsilon_0 \epsilon_r A}{W} \tag{4.15}$$

where A is the cross-sectional area of the diode, and W is the depletion region thickness. By itself the diode capacitance presents few problems. However, when connected to an external load, the RC time constant may be sufficient to limit the maximum frequency of operation. Total diode capacitances (including package capacitance) range from less than 0.8 pF for high-speed detectors, to 150 pF for low-speed, large area detectors.

Example

A Si PIN photodiode has a 0.5 μm p$^+$ layer, a 40 μm n$^-$ layer, and a 10 μm n$^+$ layer. An anti-reflection coating is used that results in 10 per cent reflection at 850 nm, and the diode has a circular cross-section of diameter 50 μm. Determine the quantum efficiency, and examine the limits on detector speed. Assume that the diode is operated with a reverse bias of 20 V.

The diode is 10 per cent reflective, and so 90 per cent of the incident light reaches the silicon. The dimensions of the p$^+$ and n$^+$ layers are such that any carriers generated in these regions will diffuse into the depletion region. In passing through the diode, 90 per cent of the light is absorbed. Thus the overall quantum efficiency of the diode is 86 per cent.

As regards the speed of the detector, the p^+ region is thin enough for us to neglect any diffusion of carriers to the I-layer. So, the main limit on the response time is the transit time of carriers across the I-layer. In order to find this transit time, we need to find the electric field strength across the I-layer. This field strength is

$$\begin{aligned} E &= \frac{20}{40 \times 10^{-6}} \\ &= 0.5 \text{ V}/\mu\text{m} \end{aligned}$$

If we take an electron mobility of 1350 $\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$, we get an electron velocity of 6.75×10^4 $\text{m}\,\text{s}^{-1}$ which results in an electron transit time of approximately 600 ps. By following a similar analysis for the holes (500 $\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$ mobility) we get a hole transit time of 1.6 ns. As we have already noted, the optical power is low close to the n^+ region, and so we can usually neglect the effects of hole transit time.

As regards the diode capacitance, the cross-sectional area of the diode is approximately 2×10^3 μm^2 which results in a depletion capacitance of 5 fF. This value is low enough for any package capacitance to dominate.

4.2.4 Photodiode circuit model

Figure 4.8 shows an equivalent circuit for a PIN photodiode, which is connected to an external load feeding an amplifier. In this diagram, the photoconductive current has been modelled as a current source, I_s, whose magnitude depends on the incident optical power. The constant current source, I_d, models the dark current, that is the leakage current and any photoconductive current due to background radiation. The shunt resistance, R_j, represents the slope of the reverse bias characteristic, and the series resistance, R_s, is that

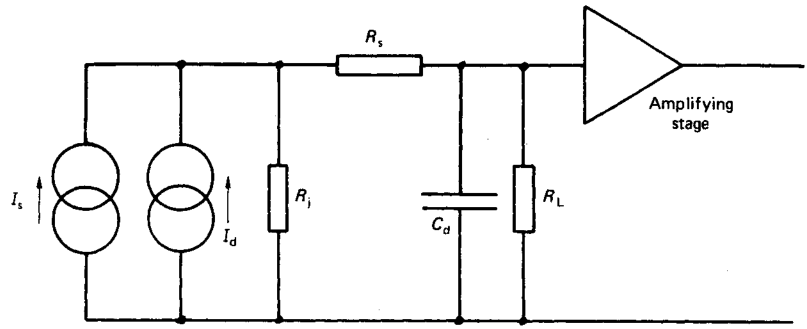


Figure 4.8 A circuit model for a typical photodiode

of the bulk semiconductor and the contact resistance. The *total* diode capacitance, C_d, models the depletion and diffusion capacitances. The load resistor, R_L, shunts this capacitance, and it is this time constant that usually limits the speed of response.

In general, we can ignore R_j and R_s, and so the bandwidth of the detector is given by

$$f = \frac{1}{2\pi R_L C_d} \tag{4.16}$$

We should also note that any following amplifier will have a time constant, and this may prove to be the limiting factor.

4.2.5 Long-wavelength PIN photodiodes

At long wavelengths, >1 μm, silicon becomes transparent. Thus detectors for 1.3 and 1.55 μm wavelengths must be made out of low band-gap materials. Germanium has a band-gap of 0.67 eV, corresponding to a cut-off wavelength of 1.85 μm, and so would appear to be a suitable material. However the low band-gap means that Ge photodiodes exhibit a high leakage current (>100 nA). As we will see later, the dark current is an additional source of noise, and so Ge PIN photodiodes are rarely used in long-haul routes.

When we discussed light sources, we saw that InGaAsP emits light in the band 1.0–1.7 μm. Thus detectors made of a similar material should respond to 1.3 or 1.55 μm light. In practice, we can use an InGaAs alloy, where the proportions of In and Ga alter the band-gap. Thus the diode can be tailored to respond to light of a specific wavelength. As an example, a diode fabricated out of $In_{0.53}Ga_{0.47}As$ has a band-gap of 0.47 eV which gives a cut-off wavelength of 1.65 μm. The dark current of these devices is usually about 10 nA.

The absorption coefficient of InGaAs at 1.3 μm is about $5 \times 10^5\ m^{-1}$, which results in a penetration depth of around 2 μm. Therefore the dimensions of a long-wavelength PIN detector are much smaller than that of a Si photodiode, leading to a better frequency response. Figure 4.9 shows the structure of a typical InGaAs photodiode.

The quantum efficiency of this particular device is quite low, ≈0.4, because the p^+ layer absorbs 40 per cent of the incident power. However, the InP substrate is transparent to light of wavelength greater than 0.92 μm. Thus illumination from the rear of the device will increase the quantum efficiency to about 90 per cent. Such a device is known as a *rear-entry* or *substrate-entry* photodiode.

The I-layer is usually doped to a level of $10^{21}\ m^{-3}$ and this, together with an ϵ_r of 14, gives a punch-through voltage of 10 V. This results in an E field of 2.5 $V\,\mu m^{-1}$ which is well above the 1 $V\,\mu m^{-1}$ required for the

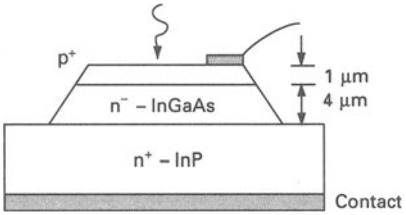


Figure 4.9 Cross-section through a typical long-wavelength PIN photodiode

carriers to reach their saturation velocity of about 1×10^5 m s^{-1}. So the transit time across the I-layer is in the region of 40 ps and, as there is little absorption in the p^+ layer, the device is inherently very fast.

The junction capacitance of a typical 50 μm diameter device is 60 fF, which is considerably less than that of a Si diode of the same dimensions. However, any package capacitance may cause the capacitance to rise to 0.8 pF or more. Thus for high-speed operation, hybrid thick-film receivers use unpackaged photodiodes.

4.3 Avalanche photodiodes (APDs)

When a p–n junction diode has a high reverse bias applied to it, breakdown can occur by two separate mechanisms: direct ionisation of the lattice atoms, *zener breakdown*; and high velocity carriers causing impact ionisation of the lattice atoms, *avalanche breakdown.* APDs use the latter form of breakdown.

Figure 4.10 shows the schematic structure of an APD. By virtue of the doping concentration and physical construction of the n^+p junction, the *E* field is high enough to cause impact ionisation. Under normal operating bias, the I-layer (the p^- region) is completely depleted. This is known as the *reach-through* condition, and so APDs are sometimes known as *reach-through APDs* or *RAPDs.*

Like the PIN photodiode, light absorption in APDs is most efficient in the I-layer. In this region, the *E* field separates the carriers, and the electrons drift into the avalanche region where carrier multiplication occurs. We should note however, that an APD biased close to breakdown could breakdown owing to the reverse leakage current. Thus APDs are usually biased just below breakdown, with the bias voltage being tightly controlled.

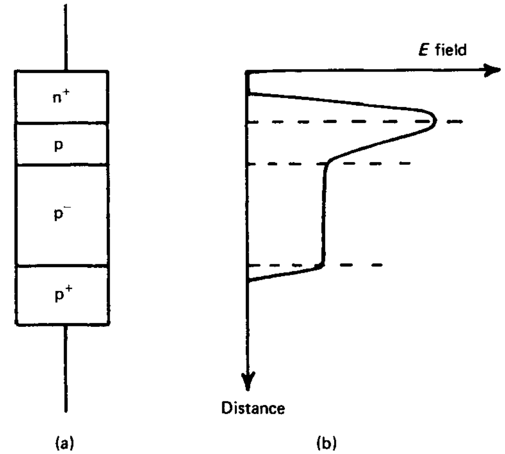


Figure 4.10 (a) APD schematic and (b) variation of electric field intensity across the diode

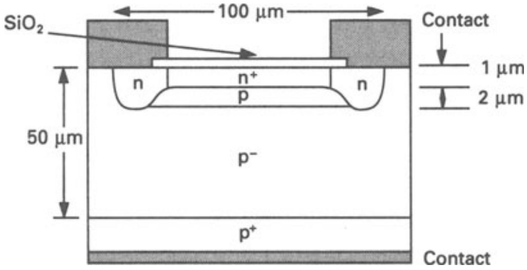


Figure 4.11 Cross-section through a typical silicon APD

4.3.1 APD structures

Figure 4.11 shows the cross-section of a typical Si APD. In order to minimise photon absorption in the n^+p region, the $n^+$and p layers are made very thin. In practice, these layers have doping concentrations of around 10^{24} and 10^{21} m^{-3}, and the p^+ and p^- layers have concentrations of 10^{24} and 10^{20} m^{-3} respectively. These parameters, together with the device dimensions, result in a reach-through voltage of ≈ 40 V, and an avalanche breakdown E field value of ≈ 18 V μm^{-1}. (The reverse breakdown voltage for a typical device lies in the range 200–300 V.) An n-type guard ring serves to increase the peripheral breakdown voltage, causing the n^+p junction to breakdown before the pn junction.

For operation at 1.3 and 1.55 µm wavelengths, we can use germanium APDs. However, as we have already noted, these diodes exhibit a high dark

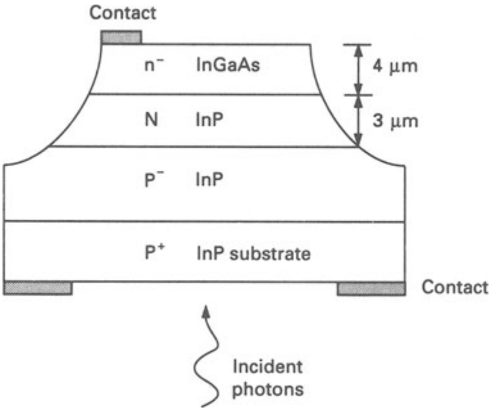


Figure 4.12 Cross-section through a typical long-wavelength heterojunction APD

current, giving a noisy detection process. In spite of their drawbacks, several different Ge APD structures are being investigated, and such devices may find applications in the future.

Like long-wavelength PIN photodiodes, APDs can be made out of InP/InGaAs, and figure 4.12 shows a typical structure. In this particular design, the InGaAs layer absorbs the light and, because InP is transparent to long-wavelength light, the device can be either front or rear illuminated. The *E*-field in the fully depleted region causes separation of the photo-generated carriers. However, because of the n^-N heterojunction, only holes cause breakdown in the N-type region.

Such APDs usually operate with a sufficiently high *E*-field in the absorbing region, >1 V μm^{-1}, to accelerate the carriers to their saturation velocity, and a field strength in the N-type large enough to cause breakdown, >20 V μm^{-1}. In practical devices, the operating fields are typically 15 V μm^{-1} and 45 V μm^{-1}, and so these conditions are satisfied. The bias voltage at which these fields occur is usually around 50 V.

4.3.2 Current multiplication

In an APD, avalanche multiplication increases the primary current, that is the unmultiplied photocurrent given by (4.12). Thus we can write the responsivity as,

$$R_0 = \frac{Mq\eta\lambda_0}{hc} \tag{4.17}$$

where M is the multiplication factor. It therefore follows that M is given by

$$M = \frac{I_m}{I_s} \tag{4.18}$$

where I_m is the average total multiplied diode current. In order for M to be large, there must be a large number of impact ionisation collisions in the avalanche region. The probability that a carrier will generate an electron–hole pair in a unit distance is known as the *ionisation coefficient* (α_e for electrons, and α_h for holes). Obviously, M is highly dependent on these coefficients which, in turn, depend upon the E-field and the device structure. After a straightforward analysis, it can be shown that M is given by

$$M = \frac{1 - k}{\exp[-(1 - k)\alpha_e W] - k} \tag{4.19}$$

where k is α_e/α_h, and W is the width of the avalanche region. So, a large M requires a low value of k. In silicon, k ranges from 0.1 to 0.01, and this leads to values of M ranging from 100 to 1000. However, in germanium and III–V materials, k ranges from 0.3 to 1 and, in practice, it is difficult to fabricate and control devices with gains above 15.

As expected, M is highly dependent on the bias voltage. An empirical relationship which shows this dependency is

$$M = \frac{1}{1 - (V/V_{br})^n} \tag{4.20}$$

where V_{br} is the device breakdown voltage, and n is an empirical constant, <1. Now, n and V_{br} are dependent on temperature as shown by

$$V_{br}(T) = V_{br}(T_0) + a(T - T_0)$$

and (4.21)

$$n(T) = n(T_0) + b(T - T_0)$$

where a and b are empirical constants, <1. So, as figure 4.13 shows, M depends on both the bias voltage and the temperature.

4.3.3 Speed limitations

Several factors will limit the speed of response of an APD; the RC time constant of the detector circuitry; the drift time of carriers to the avalanche region, t_d; the time taken to achieve avalanche breakdown, t_a; and the time taken to sweep the avalanche produced carriers through the diode, t_s. Of

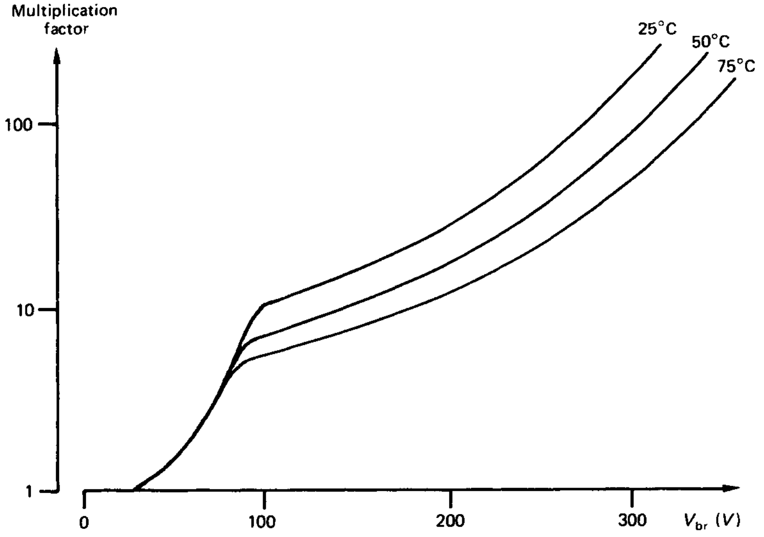


Figure 4.13 Theoretical variation of multiplication factor, M, with reverse bias voltage, V_{br}, for three different temperatures. Unity gain has been taken at V_{br} = 30 V

these four factors, t_a and t_s represent delays which are additional to those experienced with PIN photodiodes.

A full analysis of the APD response times reveals that the intrinsic time constant, τ, for an APD with $k << 1$ is given by

$$\begin{aligned}\tau &= t_d + t_a + t_s \\ &= \frac{W_i}{v_{se}} + \frac{MkW_a}{v_{se}} + \frac{1}{v_{sh}}(W_a + W_i) \qquad (4.22)\end{aligned}$$

where W_i and W_a are the widths of the intrinsic and avalanche regions, and v_{se} and v_{sh} are the saturation velocities of electrons and holes respectively. As can be seen, a fast diode requires $k << 1$. Silicon has $k \approx 0.05$, and so this is a very popular material. In general, APDs have a slower response time than an equivalent PIN photodiode, and so gain has been traded for a reduction in bandwidth. It should be noted that the *RC* time constant of the diode capacitance and the external load is likely to limit the overall frequency response.

4.4 Photodiode noise

The signal at the end of any optical link is often highly attenuated, and so any receiver noise should be as small as possible. The minimum signal-to-noise ratio, *S/N*, required for satisfactory detection is often specified for a particular application. This is dealt with in greater detail in the next chapter; however the S/N can be written as

$$\frac{\mathrm{S}}{\mathrm{N}} = \frac{<I_s^2>}{<i_n^2>_{\mathrm{pd}} + <i_n^2>_{\mathrm{c}}} \tag{4.23}$$

where $<I_s^2>$ is the mean square, *m.s.*, value of the photodiode signal current, $<i_n^2>_{\mathrm{pd}}$ is the m.s. value of the photodiode noise, and $<i_n^2>_{\mathrm{c}}$ is the m.s. value of the following amplifier noise when referred to the detector terminals. Even if the following amplifier is noiseless, there is still some photodiode noise, and it is this noise source that concerns us here.

There are three main components to the photodiode noise: *quantum noise*, $<i_n^2>_{\mathrm{Q}}$, due to quanta of light-generating packets of electron–hole pairs; thermally generated dark current, $<i_n^2>_{\mathrm{DB}}$, occurring in the photodiode bulk material; and surface leakage current, $<i_n^2>_{\mathrm{DS}}$. (There is an extra noise component due to the ambient light level causing additional dark current; however, careful shielding of the detector can reduce this to a minimum.) Thus we can write $<i_n^2>_{\mathrm{pd}}$ as

$$<i_n^2>_{\mathrm{pd}} = <i_n^2>_{\mathrm{Q}} + <i_n^2>_{\mathrm{DB}} + <i_n^2>_{\mathrm{DS}} \tag{4.24}$$

4.4.1 PIN photodiode noise

No current multiplication occurs in a PIN detector and so, with a receiver noise equivalent bandwidth of B_{eq}, we can write the detector S/N as

$$\frac{\mathrm{S}}{\mathrm{N}} = \frac{<I_s^2>}{<i_n^2>_{\mathrm{Q}} + 2qI_{\mathrm{DB}}B_{\mathrm{eq}} + 2qI_{\mathrm{DS}}B_{\mathrm{eq}}} \tag{4.25}$$

where I_{DB} and I_{DS} are the bulk and surface leakage currents respectively. However, if the leakage currents are negligible

$$\frac{\mathrm{S}}{\mathrm{N}} = \frac{<I_s^2>}{<i_n^2>_{\mathrm{Q}}} \tag{4.26}$$

With this condition, we must take account of the quantum nature of light and so the S/N defined by (4.26) is known as the *quantum limit*. We can determine $<i_n^2>_{\mathrm{Q}}$ from a knowledge of the photon statistics.

Photons arrive at the detector at random intervals, but with a constant *average* rate. So, in a certain time interval, we can expect to receive an average of m photons but, because of the random arrival of photons, we actually receive n photons. The photon arrival follows a Poisson probability distribution and so, the probability that the resultant number of detected photons is n, with an expected number of m, is

$$p(n) = \mathrm{Pos}[n, m] = \frac{m^n e^{-m}}{n!} \tag{4.27}$$

If we take a quantum efficiency of unity, the number of EHPs is n. In a digital system, a decision must be made as to whether a 1 or a 0 was sent; however, noise will corrupt the signal levels so that a 1 signal occasionally turns into a 0, and a 0 turns into a 1. In an ideal receiver, the detection of a single EHP results in a logic 1, while the absence of any signal current results in a logic 0. As the quantum noise is dependent on the *presence* of an optical signal, it will only corrupt logic 1 signals (assuming no dark current, and no other noise sources). So, the condition for a logic 1 detection error is that m photons are received, but $n = 0$ photons are detected. If we assume that the probability of sending a logic 1 is the same as for a logic 0, that is, they are *equiprobable*, we can write the probability of an error, P_e, as

$$\begin{aligned} P_e &= \frac{1}{2}\,(P(0|1) + P(1|0)) \\ &= \frac{1}{2}\left(\frac{m^0 e^{-m}}{0!} + 0\right) \\ &= \frac{1}{2}\,e^{-m} \end{aligned} \tag{4.28}$$

So, for a typical error rate of 1 bit in 10^9, we require an average of $m = 21$ EHPs. These carriers are generated by $21/\eta$ photons arriving in a bit-time $1/B$, where B is the data-rate and η is the quantum efficiency. For equiprobable 1s and 0s, the *mean* optical power required, P, is

$$P = \frac{1}{2} \times \frac{21}{\eta} \times \frac{hc}{\lambda_0} \times B \tag{4.29}$$

Example

Determine the quantum limit for a PIN photodiode, with unity quantum efficiency, that detects 34 Mbit/s at a wavelength of 850 nm. Assume an error rate of 1 bit in 10^9.

If we use (4.29) we find that the mean optical power for the specified error rate is

$$\begin{aligned} P &= 80 \text{ pW} \\ &= -71 \text{ dBm} \end{aligned}$$

This represents a very high sensitivity. However, this result ignores the noise from the photodiode dark current and following amplifier stages. In practice, these effects will tend to limit the receiver sensitivity. In spite of this, coherent detection systems (which we examine in the final chapter) can achieve sensitivities better than the direct detection quantum limit we are considering here.

Before we comment on APD noise, we should note that the spectral density of the quantum noise is simply the shot noise expression, given by

$$<i_n^2>_Q = 2q<I_s> \qquad \text{A}^2/\text{Hz} \tag{4.30}$$

where $<I_s>$ is the mean signal current. (This result arises from the statistics of the Poisson process.) Therefore we can write the quantum limited S/N as

$$\frac{\text{S}}{\text{N}} = \frac{<I_s^2>}{2q<I_s>B_{eq}} \tag{4.31}$$

4.4.2 APD noise

In an APD, the avalanche gain multiplies the *primary current*. (We define the primary current as that produced by a unity gain photodiode.) Since the gain is statistically variant, that is not all of the photo-generated carriers undergo the same multiplication, we define the *average* gain as M. As we have seen, the Poisson distribution describes photon arrival and hence the signal current. So, we can find the APD signal current by performing a convolution type process between the Poisson distribution of the primary current, and the avalanche gain distribution. The resulting expression is very complicated, and so it is common practice to approximate the APD current to a Gaussian distribution with a mean value of $<I_s>M$, and a noise current spectral density of $2q<I_s>M^2F(M)$. The term $F(M)$ is known as the *excess noise factor*, and we include it to account for the random fluctuations of the APD gain about the mean. We can approximate $F(M)$ by

$$F(M) = M^x \tag{4.32}$$

where x is an empirical constant which is less than unity. From our earlier discussion of avalanche multiplication, it should be apparent that $F(M)$

depends on the value of k and the type of carrier undergoing multiplication. Detailed analysis (McIntyre [1]) shows that $F(M)$ can be approximated by

$$F_e(M) = kM_e + \frac{(1 - k)}{M_e}(2M_e - 1) \tag{4.33}$$

for electron avalanche, and

$$F_h(M) = \frac{M_h}{k} + \frac{(1 - 1/k)}{M_h}(2M_h - 1) \tag{4.34}$$

for hole avalanche. These equations clearly show the need to fabricate devices out of materials with low values of k.

Example

A Si APD has a gain of 100 and k = 0.02. Determine the excess noise factor for electrons, and compare it with that obtained from a Ge APD with M = 20 and k = 0.5.

By applying (4.33) we find that $F_e(M)$ for electrons in the silicon APD is

$$\begin{aligned} F_e(M) &= kM_e + \frac{(1 - k)}{M_e}(2M_e - 1) \\ &= 0.02 \times 100 + \frac{(1 - 0.02)}{100}(2 \times 100 - 1) \\ &\approx 4 \end{aligned}$$

For the germanium diode however, $F_e(M)$ is

$$\begin{aligned} F_e(M) &= 0.5 \times 20 + \frac{(1 - 0.5)}{20}(2 \times 20 - 1) \\ &\approx 11 \end{aligned}$$

Thus we can see that as well as having higher gains than Ge APDs, Si APDs have a lower excess noise factor.

As regards the S/N for an APD, we have to take account of signal multiplication, and the noise terms due to surface and bulk leakage currents. Thus we can write the S/N as

$$\frac{\mathrm{S}}{\mathrm{N}} = \frac{<I_s^2>M^2}{2q<I_s>M^2F(M)B_{eq} + 2qI_{DB}M^2F(M)B_{eq} + 2qI_{DS}B_{eq} + <i_n^2>_c} \quad (4.35)$$

If we can ignore the leakage currents and assume a noiseless receiver, we can write (4.35) as

$$\frac{\mathrm{S}}{\mathrm{N}} = \frac{<I_s^2>M^2}{2q<I_s>M^2F(M)B_{eq}} \quad (4.36)$$

Comparison with the PIN equation (4.31) reveals that, because of the excess noise factor, an APD receiver cannot approach the quantum limit. Indeed, at high levels of received signal power, the use of an APD could be a disadvantage because of the signal dependent shot noise.

However, as we will presently see, we can use an APD to increase the signal-to-noise ratio of an ordinarily noisy optical receiver. If the noise from the following amplifier stage is greater than that of the detector noise, the S/N approximates to

$$\frac{\mathrm{S}}{\mathrm{N}} = \frac{<I_s^2>M^2}{<i_n^2>_c} \quad (4.37)$$

Thus, because of the M^2 term, the S/N for an APD receiver can be greater than that for a PIN receiver. As the following example demonstrates, in most APD receivers the sensitivity advantage reduces because we cannot ignore the detector noise. (Section 5.4 in the next chapter also deals with this problem.)

Example

Two receivers are available for use in an optical link: one has a total input equivalent noise current of 10^{-15} A^2, while the other has an $<i_n^2>_c$ of 10^{-18} A^2. Both receivers have a noise equivalent bandwidth of 27 MHz. Determine the S/N of the receivers if the mean received signal power is 100 nW. Assume a diode responsivity of 0.5 A/W, M = 100, $F(M)$ = 4, I_{DB} = I_{DS} = 10 nA.

As the diode responsivity is 0.5 A/W, the primary signal current is

$$\begin{aligned} <I_s> &= 0.5 \times 100 \text{ nA} \\ &= 50 \text{ nA} \end{aligned}$$

In order to find S/N, we can use (4.35) to give, for the noisy receiver,

$$\frac{\mathrm{S}}{\mathrm{N}} = \frac{<I_s^2>M^2}{2q<I_s>M^2F(M)B_{eq} + 2qI_{DB}M^2F(M)B_{eq} + 2qI_{DS}B_{eq} + <i_n^2>_c}$$

$$= \frac{(50 \times 10^{-9})^2 \times 100^2}{1.73 \times 10^{-14} + 3.5 \times 10^{-15} + 8.6 \times 10^{-20} + 1 \times 10^{-15}}$$

$$= \frac{2.5 \times 10^{-11}}{2.18 \times 10^{-14}}$$

$$= 1.15 \times 10^3$$

$$= 30.60 \text{ dB}$$

For the less noisy receiver we get an S/N of 30.80 dB. Thus we can see that there is negligible difference between the two receivers. This is because the dominant noise source is the first term in the denominator of (4.35) – the signal-dependent noise.

It is instructive to compare these sensitivities to those that would be achieved if a PIN photodiode, with the same basic parameters, were used. In this instance, the S/N for the noisy receiver is

$$\frac{\mathrm{S}}{\mathrm{N}} = \frac{(50 \times 10^{-9})^2}{4.32 \times 10^{-19} + 8.6 \times 10^{-20} + 8.6 \times 10^{-20} + 1 \times 10^{-15}}$$

$$= \frac{2.5 \times 10^{-15}}{1 \times 10^{-15}}$$

$$= 2.5$$

$$= 4.00 \text{ dB}$$

while the S/N for the less noisy receiver is 31.95 dB. So, a considerable sensitivity advantage results from using the less noisy receiver.

We can see from these figures that the signal dependent shot noise associated with an APD tends to mask the effects of receiver noise. However, in a PIN photodiode receiver, a considerable sensitivity advantage results from using a low noise receiver. We can also see that the PIN detector and a low noise receiver offers a slightly better sensitivity than the equivalent APD receiver. If we increase the mean signal power to 1 μW, the PIN detector gives a 5 dB advantage over the APD detector if the low noise receiver is used. This is due to the effect of the excess noise factor which means that an APD cannot reach the quantum limit, whereas a PIN detector can.

So, in general, we can conclude that we can use an APD to increase the sensitivity of a noisy receiver. However, a PIN detector gives a better sensitivity when detecting high power levels. The exact advantage depends on the bulk leakage current, the magnification factor, and the excess noise factor.

For further background reading to this chapter, see references [2], [3] and chapter 3 of reference [4].

5 Introduction to Receiver Design

The basic structure of an optical receiver (figure 5.1) is similar to that of a direct detection r.f. receiver: a low-noise preamplifier, the *front-end*, feeds further amplification stages, the *post-amplifier*, before filtering. An important point to note is that the pre- and post-amplifiers are usually non-saturating. (If the amplifiers did saturate, charge storage in the transistors would tend to limit the maximum detected data-rate.) Because of this, we can use the same pre- and post-amplifier combination to detect analogue or digital signals. The difference between the two receivers arises from the way they process the signals after amplification. As digital optical communications systems are quite common, most of the work presented is devoted to a performance analysis of digital receivers. However, analogue systems are used to transmit composite video and signals from optical fibre sensors, and so we will consider analogue receiver performance towards the end of this chapter.

Although preamplifier design is dealt with in the next chapter, we must make certain assumptions regarding its performance: the bandwidth must be large enough so as not to distort the received signal significantly; and its gain function must be high enough so that we can neglect any noise from the following stages. As we shall see later, the requirement to minimise the noise implies restricting the receiver bandwidth. However, a low bandwidth results in considerable inter-symbol interference, *ISI*, and so the receiver bandwidth is a compromise between minimising the noise and ISI.

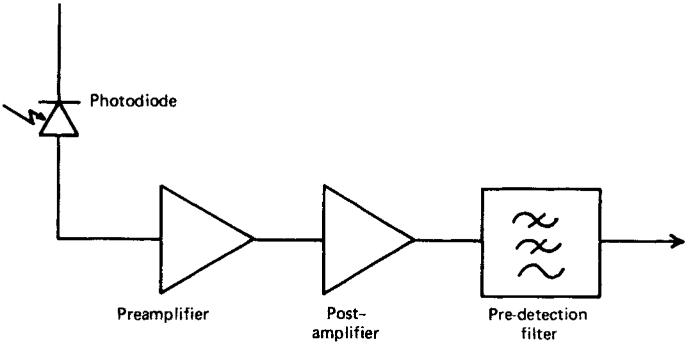


Figure 5.1 The basic structure of an optical receiver

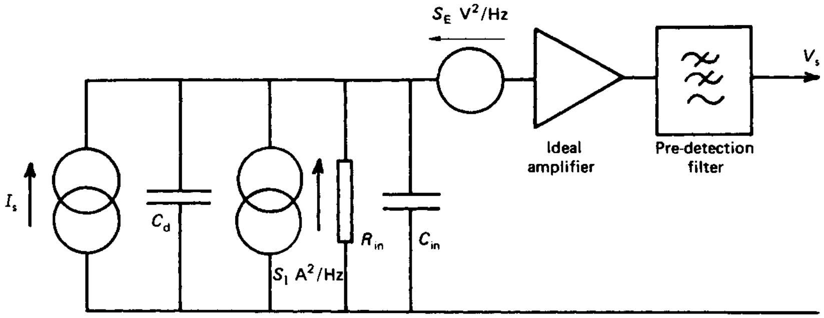


Figure 5.2 A.c. equivalent circuit of an optical receiver

5.1 Fundamentals of noise performance

In order to examine the noise performance of an optical receiver, and hence determine its sensitivity, we shall consider the receiver as a linear channel, with the a.c. equivalent circuit shown in figure 5.2.

An ideal current source, shunted by the detector capacitance, C_d, models the photodetector, which feeds the parallel combination of R_{in} and C_{in}, modelling the input impedance of the preamplifier. The pre- and post-amplifiers are modelled as a single voltage amplifier, with transfer function $A(\omega)$, the output of which feeds the pre-detection filter. If we initially neglect the photodiode noise, then the only noise in the receiver will be due to the preamplifier. A shunt noise generator S_I, with units of A^2/Hz, models the noise current due to the preamplifier first stage, and the photodiode load resistor. The series noise generator S_E, with units of V^2/Hz, models the preamplifier series noise sources. (The reason for the inclusion of this generator will become apparent when we consider preamplifier design in the next chapter.)

In order to determine the signal-to-noise ratio, *S*/*N*, at the output of the pre-detection filter, we need to find the receiver transfer function. Because the input signal is a current, I_s, and the output is a voltage, V_s, the transfer function is a *transimpedance*, $Z_T(\omega)$, given by

$$Z_T(\omega) = \frac{V_s}{I_s} \tag{5.1}$$

From figure 5.2, we see that

$$V_s = I_s Z_{in} A(\omega) H_f(\omega) \tag{5.2}$$

where Z_{in} is the total input impedance, that is the parallel combination of

R_{in} and the *total* input capacitance ($C_{\mathrm{d}} + C_{\mathrm{in}}$), and $H_{\mathrm{f}}(\omega)$ is the pre-detection filter transfer function. Thus we can express $Z_{\mathrm{T}}(\omega)$ as

$$Z_{\mathrm{T}}(\omega) = Z_{\mathrm{in}}A(\omega)H_{\mathrm{f}}(\omega) \tag{5.3}$$

If we now turn our attention to the noise sources, we can see that the series noise generator produces a m.s. *input* noise current of

$$\frac{S_{\mathrm{E}}}{[Z_{\mathrm{in}}]^2} \qquad \text{or} \qquad S_{\mathrm{E}}[Y_{\mathrm{in}}]^2 \qquad \mathrm{A}^2/\mathrm{Hz}$$

If we assume that the two noise sources are independent of each other, that is *uncorrelated*, then the total equivalent input noise current spectral density, $S_{\mathrm{eq}}(f)$, will be given by

$$S_{\mathrm{eq}}(f) = S_{\mathrm{I}} + S_{\mathrm{E}}[Y_{\mathrm{in}}]^2$$

Noting that

$$Y_{\mathrm{in}} = \frac{1}{R_{\mathrm{in}}} + \mathrm{j}\omega C_{\mathrm{T}}$$

where C_{T} is $C_{\mathrm{d}} + C_{\mathrm{in}}$, we can write $S_{\mathrm{eq}}(f)$ as

$$S_{\mathrm{eq}}(f) = S_{\mathrm{I}} + S_{\mathrm{E}}\left(\frac{1}{R_{\mathrm{in}}^{\ 2}} + (2\pi C_{\mathrm{T}})^2 f^2\right) \tag{5.4}$$

Thus we can see that the equivalent input noise current spectral density consists of two terms: a frequency-independent term; and a term that varies according to f^2. (We will return to this point in the next chapter.)

Now, with $S_{\mathrm{eq}}(f)$ given by (5.4), we can write the total m.s. output noise voltage, $<n^2>_{\mathrm{T}}$, as

$$\begin{aligned} <n^2>_{\mathrm{T}} &= \int_0^\infty S_{\mathrm{eq}}(f)[Z_{\mathrm{T}}(\omega)]^2 \,\mathrm{d}f \\ &= \left(S_{\mathrm{I}} + \frac{S_{\mathrm{E}}}{R_{\mathrm{in}}^{\ 2}}\right) \int_0^\infty [Z_{\mathrm{T}}(\omega)]^2 \,\mathrm{d}f \\ &\quad + S_{\mathrm{E}}\,(2\pi C_{\mathrm{T}})^2 \int_0^\infty [Z_{\mathrm{T}}(\omega)]^2 f^2 \,\mathrm{d}f \end{aligned} \tag{5.5}$$

Except for $Z_{\mathrm{T}}(\omega)$, which depends on the filter characteristic, we can find all the parameters in (5.5) from a knowledge of the preamplifier design, dealt with in the next chapter. In the following section, we shall examine $Z_{\mathrm{T}}(\omega)$ in greater detail. In particular, we will determine the frequency response of a digital receiver which results in the minimum output noise, while retaining an acceptable degree of ISI.

5.2 Digital receiver noise

In order to determine the integrals in (5.5) we redefine $Z_T(\omega)$ as

$$Z_T(\omega) = R_T H_T(\omega) \tag{5.6}$$

where R_T is the low-frequency transimpedance, and $H_T(\omega)$ represents the frequency dependence of $Z_T(\omega)$. If $H_p(\omega)$ is the Fourier Transform, *FT*, of the received pulse, $h_p(t)$, and $H_{out}(\omega)$ is the FT of the pulse at the output of the filter, $h_{out}(t)$, then we can express $Z_T(\omega)$ as

$$Z_T(\omega) = R_T H_T(\omega) = \frac{H_{out}(\omega)}{H_p(\omega)}$$

If we now normalise the output pulse shape, that is remove the dependency on R_T, we can write

$$H_T(\omega) = \frac{H_{out}(\omega)}{H_p(\omega)} \tag{5.7}$$

The FTs used in (5.7) depend on the bit-time of the pulses, T seconds. In order to remove this dependency, we use a normalised, dimensionless frequency variable, y, defined by

$$y = \frac{f}{B} = \frac{\omega}{2\pi B} = \frac{\omega T}{2\pi} \tag{5.8}$$

where B is the bit-rate. We can now define two new functions

$$H_p'(y) = \frac{1}{T} \times H_p(2\pi y/T) \qquad \text{and}$$

$$H_{out}'(y) = \frac{1}{T} \times H_{out}(2\pi y/T)$$

Thus the normalised receiver frequency response becomes

$$H_T'(y) = \frac{H_{out}'(y)}{H_p'(y)} \tag{5.9}$$

Because of the normalisation of $H_T(\omega)$, the integrals in (5.5) will only depend on the relative shapes of the input and output pulses. So, to return to (5.5), we can write

$$\begin{aligned} \langle n^2\rangle_T = {} & \left(S_I + \frac{S_E}{R_{in}^2}\right) R_T^2 B I_2 \\ & + (2\pi C_T)^2 S_E R_T^2 B^3 I_3 \end{aligned} \qquad (5.10)$$

where

$$I_2 = \int_0^\infty [H_T'(y)]^2 \, dy \qquad \text{and} \qquad I_3 = \int_0^\infty [H_T'(y)]^2 y^2 \, dy$$

(The inclusion of the B and B^3 terms in (5.10) accounts for the bandwidth (bit-rate) dependency of the noise. In fact, we can regard BI_2 and B^3I_3 as the noise equivalent bandwidths for the frequency-independent and f^2-dependent noise sources.) Since the signal output voltage and the r.m.s. output noise are both dependent on R_T, we can refer them to the input of the preamplifier. Thus the m.s. equivalent input noise current is given by

$$\langle i_n^2\rangle_c = \left(S_I + \frac{S_E}{R_{in}^2}\right) BI_2 + (2\pi C_T)^2 S_E B^3 I_3 \qquad (5.11)$$

If we know the required S/N, then it is a simple matter to determine the minimum signal current and hence the minimum optical power. This assumes that we know the value of the I_2 and I_3 integrals. As these depend on the shape of the input and output pulses, we must study them in greater detail. Before we consider the input pulse, let us define an output pulse shape that results in low noise and low ISI.

5.2.1 Raised-cosine spectrum pulses

At the output of the pre-detection filter, samples are taken to determine the polarity of the pulses. For minimum error rate, sampling must occur at the point of maximum signal. However, if ISI is present, then adjacent pulses will corrupt the sampled pulse amplitude, leading to an increase in detection errors. So, we require an output pulse shape that maximises the pulse amplitude at the sampling instant, and yet results in zero amplitude at all other sampling points, that is at multiples of $1/B$ where B is the data-rate.

A sin$x$/x pulse shape will satisfy the ISI requirement. Figure 5.3 shows a sequence of sin$x$/x pulses and, it should be evident that the amplitude of the precursors and tails due to adjacent pulses is zero at the pulse centres. So, the ISI is zero at the sampling instant. A further advantage of these pulses is that the pulse spectrum is identical to the frequency response of an ideal

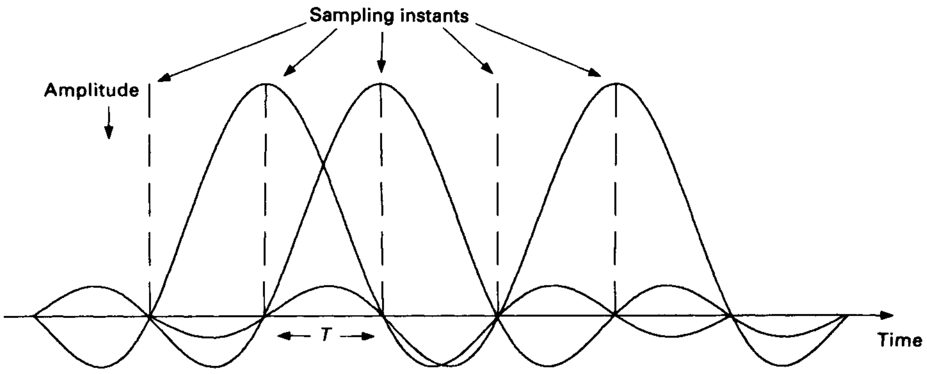


Figure 5.3 A sequence of sin*x*/*x* pulses

low-pass filter having a bandwidth of *B*/2. As this is the lowest possible bandwidth for a data-rate of *B*, the use of such an output pulse shape results in minimum receiver noise.

There are, however, several difficulties with such an output pulse shape:

(1) A receiver transfer function that results in sin*x*/*x* shape output pulses for a certain input pulse would be very intolerant of any changes in the input pulse shape. Even if the received pulse shape is fixed, variations in component values may cause the bandwidth of the pre-detection filter to reduce, leading to ISI at the sampling instant.

(2) It is important to sample at precisely the centre of the pulses, because ISI occurs either side. In practice, the rising edge of the clock varies either side of a mean, a phenomenon known as *clock jitter*, and this results in some ISI. (We can minimise jitter by careful design of the clock extraction circuit; however, some jitter is always present on the recovered clock.)

(3) A further disadvantage is that we are considering an ideal sin*x*/*x* pulse shape. In practice this is impossible to achieve.

From the foregoing, it should be evident that ISI is the major difficulty. The precursors and tails of the sin*x*/*x* pulses are due to the steep cut-off of the pulse spectrum. So, if we specify a pulse shape with a shallower cut-off spectrum, we can minimise the ISI either side of the sampling instant, so leading to more jitter tolerance. As we will see presently, we can only obtain this advantage at the cost of a reduction in S/N ratio.

Let us consider a pulse shape, $h_{\text{out}}(t)$, given by (5.12):

$$h_{\text{out}}(t) = \left(\frac{\sin \pi Bt}{\pi Bt}\right) \times \frac{\cos \pi rBt}{1 - (2rBt)^2} \tag{5.12}$$

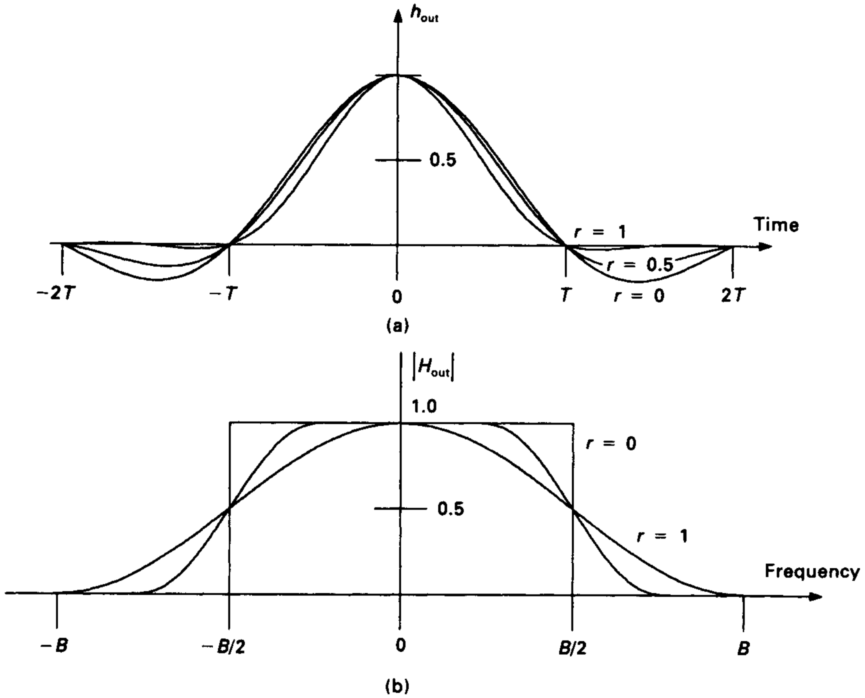


Figure 5.4 (a) A selection of raised-cosine spectrum pulses and (b) corresponding spectra

As can be seen, a factor that decreases rapidly with time has modified the sin*x*/*x* response of the ideal low-pass filter. Thus the precursors and tails are considerably reduced, leading to more jitter tolerance, and low ISI. The spectrum of these pulses, $H_{out}(f)$, is given by

$$
\begin{aligned}
H_{out}(f) &= 1 && |f| < (1-r)\frac{B}{2} \\
&= \frac{1}{2}\left[1 + \cos\{(\pi|f| - \pi f_1)/rB\}\right] && (1-r)\frac{B}{2} < |f| < (1+r)\frac{B}{2} \\
&= 0 \qquad \text{elsewhere} && (5.13)
\end{aligned}
$$

where f_1 is $(1 - r)B/2$, and r is known as the *spectrum roll-off factor*. Figure 5.4 shows the normalised pulse shapes and spectra for $r = 0$, 0.5 and 1. As can be seen, the spectra are similar to a cosine that has been shifted up by a d.c. level, and so these pulses are known as *raised-cosine spectrum* pulses. The value of roll-off factor affects both the ISI and the receiver noise: a large roll-off factor gives minimum ISI at the expense of bandwidth, and hence noise; the reverse is true for a low roll-off factor ($r = 0$ yields $\sin x/x$ pulses). In practice, ISI is the more important parameter, and so the output pulses of the preamplifier–filter combination have a full-raised cosine spectrum, that is $r = 1$. Hence the normalised spectrum of the output pulses is

$$\begin{aligned} H_{\text{out}}'(y) &= \frac{1}{2T} \times \{1 + \cos\pi y\} \qquad 0 < |y| < 1 \\ &= 0 \quad \text{elsewhere} \end{aligned} \tag{5.14}$$

Provided we know the input pulse shape, we can find the values of the I_2 and I_3 integrals using full-raised cosine spectrum output pulses.

5.2.2 Determination of I_2 and I_3

As we saw in section 2.4, the received pulse shape, $h_p(t)$, depends on the characteristics of the optical link: it may be rectangular, Gaussian, or exponential in form. To complicate matters further, the received pulses may occupy only part of the time-slot, that is short-width pulses. We have to account for all these factors, when calculating the values of I_2 and I_3. In a now classic paper, Personick [1] evaluated the integrals for all three different received pulse shapes, and interested readers are referred to the Bibliography for further details. Here we shall consider rectangular and, for comparison purposes only, Gaussian shape pulses. The normalised FTs of these pulses are

$$H_p'(y) = \frac{1}{2} \times \frac{\sin\alpha\pi y}{\alpha\pi y} \qquad \text{for rectangular pulses} \tag{5.15}$$

and

$$H_p'(y) = \frac{1}{T} \times \exp\{-(2\pi\beta y)^2/2\} \qquad \text{for Gaussian pulses} \tag{5.16}$$

where β is a measure of the pulse width. The parameter α in (5.15) is the fraction of the time-slot occupied by the rectangular pulses. If $\alpha = 1$, the pulses fill the whole of the slot, and we have *full-width* or *non-return-to zero*, *NRZ*, rectangular pulses. For Gaussian shaped pulses, the equivalent parameter is γ, defined by

$$\gamma = \int_{-T/2}^{T/2} h_p(t)\mathrm{d}t \tag{5.17}$$

These pulses are shown in figure 5.5.

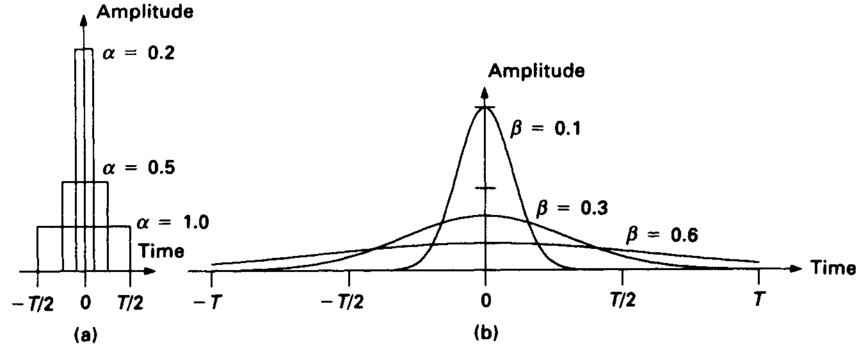


Figure 5.5 (a) Rectangular and (b) Gaussian shape pulses with various pulse widths

Table 5.1 Values of I_2, I_3 and γ for differing rectangular and Gaussian input pulse shapes

Rectangular input pulses

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
I_2	0.376	0.379	0.384	0.392	0.403	0.417	0.436	0.463	0.501	0.564
I_3	0.030	0.031	0.032	0.034	0.036	0.040	0.044	0.053	0.064	0.087

Gaussian input pulses

β	0.1	0.2	0.3	0.4	0.5	0.6
I_2	0.376	0.379	0.384	0.392	0.403	0.417
I_3	0.030	0.031	0.032	0.034	0.036	0.040
γ	1.000	0.988	0.904	0.789	0.683	0.595

By using these input pulse shapes, together with raised-cosine spectrum output pulses, we can find the I_2, I_3 and γ integrals by numerical integration. Table 5.1 summarises the results.

We should note that, for Gaussian input pulses, the values of I_2 and I_3 increase rapidly for $\beta > 0.5$. We should expect this because a high value of β results in considerable pulse spreading (note the values of γ). Thus ISI occurs *before* filtering, and the receiver must re-shape the pulses by emphasising the high-frequency components, resulting in an increase in noise.

We can determine the optimum receiver frequency response by dividing the output pulse shape (raised-cosine spectrum pulses) by the input pulse shape (rectangular or Gaussian in shape). For full-width rectangular input pulses, the optimum transfer function is approximated by a single-pole frequency response preamplifier, with a −3 dB cut-off at $B/2$ Hz, feeding a third-order Butterworth filter having a cut-off frequency of 0.7B Hz. (A wideband post-amplifier is usually inserted between the preamplifier and the filter.) In applications where the noise performance is not critical, for example short-haul links, the pre-detection filter is often omitted.

Provided the required S/N is known, we can calculate the receiver noise and hence the receiver sensitivity. In a digital receiver, we can predict the S/N from a knowledge of the decision-making circuitry and the binary signal probabilities. This is the subject of the next section.

5.2.3 Statistical decision theory

In a digital receiver, the output of the pre-detection filter consists of a sequence of raised-cosine spectrum pulses in the presence of additive preamplifier noise. The task of any processing circuitry is to determine, with the minimum uncertainty, whether a 1 or a 0 was received. As figure 5.6 shows, this is done by a *threshold crossing* device, or comparator, feeding a D-type flip-flop.

The circuit operation is best explained by examining the eye diagrams at certain relevant points. (An eye diagram is produced by observing the data stream on an oscilloscope which is triggered by the data clock. Because a complete cycle of the clock corresponds to one bit of data, the eye diagram will show the rising and falling edges of the data, as well as the logic 1 and logic 0 levels.) The eye at the input to the comparator clearly shows the slow rising and falling edges which are due to the limited receiver bandwidth. The effect of the preamplifier noise is to reduce the height and width

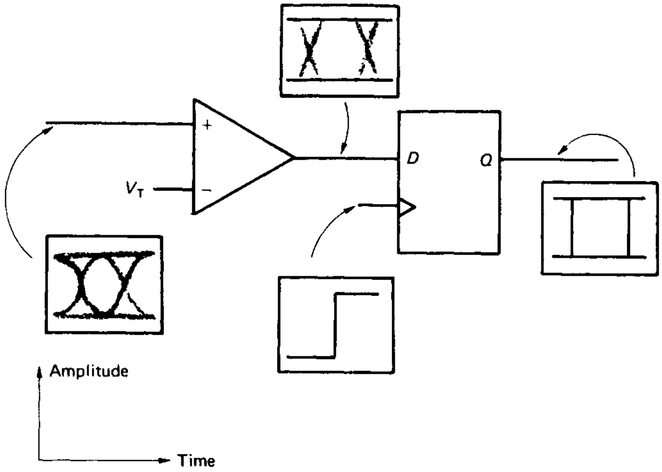


Figure 5.6 Eye diagrams and schematic diagram of a threshold crossing detector and central decision gate

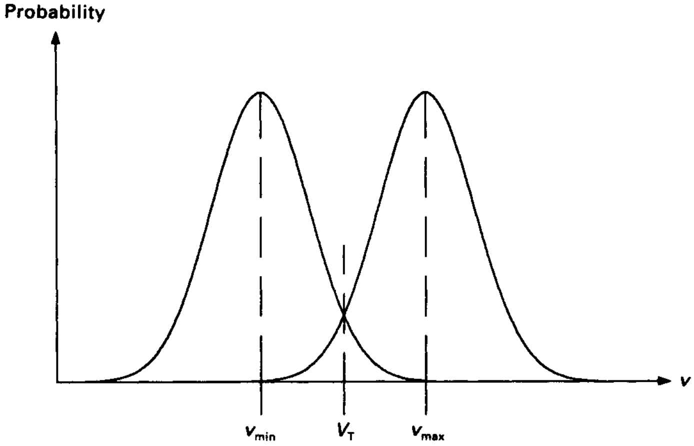


Figure 5.7 Probability density function plot for logic 1 and logic 0 pulses in the presence of additive Gaussian noise

of the eye, and so the comparator acts to 'clean up' the data. As we shall see presently, the optimum threshold level is mid-way between the logic 1 and 0 levels.

At the output of the comparator, all uncertainty about the level of the pulses has been removed; there is no observable noise at the centre of the eye. However the width of the eye is still affected by noise, and errors could result if sampling occurs close to the cross-over regions. Evidently the point of least uncertainty is the centre of the eye. Thus the clock to the *D*-type flip-flop is set to latch the data through to the output at the centre of the eye, so-called *central decision detection.* (The precise position of the clock rising edge can be set by using propagation delays through gates, and employing various lengths of co-axial cable.) The output of the *D*-type has no uncertainty associated with it and so a decision has been made, rightly or wrongly, about the received signal.

In order to evaluate the probability of a detection error for a certain S/N, we need to examine the noise-corrupted signal, at the input of the comparator, in greater detail. If we assume that 1s and 0s are equiprobable, then we can draw a probability density function plot of the data at the input to the comparator as shown in figure 5.7. In this figure, v_{max} and v_{min} represent the *received* signal levels at the output of the pre-detection filter, while V_T is the threshold voltage. So, any signal voltage above V_T is received as a logic 1, and any below V_T is a logic 0.

In this figure, the area under the logic 0 plot to the right of V_T represents the probability that a zero is received as a one, $P_{e0|1}$. Similarly, the area to the left of V_T is the probability that a logic 1 becomes a logic 0, $P_{e1|0}$. If the noise has a Gaussian distribution

$$P_{e0|1} = \frac{1}{\sqrt{2\pi\sigma_{off}{}^2}} \int_{V_T}^{\infty} \exp\{-(v - v_{min})^2/2\sigma_{off}{}^2\} \, dv \tag{5.18}$$

and

$$P_{e0|1} = \frac{1}{\sqrt{2\pi\sigma_{on}{}^2}} \int_{-\infty}^{V_T} \exp\{-(v_{max} - v)^2/2\sigma_{on}{}^2\} \, dv \tag{5.19}$$

where σ_{off} and σ_{on} are the r.m.s. noise voltage, at the comparator input, for logic 0 and logic 1 pulses. (The difference between the individual noise voltages accounts for signal-dependent shot noise. Although we are neglecting this noise source for the present, we include these terms for reasons of brevity.) As the probabilities of sending a logic 1 or logic 0 are identical and equal to 1/2, the total error probability, P_e, is

$$P_e = 0.5(P_{e0|1} + P_{e1|0}) \tag{5.20}$$

If we neglect signal-dependent shot noise, $\sigma_{off} = \sigma_{on} = \sigma$. In such circumstances, the optimum threshold voltage lies mid-way between v_{max} and v_{min}. The reason for this is that if we bias V_T to the left, $P_{e0|1}$ will increase at the expense of $P_{e1|0}$, whereas the opposite is true if we bias V_T slightly to the right. So, with these assumptions, $P_{e0|1} = P_{e1|0}$ and

$$\begin{aligned} P_e &= P_{e0|1} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{V_T}^{\infty} \exp\{-(v - v_{min})^2/2\sigma^2\} \, dv \end{aligned} \tag{5.21}$$

If we change variables by letting

$$x = \frac{v - v_{min}}{\sigma}$$

we get

$$P_e = \frac{1}{\sqrt{2\pi}} \times \int_{Q}^{\infty} \exp(-x^2/2) \, dx \tag{5.22}$$

where

$$Q = \frac{V_T - v_{min}}{\sigma} \tag{5.23}$$

Since V_T lies mid-way between v_{min} and v_{max}

$$V_T = \frac{v_{min} + v_{max}}{2}$$

and

$$Q = \frac{v_{max} - v_{min}}{2\sigma} \tag{5.24}$$

So, provided we know the signal voltage levels and the r.m.s. noise voltage at the input to the comparator, we can determine the error probability from (5.22). Although we can evaluate this by numerical methods, a more convenient solution is to express it in terms of the widely tabulated *complementary error function, erfc*, as

$$P_e = \frac{1}{2}\,\mathrm{erfc}(Q/\sqrt{2}) \tag{5.25}$$

where

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} \exp(-y^2)\,dy$$

For $Q > 2$, we can approximate (5.25) to

$$P_e = \frac{1}{Q\sqrt{2\pi}} \left(1 - \frac{0.7}{Q^2}\right) \exp(-Q^2/2) \tag{5.26}$$

So, a Q value of 6 results in an error-rate of 1 bit in 10^9, that is $P_e = 1 \times 10^{-9}$, and figure 5.8 shows the variation of P_e with Q.

Let us now consider the parameter Q in further detail. As defined by (5.24), all the parameters are directly dependent on the low-frequency transimpedance, R_T. So we can divide throughout by R_T to give

$$Q = \frac{I_{max} - I_{min}}{2\sqrt{\langle i_n^2 \rangle_c}}$$

or

$$\frac{I_{max} - I_{min}}{2} = Q\sqrt{\langle i_n^2 \rangle_c} \tag{5.27}$$

where $\langle i_n^2 \rangle_c$ is the m.s. equivalent input noise current, as defined by (5.11), and I_{max} and I_{min} are the maximum and minimum diode currents resulting

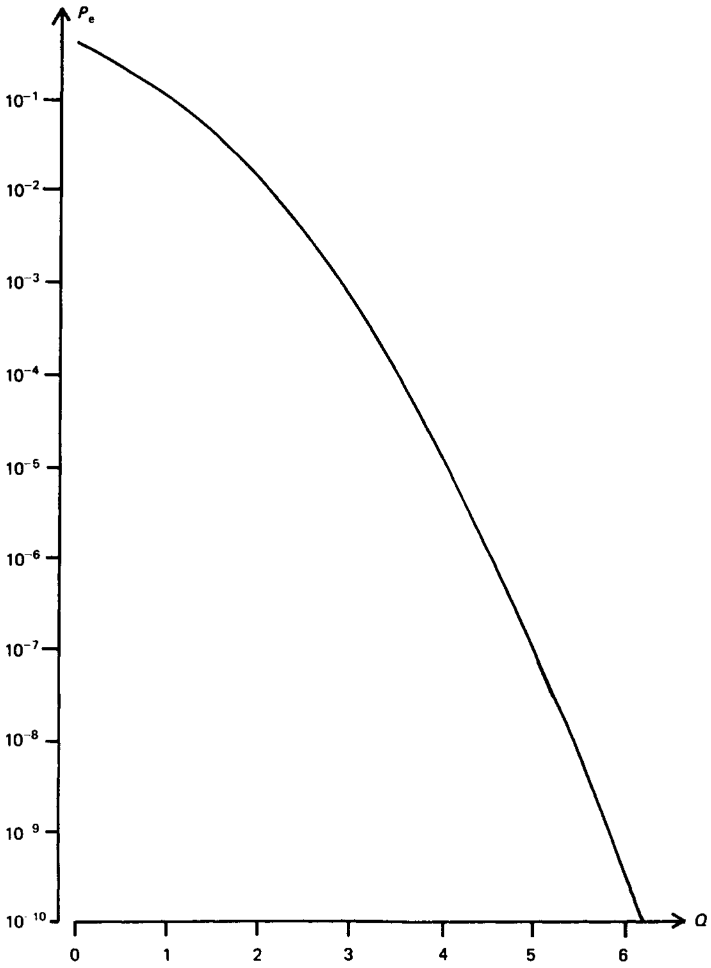


Figure 5.8 Graph of error probability, P_e, against signal-to-noise ratio parameter, Q, for a threshold crossing detector

from the different light levels. As figure 5.9 shows, I_{min} is not equal to zero. This results from imperfect extinction of the light source, so-called *non-zero extinction*. Under these conditions, the mean photodiode current is $I_{max}/2$, and so the receiver sensitivity, P, is

$$P = \frac{I_{max}}{2R_0} = \frac{I_{max} - I_{min}}{2R_0} \times \frac{I_{max}}{I_{max} - I_{min}}$$

$$= \frac{Q\sqrt{<i_n^2>_c}}{R_0} \frac{1}{1-\epsilon} \qquad (5.28)$$

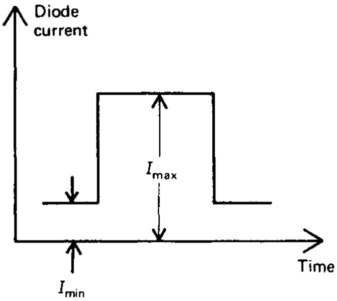


Figure 5.9 Illustrative of a non-zero extinction ratio

where $\epsilon = I_{min}/I_{max}$ is known as the *extinction ratio.* (We will return to this point later.) We should note, however, that the error-rate is dependent on the *difference* between the two light levels.

Most modern light sources have a very small extinction ratio, and so I_{min} is low in comparison with I_{max}. If we take I_{min} equal to zero, Q becomes the mean signal to r.m.s. noise ratio, and so

$$P = \frac{1}{R_0} Q\sqrt{<i_n^2>_c} \tag{5.29}$$

Hence, provided the signal-dependent noise is negligible, we can find the sensitivity from (5.29). In the next section, we will examine the effect of photodiode noise on receiver sensitivity.

Example

An optical receiver detects full-width rectangular pulses at a data-rate of 10 Mbit/s. The transfer function of the receiver and pre-detection filter is such that the output pulses have an ideal raised-cosine spectrum. The data consists of equiprobable 1s and 0s, and any signal-dependent noise is negligible when compared with the preamplifier noise. The preamplifier noise consists of a frequency-independent term of magnitude 5 × 10^{-24} A^2/Hz, and an f^2 noise term of magnitude 2 × 10^{-18} V^2/Hz^3. The input impedance of the preamplifier can be taken to be 10 kΩ, and the input capacitance is 3 pF. Determine the sensitivity of the receiver assuming 850 nm wavelength light, and a PIN photodiode with a quantum efficiency of 90 per cent.

The receiver detects 10 Mbit/s, full-width rectangular pulse data. So, the values of I_2 and I_3, from table 5.1, are

$$I_2 = 0.564$$

$$\text{and } I_3 = 0.087$$

We can now use (5.11) to give

$$\begin{aligned}
<i_n^2>_c &= \left(S_I + \frac{S_E}{R_{in}^2}\right)BI_2 + (2\pi C_T)^2 S_E B^3 I_3 \\
&= 2.82 \times 10^{-17} + 6.17 \times 10^{-20} \\
&= 2.82 \times 10^{-17}\ \text{A}^2
\end{aligned}$$

These figures are typical for a receiver operating under these conditions. We can see that the series noise generator has a negligible effect on the receiver noise. However, we should note that, because of the B^3 dependency, this noise source becomes dominant at high data-rates.

To find the required optical power, we need to find the responsivity of the detector. From (4.13) we find that

$$R_0 = 0.62\ \text{A/W}$$

and so, from (5.29), we find that we need 51.4 nW (−42.89 dBm) of optical power for an error rate of 1 bit in 10^9 ($Q = 6$).

5.2.4 Photodiode noise

As discussed in chapter 4, the photodiode noise falls into two main categories – invariant dark current noise, and signal-dependent shot noise. In a PIN receiver the signal-dependent noise is often insignificant compared with the circuit noise, but with an APD receiver we cannot ignore the signal noise. As we saw in chapter 4, it is common practice to approximate the APD signal current to a Gaussian random variable. Hence the sensitivity analysis we have just done will be valid for an APD receiver.

In order to simplify the following work, we will assume that the extinction ratio is zero, that is I_{min} *is zero*. So, by making use of the photodiode noise equations derived in chapter 4, the noise current spectral density of an APD can be written as

$$S_{Id} = 2qI_{DB}M^2F(M) + 2qI_{DS} \tag{5.30}$$

and

$$S_{IS} = 2q<I_s>M^2F(M) \tag{5.31}$$

where $\langle I_s \rangle$ is the average signal current. We can treat these noise sources in the same manner as the preamplifier shunt noise source. Thus the equivalent input noise current due to the photodiode is

$$\langle i_n^2 \rangle_{pd} = (S_{Id} + S_{Is})BI_2 \tag{5.32}$$

Now, $\langle I_s \rangle$ is dependent on the presence of an optical pulse in a particular time slot. When a pulse is present in the time slot, $\langle I_s \rangle$ is I_{max}, while $\langle I_s \rangle$ is zero when no pulse is present (assuming complete extinction of the source). So, the equivalent input noise currents for logic 1 and logic 0 pulses are

$$\langle i_n^2 \rangle_1 = 2qI_{max}M^2F(M)BI_2 + \langle i_n^2 \rangle_T \tag{5.33}$$

and

$$\langle i_n^2 \rangle_0 = \langle i_n^2 \rangle_T \tag{5.34}$$

where $\langle i_n^2 \rangle_T$ is the *total*, signal-independent, equivalent input noise current which includes the noise from the photodiode dark currents and any preamplifier noise. As the noise for logic 1 and logic 0 signals is different, the probability density plots of figure 5.7 will be different. If we account for this, and I_{min} is zero, Q will be given by

$$Q = \frac{I_{max}}{\sqrt{\langle i_n^2 \rangle_1} + \sqrt{\langle i_n^2 \rangle_0}} \tag{5.35}$$

Thus the mean optical power required is

$$P = \frac{Q}{2MR_0}\left(\sqrt{\langle i_n^2 \rangle_1} + \sqrt{\langle i_n^2 \rangle_0}\right) \tag{5.36}$$

If we substitute for $\langle i_n^2 \rangle_1$ and $\langle i_n^2 \rangle_0$, we get

$$\begin{aligned} P &= \frac{Q}{2MR_0}\left([2qI_{max}M^2F(M)BI_2 + \langle i_n^2 \rangle_T]^{\frac{1}{2}} + \langle i_n^2 \rangle_T^{\frac{1}{2}}\right) \\ &= \frac{Q}{2MR_0}\left([4qPR_0M^2F(M)BI_2 + \langle i_n^2 \rangle_T]^{\frac{1}{2}} + \langle i_n^2 \rangle_T^{\frac{1}{2}}\right) \end{aligned} \tag{5.37}$$

After some lengthy rearranging, we can express the sensitivity as

$$P = \frac{Q}{R_0}\left(\frac{\langle i_n^2 \rangle_T^{\frac{1}{2}}}{M} + qBI_2QF(M)\right) \tag{5.38}$$

Now, the first term in the brackets is inversely proportional to the avalanche gain, while the second term is, indirectly, dependent on M. Thus there must be an optimum value of M which minimises the required optical power. In order to find this optimum, we differentiate (5.38) with respect to M, equate the result to zero, and solve to find M_{opt}. (To simplify the derivation, we ignore the multiplied dark current. The effect of this approximation is not very dramatic.) Omitting the straightforward but lengthy mathematics, M_{opt} is given by

$$M_{\text{opt}} = \frac{1}{k^{\frac{1}{2}}} \left(\frac{\langle i_n^2 \rangle_T^{\frac{1}{2}}}{qBI_2Q} + k - 1 \right)^{\frac{1}{2}} \tag{5.39}$$

where k is the APD carrier ionisation ratio. In practice, because of all the approximations, (5.39) will only give an indication of the optimum gain. As we can vary the APD gain by altering the bias voltage, the optimum gain is often determined experimentally when the optical link is installed.

Example

In the optical receiver described in the previous example, the PIN detector is replaced by a silicon APD with the same quantum efficiency, a gain of 100, $k = 0.02$, and $I_{\text{DB}} = I_{\text{DS}} = 10$ nA. Determine the receiver sensitivity, the optimum APD gain, and the sensitivity at this gain.

In order to find the receiver sensitivity, we must use (5.38)

$$P = \frac{Q}{R_0} \left(\frac{\langle i_n^2 \rangle_T^{\frac{1}{2}}}{M} + qBI_2QF(M) \right)$$

Now, for an error rate of 1 in 10^9 pulses, $Q = 6$. The responsivity of the APD is the same as the PIN detector used previously, that is $R_0 = 0.62$ A/W. We also need to find the excess noise factor of the APD. By using (4.33) we find that $F(M) = 4$ for electrons producing avalanche gain. We also need to find the total signal-independent noise $\langle i_n^2 \rangle_T$.

If we assume that the source is totally extinguished, we can use (5.34) to give

$$\begin{aligned} \langle i_n^2 \rangle_T &= \langle i_n^2 \rangle_0 \\ &= (2qI_{\text{DB}}M^2F(M) + 2qI_{\text{DS}})BI_2 + \langle i_n^2 \rangle_c \\ &= 5.16 \times 10^{-16} + 2.02 \times 10^{-17} \\ &= 5.36 \times 10^{-16}\ \text{A}^2 \end{aligned}$$

From these figures we can see that the APD noise dominates over the preamplifier noise. This is due to the large amount of multiplied bulk leakage current.

So, the required optical power is

$$\begin{aligned} P &= \frac{6}{0.62}\left(2.3 \times 10^{-10} + 0.22 \times 10^{-10}\right) \\ &= \frac{6}{0.62} \times 2.52 \times 10^{-10} \\ &= 2.4 \text{ nW or } -56.13 \text{ dBm} \end{aligned}$$

To find the optimum avalanche gain we use (5.39) to give

$$\begin{aligned} M_{\text{opt}} &= \frac{1}{k^{\frac{1}{2}}}\left(\frac{\langle i_n^{2}\rangle_T^{\frac{1}{2}}}{qBI_2Q} + k - 1\right)^{\frac{1}{2}} \\ &= \frac{1}{\sqrt{0.02}}\left(1.2 \times 10^{3} - 0.98\right)^{\frac{1}{2}} \\ &= 244 \end{aligned}$$

The sensitivity of the receiver with this avalanche gain is

$$\begin{aligned} P &= \frac{6}{0.62}\left(0.95 \times 10^{-10} + 0.22 \times 10^{-10}\right) \\ &= 1.1 \text{ nW or } -59.46 \text{ dBm} \end{aligned}$$

From these calculations we can see that the use of an APD increases the sensitivity of the receiver. We can also see that, because of the small light levels, the signal-dependent APD noise is minimal when compared with the noise from the multiplied dark current. (In section 5.4 we will compare PIN and APD receivers in greater detail.)

5.2.5 Timing extraction

The sensitivity analysis just presented, assumed central decision detection. As we saw earlier, the *D*-type flip-flop requires a clock of period equal to the time-slot width. We could transmit this clock as a separate signal, but it is more usual to extract the clock from the received data. This is the function of the timing extraction circuit shown in figure 5.10.

The input to this circuit, taken from the threshold crossing detector, is

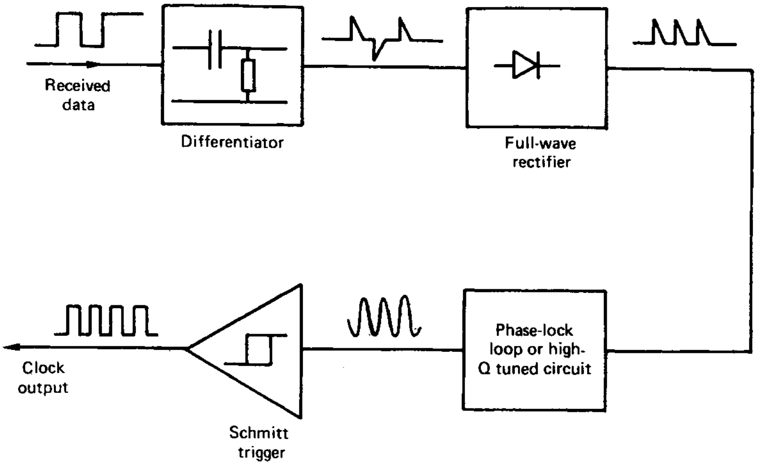


Figure 5.10 Schematic diagram of a timing extraction circuit

first differentiated and then full-wave rectified. These two operations result in a series of pulses with the same period as that of the required clock signal. It is then a simple matter to use a phase-lock-loop, *PLL*, or a high *Q* tuned circuit, to extract the clock required by the decision gate.

So long as there are a large number of data transitions, the circuit will maintain a clock to the flip-flop. However with the NRZ signalling format we are considering, a long sequence of 1s or 0s will cause a loss of the clock. This is because the PLL, or tuned circuit, will not receive any pulses. One solution to this problem is to use an alternative signalling format, such as *bi-phase* (or *Manchester*) coding. With this code, each time-slot contains a data transition regardless of the logic symbol (figure 5.11) and this increases the timing content. The major disadvantage of this code is that the pulse width is half that of full-width pulses, resulting in a doubling of the required bandwidth.

Although bi-phase coding is often used in low bit-rate links and local

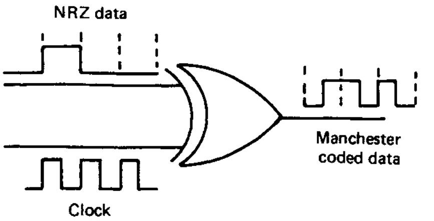


Figure 5.11 Generation of Manchester coded data using an exclusive-OR gate

area networks, *LANs*, the doubling in data-rate makes this format unattractive for use in high-speed telecommunications links. For this application, *block coding* of the NRZ data is used. With this type of coding, a look-up table converts *m* bits of input data into *n* bits of output data ($n > m$). Such codes are known as *mBnB* block codes, and they enable designers to increase the timing content of NRZ signals by limiting the maximum number of consecutive 1s or 0s.

Block coding of random data also helps to alleviate *base-line* wander, which causes the amplitude of a long sequence of like symbols to sag; in extreme circumstances, the amplitude of a long sequence of ones drops below the threshold level, and the error-rate increases. Base-line wander is due to a poor receiver low-frequency response filtering out the strong d.c. content of a long sequence of ones or zeros. For block-coded data to exhibit zero d.c. content, we must limit the maximum number of consecutive like symbols, and ensure that the number of coded ones and zeros is equal. Such codes are known as *zero disparity* block codes, and table 5.2 illustrates the 5B6B code.

Close examination of the table shows that, when coded, the maximum number of consecutive like symbols is six, and this aids timing extraction. We can alleviate base-line wander by using the *balanced disparity* code at the top of the table; each coded data word has an equal number of ones and zeros. However, the right-hand alphabet at the bottom of the table has a *positive disparity* of two (the number of ones exceeds the number of zeros by two) while the left-hand alphabet has an equal *negative disparity*. Every time a word is coded from the bottom of the table, the alphabet is changed to maintain zero mean disparity. As an example, if 00000 is encoded into 101000 by the left-hand alphabet, then the *running disparity* is minus two. The next time the bottom alphabet is selected, the transmitted word must come from the right-hand alphabet, which has a disparity of plus two. In this way, the coded data have zero mean disparity and hence zero d.c. content. If the lower cut-off frequency of the receiver is less than that of the coded data, we can eliminate base-line wander.

A useful feature of line coded data is that the spectrum has a lower cut-off frequency below which there are no signal components. Hence *supervisory channels* can use the empty low-frequency spectrum. Such channels are required for reporting on the state of various system components, and to send control data to repeaters and terminal equipment.

One disadvantage of block codes is that, because of their inbuilt redundancy, the encoded data-rate is *n*/*m* times the original data-rate. However, this increase is significantly less than that caused by bi-phase coding, and so high-speed links often use block codes.

Table 5.2 5B6B translation table

Input word	*Output word*	
00011	000111	
00101	001011	
00110	001101	
00111	001110	
01001	010011	
01010	010101	
01011	010110	
01100	011001	
01101	011010	
01110	011100	
10001	100011	
10010	100101	
10011	100110	
10100	101001	
10101	101010	
10110	101100	
11000	110001	
11001	110010	
11010	110100	
11100	111000	
00000	101000	010111
00001	011000	100111
00010	100100	011011
00100	010100	101011
01000	001100	110011
10000	100010	011101
01111	010010	101101
10111	001010	110101
11011	000110	111001
11101	010001	101110
11110	001001	110110
11111	000101	111010

5.3 Analogue receiver noise

Thus far we have only considered digital optical transmission links. However, some optical links transmit analogue information, for example composite video signals and analogue information from optical fibre sensors. Consequently, this section is concerned with analogue receiver noise. Although we use the term analogue receiver, the only difference between analogue and digital receivers is in the way the signals are processed after the post-amplifier. Depending on the modulation format, there may be some form of pre-detection filter prior to recovery of the baseband signal.

Let us consider sinusoidal amplitude modulation of the light, with a received optical power, $p(t)$, given by

$$p(t) = P_r (1 + ms(t)) \tag{5.40}$$

where P_r is the average received optical power, $s(t)$ is the modulating signal, and m is the modulation depth. For an APD, this signal produces a photodiode current, $i_s(t)$, given by

$$i_s(t) = R_0 M p(t) \tag{5.41}$$

and so the m.s. signal current, ignoring a constant d.c. term, is

$$<I_s^2> = \frac{1}{2} (R_0 M m P_r)^2 \tag{5.42}$$

From our discussions about digital receiver noise, we can write the equivalent input m.s. noise current as

$$\begin{aligned} <i_n^2>_T &= \int_0^{B_{eq}} 2qR_0P_rM^2F(M)\,df \\ &\quad + \int_0^{B_{eq}} 2qI_{DB}M^2F(M)\,df + <i_n^2>_c \end{aligned} \tag{5.43}$$

where B_{eq} is the noise equivalent bandwidth of the receiver, given by

$$B_{eq} = \int_0^\infty [H_T(\omega)]^2\,df \tag{5.44}$$

Performing the integrations in (5.43) yields

$$<i_n^2>_T = 2q(I_{DB} + R_0P_r)M^2F(M)B_{eq} + <i_n^2>_c \tag{5.45}$$

Now, the preamplifier noise current is given by

$$<i_n^2>_c = \int_0^{B_{eq}} \left(S_I + \frac{S_E}{R_{in}^2}\right) df + \int_0^{B_{eq}} S_E \times (\omega C_T)^2\,df$$

or

$$<i_n^2>_c = \left(S_I + \frac{S_E}{R_{in}^2}\right) B_{eq} + (2\pi C_T)^2 S_E \frac{B_{eq}^3}{3} \tag{5.46}$$

and so the signal-to-noise ratio is

$$\frac{\mathrm{S}}{\mathrm{N}} = \frac{<I_s^2>}{<i_n^2>_\mathrm{T}} = \frac{1}{2} \times \frac{(R_0 MmP_r)^2}{2q(I_\mathrm{DB} + R_0P_r)M^2F(M)B_\mathrm{eq} + <i_n^2>_\mathrm{c}} \tag{5.47}$$

As with the digital receiver, there is an optimum value of avalanche gain. We can find this value by differentiating (5.47) with respect to M, equating the result to zero, and solving for M. Thus we can find the optimum gain from

$$M_\mathrm{opt}^{2+x} = \frac{<i_n^2>_\mathrm{T}}{q(I_\mathrm{DB} + R_0P_r)B_\mathrm{eq}} \tag{5.48}$$

where we have made use of $F(M) = M^x$, and $<i_n^2>_\mathrm{T}$ is as defined previously. As with a digital receiver, the theoretical value of M_opt is only an indication of the optimum. When the receiver is commissioned, we can alter the APD bias to give the optimum S/N.

We have now completed our theoretical study of digital and analogue receivers. Before we go on to consider various preamplifier designs, we shall perform some sensitivity calculations, and describe a way of predicting the sensitivity from experimental data.

5.4 Comparison of APD and PIN receivers

In this section we will calculate the analogue and digital sensitivity of a receiver employing a PIN photodiode, and compare it with that of the same receiver using an APD. We will assume the receivers to have a bandwidth of 17 MHz, which allows for the detection of 34 Mbit/s digital data (corresponding to 512, 64 kbit/s PCM voice channels) with a NRZ format. We will consider two levels of preamplifier noise: 10^{-15} A^2, a somewhat noisy design; and 10^{-18} A^2, a typical state-of-the-art design.

The responsivity of both detectors will be taken as 0.5 A/W at 850 nm. We will assume that the noise from the PIN leakage current is negligible in comparison with other noise sources, while the APD surface and bulk leakage currents will be taken to be identical and equal to 10 nA. For the APD, we take a multiplication factor of 100 and an excess noise factor of 4. In order to simplify the work, we will assume that the source is completely extinguished.

If we initially consider the noisy preamplifier with a PIN photodiode, then use of (5.29) results in a sensitivity, for a 10^{-9} error-rate, of

$$\begin{aligned} P &= \frac{6}{0.5}\sqrt{10^{-15}}\ \mathrm{W} \\ &= 380\ \mathrm{nW} \qquad \text{or } -34.21\ \mathrm{dBm} \end{aligned}$$

Table 5.3 Comparison of digital and analogue receiver performance using PIN and APD detectors. The terms in brackets are the avalanche gain of the APD. The second set of figures in the APD columns relate to the optimum avalanche gain

Detector	*PIN*		*APD*	
Preamplifier noise level (A^2)	10^{-15}	10^{-18}	10^{-15}	10^{-18}
Digital receiver sensitivity (dBm)	–34.21	–49.21	–51.00 (100) –51.09 (150)	–51.65 (100) –52.16 (50)
Analogue receiver S/N (dB)	13.00	37.78	33.43 (100) 34.36 (32)	33.48 (100) 38.27 (1.6)

If we replace the PIN detector by the APD, we must first calculate the total signal-independent shot noise. Thus

$$\begin{aligned} <i_n^2>_T &= <i_n^2>_c + 2qI_{DS}I_2B + 2qI_{DB}M^2F(M)I_2B \\ &= 10^{-15} + (1 + 4 \times 10^4)\, 6.14 \times 10^{-20} \\ &= 3.45 \times 10^{-15} \text{ A}^2 \end{aligned}$$

As can be seen, the surface leakage current shot noise is insignificant when compared with the bulk leakage current shot noise. By substituting $<i_n^2>_T$ into (5.38) we get

$$\begin{aligned} P &= \frac{6}{0.5}\left(\frac{(3.45 \times 10^{-15})^{\frac{1}{2}}}{100} + 7.4 \times 10^{-11}\right) \\ &= 8 \text{ nW} \qquad \text{or } -51.00 \text{ dBm} \end{aligned}$$

Thus an APD will significantly increase the receiver sensitivity. However, table 5.3 shows that the advantage is reduced if we use the lower noise preamplifier. Before we comment further on these results, let us compare the performance of the receivers when detecting analogue signals.

In the following calculations, we will take an average received power, P_r, of −33 dBm (or 500 nW), and a modulation depth of 0.8. The first step in the calculation of S/N is to find the noise equivalent bandwidth of the preamplifier.

If the receivers have a single-pole frequency response, then B_{eq} is

$$
\begin{aligned}
B_{\mathrm{eq}} &= \int_0^\infty \left(\frac{1}{1 + \mathrm{j}f/f_0}\right)^2 \mathrm{d}f \\
&= \int_0^\infty \frac{1}{1 + (f/f_0)^2}\, \mathrm{d}f \\
&= \frac{\pi}{2} \times f_0
\end{aligned}
$$

where f_0 is the -3 dB frequency of the receiver. (For an ideal low-pass filter, B_{eq} is simply f_0.) Thus, for the receivers considered,

$$B_{\mathrm{eq}} = 17\frac{\pi}{2} \approx 27 \text{ MHz}$$

By using this result, together with the parameters previously quoted, we can find the receiver sensitivity from (5.47). Table 5.3 summarises the results.

Examination of table 5.3 shows that the use of an APD with a noisy digital receiver results in a significant increase in sensitivity, compared with that obtained with a PIN detector. However, if the receiver noise is low, the advantage is considerably reduced. This is because the APD noise dominates the total receiver noise, and so any reduction in preamplifier noise will not produce a significant change in sensitivity. With a PIN detector, however, the preamplifier noise is dominant, and so a reduction in preamplifier noise causes a large change in sensitivity.

With noisy analogue receivers, an APD detector is preferable to a PIN. However, with the low-noise analogue receiver, the use of an APD is a disadvantage. We should expect this because the average received power produces a standing photocurrent which, in an APD receiver, results in a high level of multiplied shot noise. In general, we can conclude that the use of an APD will increase the sensitivity of a *noisy* preamplifier.

In the next section we will consider ways of measuring receiver sensitivity. Also presented is a method of determining the sensitivity from a knowledge of the output noise characteristic, and the receiver transfer function.

5.5 Measurement and prediction of receiver sensitivity

5.5.1 Measurement of receiver sensitivity

The sensitivity of an optical receiver (that is, the preamplifier and associated signal processing circuitry) detecting digital data can be measured directly with an error-rate test set. This equipment comprises a pseudo-random binary sequence, *PRBS*, generator, and an error-rate detector. The PRBS generator modulates a light source, the output of which is coupled to the receiver photodiode. The output of the receiver *D*-type flip-flop is then applied to the error detector. This instrument compares the detected PRBS with the transmitted sequence and counts the number of errors in a certain time interval. It is then a simple matter to find the probability of an error, P_e.

We can find the mean optical power resulting in the measured error-rate by monitoring the photodiode current, and then dividing by the responsivity. Attenuators placed in the optical path will vary the received power, and hence the number of errors. If a graph of P_e against optical power is then plotted, the required power for a specified error-rate can be easily found. (This graph will take the form of figure 5.8.)

As previously noted, the low-level light signal is unlikely to be zero, and so the ammeter monitoring the photodiode current will read $I_{max}/2$ and not $(I_{max} - I_{min})/2$. We can convert the ammeter reading into the current we require by using the following formula:

$$\frac{I_{max} - I_{min}}{2} = \frac{I_{max}}{2}(1 - \epsilon) \tag{5.49}$$

where ϵ is the extinction ratio.

So, the use of an error-rate test set allows us to measure the sensitivity of an optical receiver directly. Unfortunately error-rate equipment can be expensive, particularly if the data-rate is high (>140 Mbit/s). Most laboratories have a spectrum analyser, and the next section presents a method for predicting the sensitivity from measurements taken with this instrument.

5.5.2 Prediction of receiver sensitivity

The theoretical prediction of receiver sensitivity relies on calculating the noise spectral density at the input of the receiver. We can make use of this information to predict the sensitivity from a knowledge of the preamplifier output noise characteristic.

Most modern spectrum analysers have a facility for measuring noise spectral density. We can find the output noise voltage spectral density of a preamplifier by boosting the output noise using a cascade of wideband amplifiers. If we divide this noise characteristic by the total transimpedance, we get the equivalent input noise current spectral density. It is then a simple matter to perform a curve fitting routine to determine the values of the frequency invariant, and the f^2 variant noise current spectral density components. We can use these parameters, in place of the analytical coefficients in (5.11), to determine the equivalent input noise current, and hence the receiver sensitivity.

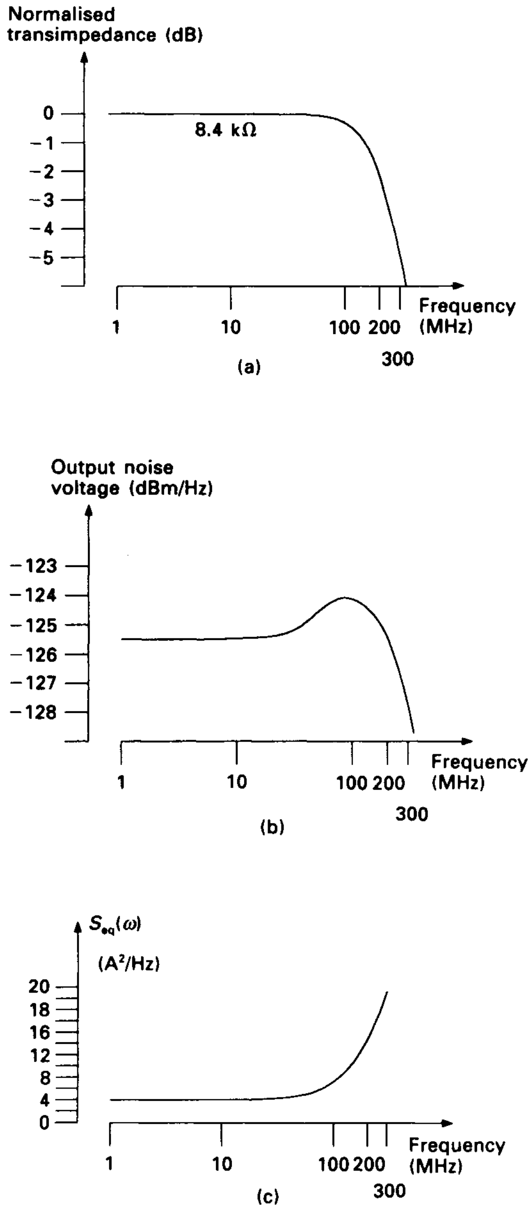


Figure 5.12 (a) Transimpedance relative to mid-band; (b) output noise spectral density; (c) equivalent input noise current spectral density of a PINBJT transimpedance preamplifier. All graphs are based on experimental data

As an example, let us consider the output noise voltage spectral density shown in figure 5.12a. When divided by the transimpedance, figure 5.12b, the equivalent input noise current spectral density takes the form of figure 5.12c. From this figure, the frequency invariant coefficient is 4 × 10^{-24} A^2/Hz, while the f^2 variant term is approximately 2.1 × 10^{-40} A^2/Hz^3. Hence we can predict the sensitivity of this receiver from

$$P = \frac{Q}{R_0}\sqrt{4 \times 10^{-24} I_2 B + 2.1 \times 10^{-40} I_3 B^3} \qquad (5.50)$$

This method relies on an accurate knowledge of the receiver transimpedance. The low-frequency transimpedance can be obtained by injecting a constant current into the preamplifier. We also need to know the frequency response of the receiver. We can obtain this by connecting a sweep generator to the constant-current source, and monitoring the output on the spectrum analyser. An alternative method is to modulate a light source with the output from the sweep generator, and connect a spectrum analyser to the preamplifier output as before.

Apart from experimental errors, the major source of error results from the use of a theoretically ideal pre-detection filter. In spite of this, the predicted sensitivity can be within 1 dB of the actual receiver sensitivity.

For further background reading, see references [1] and [2].

6 Preamplifier Design

In the previous chapter we assumed that the detector–preamplifier combination, which we shall now call the receiver, had a bandwidth of at least 0.5 times the bit-rate, or the baseband bandwidth for analogue signals, and low-noise. In this chapter we will consider the design and analysis of various preamplifier circuits, with the aim of optimising these characteristics. We shall consider the two most common types of preamplifier – the *high input impedance* design and the *transimpedance* design. In the noise analyses presented, we will only consider the performance of preamplifiers receiving digital signals.

The high input impedance preamplifier is the most sensitive design currently available and, as such, finds applications in long-wavelength, long-haul routes. The high sensitivity is due to the use of a high input resistance preamplifier (typically >1 MΩ) which results in exceptionally low thermal noise. The high resistance, in combination with the receiver input capacitance, results in a very low bandwidth, typically <30 kHz, and this causes integration of the received signal; indeed, these receivers are commonly called *integrating front-end* designs. A differentiating, *equalising* or *compensating*, network at the receiver output corrects for this integration.

In contrast, the transimpedance design relies on negative feedback to increase the bandwidth of the open-loop preamplifier, and so a compensation circuit is not normally required. Although the resulting receiver is often not as sensitive as the integrating front-end design, this type of preamplifier does exhibit a high dynamic range and is usually cheaper to produce.

Both types of preamplifier can use either field effect transistors, *FETs*, or bipolar junction transistors, *BJTs*, as the input device. FET input receivers are usually more sensitive than BJT input receivers; however, as we shall see later, the situation can change at high data-rates (typically greater than 1 Gbit/s). When we examine the integrating front-end receiver, we will consider a FET input design whereas, when we consider a transimpedance receiver, we will use a BJT. These are the most common configurations in current use.

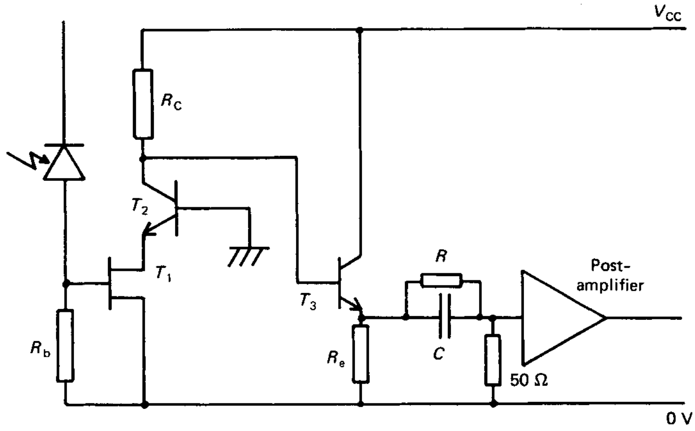


Figure 6.1 A basic PINFET optical receiver with equalisation network and post-amplifier

6.1 High input impedance preamplifiers

As these designs rely on a very high input resistance to produce a sensitive receiver, the choice of front-end transistor is important. BJTs have a relatively low input resistance and so are seldom used. On the other hand, FETs exhibit a very large input resistance, and so these are the obvious choice for the front-end device. Integrating front-end preamplifiers usually consist of a PIN photodiode feeding a FET input preamplifier. The resulting circuit is commonly known as a *PINFET* receiver, and a typical design is shown in simplified form in figure 6.1. (For reasons of clarity, we have not included the biasing components.)

The front-end is a common-source, *CS*, stage feeding a common-base, *CB*, stage (as shown in figure 6.1). This configuration, known as a *cascode* amplifier, results in a low input capacitance and a high voltage gain. (It is not necessary to use BJTs at all; some designs use FETs throughout, and can be fabricated as gallium arsenide integrated circuits.)

Also shown in figure 6.1 is the compensation network which has a zero at the same frequency as the front-end pole, and a pole which determines the receiver bandwidth. The 50 Ω load resistor across the output of the compensation network represents the input resistance of any following amplifier.

We shall now examine the frequency response (taking account of the compensating network), noise characteristic and dynamic range of a PINFET receiver. Although we only consider FET input designs, a similar analysis would also apply to PINBJT receivers.

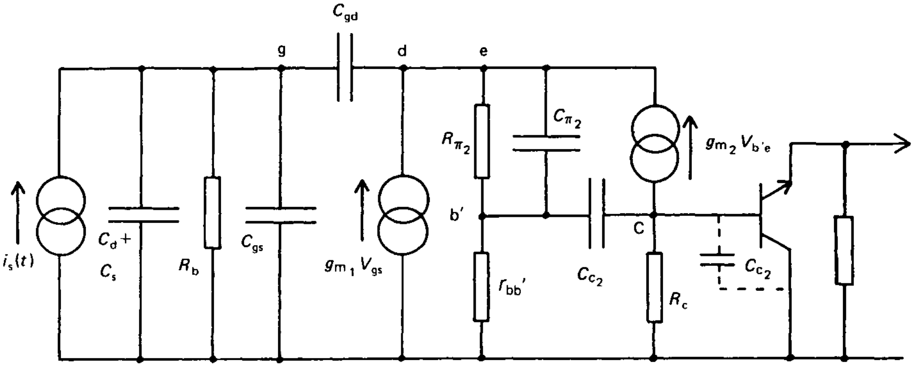


Figure 6.2 A.c. equivalent circuit of a PINFET receiver

6.1.1 Frequency response

The time constants associated with the front-end and the cascode load determine the frequency response of the PINFET receiver shown in figure 6.1. The first time constant, τ_{in}, causes a pole which is usually located at about 30 kHz, and it is this that the compensating network must counteract.

We can determine the relevant time constants by drawing the a.c. equivalent circuit, and then finding the resistance in parallel with each capacitance. Thus, with reference to figure 6.2, the front-end pole, s_1, is

$$s_1 = \frac{1}{\tau_{in}} \tag{6.1}$$

where

$$\tau_{in} = R_b C_{in} \tag{6.2}$$

Here C_{in} is the total input capacitance, which is the sum of the diode capacitance, C_d; the FET gate–source capacitance, C_{gs}; the stray input capacitance, C_s; and the *Miller* capacitance, $(1 - A_1)C_{gd}$. The parameter C_{gd} is the FET gate–drain capacitance, and A_1 is the voltage gain of the FET stage, given by

$$A_1 = -g_{m1} r_{e2}$$

or

$$A_1 = -g_{m1} \frac{V_T}{I_{e2}} \tag{6.3}$$

where $V_T = 25\,\text{mV}$ and g_{m1} is the FET *transconductance*. Thus C_{in} will be given by

$$C_{in} = C_d + C_{gs} + C_s + \frac{(1 + g_{m1}V_T)C_{gd}}{I_{e2}} \tag{6.4}$$

Now, although the FET is usually biased at about 15 mA, g_{m1} is typically 15 mS (considerably lower than that achievable by a BJT operating at the same current). This is because, unlike a BJT, the g_m of a FET is relatively independent of bias current. This low g_m, together with the load resistance of r_{e2}, means that A_1 will be very low. However, the gain of the common-base stage, A_2, is $g_{m2}R_c$, where g_{m2} is the transconductance of the CB stage, and so the total voltage gain, A_0, may well be high ($A_0 = A_1A_2$).

It is important to minimise C_{in} because this will allow for the use of a larger value of R_b, for the same pole location, and hence a reduction in thermal noise. As we shall see in the next section, a small input capacitance will also reduce the preamplifier noise. Most PINFET receivers use unpackaged devices in a hybrid thick-film construction, which results in a very low input capacitance, and hence low noise. However, this technique usually involves higher production costs.

To equalise for the front-end integration, the compensating network should have a zero at the same frequency as the input pole. We can find the location of the compensating zero by noting that the transfer function, $H_{eq}(\omega)$, of the compensation network is

$$H_{eq}(\omega) = \left[\frac{50}{50 + R}\right]\left[\frac{1 + j\omega CR}{1 + j\omega CR50/(50 + R)}\right] \tag{6.5}$$

and so CR should equal τ_{in}. The value of the pole in the denominator of (6.5) sets the bandwidth of the receiver and, if the value of R is high enough, we can approximate it to $1/C50$.

At frequencies below s_1, the compensation network acts as a potential divider and so the transimpedance of the receiver can be quite low (< 10 kΩ). This means that, when referred to the receiver input, any noise from the following amplifier can be quite high, and this may reduce the sensitivity. We can regain the sensitivity by increasing the transimpedance, and this is most easily done by increasing the preamplifier voltage gain. In view of this, most practical PINFET receivers employ additional, low-noise amplification stages prior to compensation.

If we compensate for the front-end pole, the next major pole, s_2, is that associated with the cascode load time constant, τ_c. If s_2 is lower in frequency than the compensation network pole, then the receiver will fail; thus it is important to determine τ_c, and hence s_2. By referring to figure 6.2, we can see that τ_c approximates to

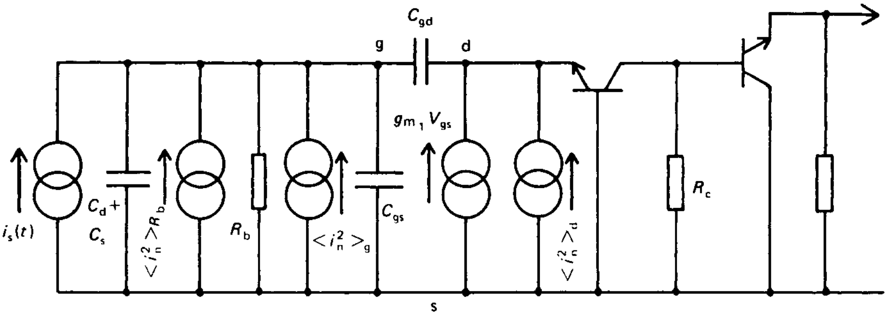


Figure 6.3 Noise equivalent circuit of a PINFET receiver

$$\tau_c = R_c 2C_c \text{ and so} \tag{6.6}$$

$$s_2 = \frac{1}{2R_cC_c} \tag{6.7}$$

where C_c is the BJT collector–base capacitance. This pole location is an approximation because we are neglecting the base-spreading resistance, r_{bb}', and the loading effect of the output emitter follower.

6.1.2 Noise analysis

In this preamplifier there are three sources of noise: thermal noise from R_b; thermal noise due to the channel conductance; and shot noise from the gate leakage current, I_g. It may be recalled from the previous chapter, that we need to find the total equivalent input noise current in order to determine the receiver sensitivity.

We can see from figure 6.3 that the I_g generator and the R_b generator are both connected to the input node, and so are easily dealt with. However, we must refer the channel conductance thermal noise to the input by some means. To do this, we note that the $<i_n^2>_d$ generator produces a noise current spectral density of $4kT\Gamma g_m$ A^2/Hz, in a short circuit placed across the drain and source. (The parameter Γ is known as the FET constant. It has an approximate value of 0.7 for Si and 1.1 for GaAs FETs.) We can refer this current to the input of the FET by dividing by the transconductance. So, a gate–source noise voltage generator of spectral density $4kT\Gamma/g_m$ V^2/Hz will produce an m.s. short-circuit output current equal to the channel conductance noise current. As this generator drives the input admittance of the short-circuited transistor, we can write the *total* equivalent input m.s. noise current, $<i_n^2>_c$, as

$$<i_n^2>_c = \frac{4kTI_2B}{R_b} + 2qI_gI_2B + \frac{4kT\Gamma}{g_{m1}}\left[\frac{I_2B}{R_b^2} + (2\pi C_T)^2 I_3 B^3\right] \tag{6.8}$$

where C_T is given by

$$C_T = C_d + C_{gs} + C_{gd} + C_s \tag{6.9}$$

We should note that, as a result of short-circuiting the drain and source, the Miller capacitance does not appear in (6.9). We can simplify (6.8) if we assume that R_b is very large, and the gate leakage current is very low. Thus the receiver noise becomes

$$<i_n{}^2>_{min} = \frac{4kT\Gamma}{g_{m1}}(2\pi C_T)^2 I_3 B^3 \tag{6.10}$$

which represents the minimum amount of noise available from an ideal PINFET receiver. This clearly shows the need to minimise the input capacitance and use high g_m FETs.

6.1.3 Dynamic range

The integration of the received signal at the front-end restricts the dynamic range of PINFET receivers – a long sequence of 1s will cause the gate voltage, V_g, to ramp up and this may disrupt the biasing levels, so causing the receiver to fail. In digital receivers, line coded data can correct for this integration. For example, if we consider the 5B6B code discussed in section 5.2.5, the maximum number of consecutive ones will be six, and this will cause the gate voltage, V_g, to rise to a certain level. However, six zeros will eventually follow the six ones, to maintain a zero symbol disparity, and so V_g will ramp down again.

Unfortunately line coding cannot take account of variations in input power level, which will also affect analogue receivers. The solution is to use an automatic gain control, or *agc*, circuit which prevents the receiver from saturating (that is, it keeps the bias conditions constant). PINFET receivers with this facility have dynamic ranges in excess of 20 dB.

6.1.4 Design example

A PINFET receiver is to operate at a data-rate of 140 Mbit/s. The input stage of the design is a cascode arrangement, with a bias current of 15 mA. The g_m of the GaAs FET at this bias current is 15 mS, and the gate leakage current is 15 nA. The receiver is to have a total voltage gain of 100. Microwave bipolar transistors, with a collector–base capacitance of 0.3 pF, are used after the FET input. The design is fabricated on a hybrid thick-film circuit, resulting in a total input capacitance of 0.5 pF. Complete the design of the receiver, and estimate the receiver sensitivity assuming an ideal predetection filter, and an error rate of 1 in 10^9 bits. Take a responsivity of unity.

As this is a PINFET design, the bias resistor on the gate of the FET front-end needs to be as high as possible. The highest practical resistance is 10 MΩ which results in a front-end pole at 32 kHz. The equalising network must compensate for this pole. From equation (6.5) we find

$$\frac{1}{2\pi RC} = 32 \times 10^3$$

and so

$$RC = 5\ \mu\text{s}$$

We require a minimum equalised bandwidth of 70 MHz, and so the pole in (6.5) must be at this frequency. Thus

$$\frac{50 + \mathrm{R}}{2\pi CR50} = 70 \times 10^6$$

and so

$$R = 110\,\text{k}\Omega \text{ and } C = 45\,\text{pF}$$

As it is unlikely that components of the exact value will be available, it is normal practice to make the resistor, or capacitor, variable.

The design of the equaliser assumes that the receiver has a very wide bandwidth when the front-end pole is equalised. Thus we need to examine the equalised receiver bandwidth. This bandwidth will be set by the time constant associated with the cascode load and the input capacitance of the following stage. Now, the common-source front-end sees the input resistance of the common-base stage as a load. Taking a cascode bias current of 15 mA, and a FET transconductance of 15 mS, we can use equation (6.3) to give

$$\begin{aligned} A_1 &= g_{m1}\frac{V_T}{I_{e2}} \\ &= 15 \times 10^{-3} \times \frac{25 \times 10^{-3}}{15 \times 10^{-3}} \\ &= 25 \times 10^{-3} \end{aligned}$$

As the total voltage gain should be 100, the CB stage needs a gain of 4×10^3 resulting in a load resistor of 6.7 kΩ. Such a high value of R_c may cause difficulties with biasing conditions and limit the compensated bandwidth. Thus we will take R_c equal to 400 Ω. This results in A_0 being 6, and so we must use further amplification stages, with a combined gain of 17.

The microwave BJT transistors used on the design have a C_c of 0.3 pF. Thus we can use equation (6.6) to give a compensated bandwidth of

$$\begin{aligned} f_2 &= \frac{1}{2\pi\tau_c} \\ &= \frac{1}{2\pi \times 400 \times 2 \times 0.3 \times 10^{-12}} \\ &= 663 \text{ MHz} \end{aligned}$$

This assumes that the stage following the cascode does not introduce any loading effects. As we saw earlier, a high-voltage gain is important, and so a common-emitter amplifier could follow the cascode stage. However, this will tend to increase the capacitance seen by the cascode load, and so limit the equalised bandwidth. Thus the cascode should be followed by an emitter follower, which can then feed further common-emitter gain stages.

We now need to examine the noise performance of this design. If we ignore the photodiode leakage current, then the total equivalent input m.s. noise current will be given by (6.8). Thus

$$\begin{aligned} \langle i_n^2 \rangle_c &= \frac{4kTI_2B}{R_b} + 2qI_gI_2B + \frac{4kT\Gamma}{g_{m1}} \left\{ \frac{I_2B}{R_b^2} + (2\pi C_T)^2 I_3 B^3 \right\} \\ &= 1.3 \times 10^{-19} + 3.8 \times 10^{-19} + 1.2 \times 10^{-18}(7.9 \times 10^{-7} + 2.4) \\ &= 3.4 \times 10^{-18} \text{ A}^2 \end{aligned}$$

where we have taken C_T to be 0.5 pF and $\Gamma = 1.1$. This results in a sensitivity of −49.56 dBm, for R_0 equal to 1 and an error rate of 1 in 10^9. As we can see from these figures, the noise from the bias resistor and the gate leakage current is not very significant when compared with the channel noise. Indeed, as the bit-rate increases, the channel noise becomes even more dominant, and so equation (6.10) will accurately predict $\langle i_n^2 \rangle_c$.

6.2 Transimpedance preamplifiers

An ideal *transimpedance* amplifier supplies an *output voltage* which is directly proportional to the *input current*, and independent of the source and

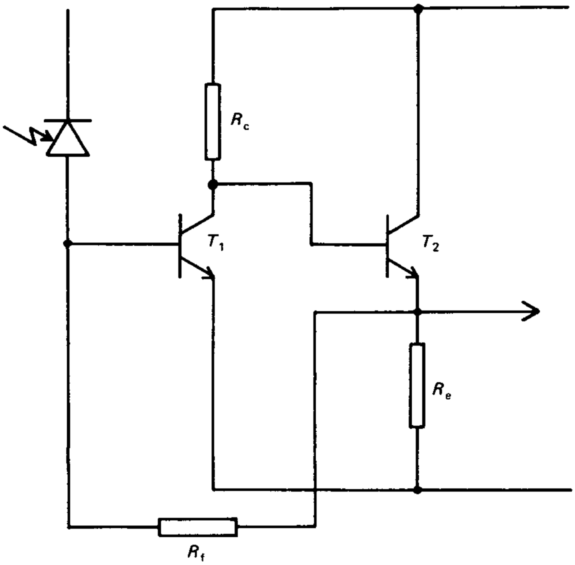


Figure 6.4 A simple common-emitter/common-collector, shunt feedback transimpedance receiver

load impedance. We can closely approximate the ideal amplifier by using negative feedback techniques to reduce the input impedance. If the open-loop amplifier is ideal, it has infinite input and zero output resistance, then the transfer function of the feedback amplifier equals the impedance of the feedback network. As well as a predictable transfer function, a transimpedance preamplifier also exhibits a large closed-loop bandwidth and so, in general, integration of the detected signal does not occur. Apart from the obvious advantage of not requiring a compensation network, the high bandwidth also results in a dynamic range which is usually larger than that of a PINFET receiver.

The choice of front-end transistor is entirely at the discretion of the designer; however, as we considered a FET input preamplifier previously, we will only examine BJT input transimpedance designs.

Figure 6.4 shows the circuit diagram of a simple, common-emitter, *CE*, common-collector, *CC*, shunt feedback preamplifier. Comparison with the PINFET receiver reveals that a feedback resistor, R_f, replaces the bias resistor and, by virtue of Miller's theorem, this resistance appears at the input as $R_f/1 - A_0$. Thus even if R_f is high, in order to reduce thermal noise, the input resistance will be less than that of the PINFET, and this will result in a higher bandwidth. (A_0 is the *open-loop* voltage gain which we can find by breaking the feedback loop, loading the circuit with R_f at both ends of the loop, and then calculating the voltage gain. Fortunately, the open-loop voltage gain is almost the same as the closed-loop voltage gain, and we shall use this approximation.)

In this design, A_0 is the product of the front-end gain and the second-stage attenuation (which we shall assume to be negligible). As we shall see later, A_0 should be as high as possible to achieve a large bandwidth; however, a high value of voltage gain may cause instability, and so most transimpedance designs have voltage gains of less than 100.

We will now proceed to examine the frequency response and noise performance of this design. Although we will consider a CE/CC shunt feedback preamplifier, the same methods can be applied to other designs.

6.2.1 Frequency response

We can find the transfer function of a transimpedance preamplifier by applying standard feedback analysis using *impedances* rather than voltages. Thus, we can relate the closed-loop transfer function, $Z_c(s)$, to the open-loop transfer function, $Z_o(s)$, and the feedback network transfer function, $Z_f(s)$, by

$$\frac{1}{Z_c(s)} = \frac{1}{Z_o(s)} - \frac{1}{Z_f(s)} \tag{6.11}$$

The open-loop transimpedance is given by

$$Z_o(s) = A_0(s) \times \frac{R_{in}R_f}{R_{in} + R_f}$$

where $A_0(s)$ signifies that A_0 is frequency dependent. In order to simplify the mathematics, we shall assume that the input resistance, R_{in}, is high. Therefore

$$Z_o(s) = A_0(s)R_f \tag{6.12}$$

From our discussion of the PINFET cascode receiver, it should be apparent that $A_0(s)$ has two *open-loop* poles: one associated with the input time constant, τ_{in}; and one due to the time constant of the CE stage load, τ_c. If we assume that $\tau_{in} >> \tau_c$ (that is, the front-end pole is dominant) we can write $A_0(s)$ as

$$A_0(s) = \frac{A_0}{(1 + s\tau_{in})} \tag{6.13}$$

where A_0 is $-g_{m1}R_c$, and

$$\tau_{in} = R_f(C_d + C_s + C_f + C_{\pi 1} + (1 - A_0)C_{c1}) \tag{6.14}$$

(The inclusion of the parasitic feedback capacitance, C_f, in (6.14) arises from the use of the *open-loop* time constant; that is, the feedback network is placed across the input node.) Thus $Z_o(s)$ is

$$Z_o(s) = \frac{A_0 R_f}{1 + s\tau_{in}} \tag{6.15}$$

Also, $Z_f(s)$ is given by

$$Z_f(s) = \frac{R_f}{1 + s\tau_f} \tag{6.16}$$

where τ_f is the feedback circuit time constant, $R_f C_f$. So, we can write (6.11) as

$$\frac{1}{Z_c(s)} = \frac{(1 + s\tau_{in})}{A_0 R_f} - \frac{(1 + s\tau_f)}{R_f} \tag{6.17}$$

or

$$Z_c(s) = \frac{A_0 R_{eff}}{1 + sR_{eff}(C_d + C_s + C_{\pi 1} + (1 - A_0)(C_{c1} + C_f))} \tag{6.18}$$

where

$$R_{eff} = \frac{R_f}{1 - A_0} \tag{6.19}$$

It is interesting to note that if A_0 is large enough, (6.18) will reduce to

$$Z_c(s) = \frac{R_f}{1 + sR_f(C_{c1} + C_f)} \tag{6.20}$$

which is the transimpedance for an ideal amplifier. In practice, this condition is difficult to achieve. This is because a large voltage gain may cause the preamplifier to become unstable, owing to the movement of the closed-loop poles within the feedback loop.

We could have obtained equation (6.18) directly by applying Miller's theorem to the feedback loop. However, if the receiver has two significant poles within the feedback loop, that is τ_c is not << τ_{in}, then we must perform the previous analysis with the two-pole version of A_0. So

$$A_0(s) = \frac{A_1}{(1 + s\tau_{in})} \times \frac{A_2}{(1 + s\tau_c)} \tag{6.21}$$

We will return to this point when we consider common-collector input preamplifiers.

6.2.2 *Noise analysis*

If a transimpedance preamplifier has a FET input stage, then the noise characteristic will be the same as for the PINFET, provided we replace R_b by R_f in (6.8). However, the transistor noise sources for a BJT are: the base current shot noise, $2qI_b$ A^2/Hz; the base-spreading resistance thermal noise, $4kTr_{bb'}$ V^2/Hz; and the collector current shot noise, $2qI_c$ A^2/Hz.

The base current shot noise and the feedback resistor thermal noise appear as current generators connected to the input node, and so can be easily accounted for. In addition, we can treat the collector current shot noise in a similar manner to the channel noise of the FET. However, the $r_{bb'}$ noise generator is a series generator, and so we must divide by the source impedance to convert it to a current generator. Thus, if we neglect r_π, we can write the total equivalent input m.s. noise current for a digital receiver as

$$<i_n^2>_c = \frac{4kTI_2B}{R_f} + 2qI_bI_2B + \frac{2qI_{c1}}{g_{m1}^2}\left[\frac{I_2B}{R_f^2} + (2\pi C_T)^2I_3B^3\right]$$
$$+ 4kTr_{bb'}\left[\frac{I_2B}{R_f^2} + (2\pi C_1)^2I_3B^3\right] \qquad (6.22)$$

where $C_1 = C_d + C_s + C_f$. If R_f is made very large, so that we can neglect its noise, (6.22) becomes

$$<i_n^2>_c = 2qI_bI_2B + \frac{2qI_{c1}}{g_{m1}^2}(2\pi C_T)^2I_3B^3$$
$$+ 4kTr_{bb'}(2\pi C_1)^2I_3B^3 \qquad (6.23)$$

which represents the minimum noise in a bipolar digital receiver, be it a transimpedance design or an integrating front-end design. Comparison with the corresponding PINFET equation (6.10) shows that there are two extra terms: the I_b noise term; and the $r_{bb'}$ term. We can only reduce these terms by employing high gain, low $r_{bb'}$ transistors.

It is interesting to note that the I_b shot noise term is proportional to I_c, while the collector current shot noise term is inversely proportional to I_c. (This can best be seen by substituting for g_m as I_c/V_T.) Thus there should be an optimum value of collector current, $I_{c,\,opt}$, that minimises the total noise. We can find this optimum by differentiating (6.22) with respect to I_c, and equating the result to zero. Hence, $I_{c,\,opt}$ is given by

$$I_{c,\,opt} = 2\pi V_TC_T\beta^{\frac{1}{2}}B(I_3/I_2)^{\frac{1}{2}} \qquad (6.24)$$

For analogue receivers, the equivalent equation is

$$I_{\mathrm{c,\,opt}} = 2\pi V_{\mathrm{T}} C_{\mathrm{T}} \beta^{\frac{1}{2}} B_{\mathrm{eq}}/\sqrt{3} \tag{6.25}$$

(It should be noted that we are assuming C_{T} to be independent of bias. In reality, $C_{\pi 1}$ varies with bias, and so we have to find $I_{\mathrm{c,\,opt}}$ by constant iteration.) If we substitute (6.24) back into (6.23), then the minimum noise from a bipolar front-end preamplifier will be

$$\begin{aligned} <i_{\mathrm{n}}^2>_{\mathrm{c,\,min}} &= (8\pi kT)(C_{\mathrm{T}}/\beta^{\frac{1}{2}})(I_2 I_3)^{\frac{1}{2}} B^2 \\ &\quad + 4kTr_{\mathrm{bb}'}(2\pi C_1)^2 I_3 B^3 \end{aligned} \tag{6.26}$$

Comparison with the minimum noise from a PINFET receiver, (6.10), shows that, *provided the* $r_{\mathrm{bb}'}$ *noise is insignificant*, the m.s. noise from a BJT front-end receiver increases as the square of the data-rate, whereas the m.s. noise from a PINFET receiver increases as the cube of the data-rate. So, although a PINFET receiver may be more sensitive than a BJT receiver at low data-rates, at high data-rates (typically >1 Gbit/s) the BJT receiver can be more sensitive. The f_{T}, β and $r_{\mathrm{bb}'}$ of the transistor will determine the exact crossover point.

6.2.3 Dynamic range

If the bandwidth of a transimpedance preamplifier is high enough so that no integration takes place, then the dynamic range can be set by the maximum voltage swing available at the preamplifier output. As the output stage is normally an emitter follower, running this stage at a high current will increase the voltage swing.

If the final stage is not the limiting factor, the dynamic range will be set by the maximum voltage swing available from the gain stage. With the design considered, the collector–base voltage of the front-end is a V_{be} (0.75 V) when there is no diode current. In a digital system, this results in a maximum peak voltage of approximately 1 V. The optical power at which this occurs can be easily found by noting that the peak signal voltage will be $I_{\mathrm{max}} R_{\mathrm{eff}} A_{\mathrm{o}}$. Thus it is a simple matter to find the maximum input current, and hence the maximum optical power. In most practical receivers, the dynamic range is greater than 25 dB.

6.2.4 Design example

A bipolar transimpedance receiver is constructed using state-of-the-art surface mount components on a p.c.b. The diode capacitance is dominated by the package, and has a value of 0.8 pF. The microwave transistors used in the design have a collector–base capacitance of 0.3 pF, an f_{T} of 4 GHz, a current gain of 120, and an r_{bb} value of 10 Ω. The front-end transistor is biased with 2 mA of collector current. Complete the design, and estimate the sensitivity if the receiver is to detect 140 Mbit/s data with an error rate of 1 in 10^9 bits.

In order to find the receiver bandwidth, we must initially determine the location of the open-loop poles. Let us first take a front-end voltage gain of 20 and a feedback resistance of 4 kΩ. As the receiver is fabricated on a p.c.b. using surface mount components, the parasitic capacitance associated with the feedback resistor is 0.1 pF. With these parameters, we find that the front-end time constant is, from equation (6.14)

$$\begin{aligned}\tau_{in} &= R_f(C_d + C_s + C_f + C_{\pi 1} + (1 - A_0)C_{c1})\\ &= 48\,\text{ns}\end{aligned}$$

The second pole is due to the collector load of the front-end transistor interacting with the input impedance of the following stage. We require a voltage gain of 20 from the front-end, and so the value of the collector resistor must be 250 Ω with a bias current of 2 mA. As the second stage is an emitter follower, the input resistance will be very high, while the input capacitance will be 0.3 pF. Thus the second pole will have a time constant given by

$$\begin{aligned}\tau_c &= 250 \times 0.3 \times 10^{-12}\\ &= 75\,\text{ps}\end{aligned}$$

As this is far lower than the front-end time constant, we can assume that the design has a single pole transfer function. (In reality we should use the two-pole form, and then determine whether the response is single pole in form. We will assume that this has been done!) Thus we can use equation (6.18) to give the bandwidth of the receiver as

$$\begin{aligned}f_{3\,\text{dB}} &= \frac{1}{2\pi R_{\text{eff}}(C_d + C_s + C_{\pi 1} + (1 - A_0)(C_{c1} + C_f))}\\ &= \frac{1}{2\pi 190(0.8 + 0 + 2.9 + 8.4) \times 10^{-12}}\\ &= 70\,\text{MHz}\end{aligned}$$

As the receiver is to detect 140 Mbit/s data, this bandwidth is correct. If we design the receiver to have a very high voltage gain, we can use equation (6.20) which gives a bandwidth of 400 MHz. This clearly shows the desirability of a high voltage gain. However, we should remember that this preamplifier has a two-pole response, and so stability requirements may limit the maximum voltage gain.

Let us now examine the noise performance of this design. The equivalent input noise current is, from equation (6.22)

$$\begin{aligned}<i_n{}^2>_c &= \frac{4kTI_2B}{R_f} + 2qI_bI_2B + \frac{2qI_{c1}}{g_{m1}{}^2}\left\{\frac{I_2B}{R_f^2} + (2\pi C_T)^2 I_3 B^3\right\} \\ &\quad + 4kTr_{bb'}\left\{\frac{I_2B}{R_f^2} + (2\pi C_1)^2 I_3 B^3\right\} \\ &= 3.3 \times 10^{-16} + 4.2 \times 10^{-16} + 6.7 \times 10^{-22} + 2.0 \times 10^{-18} \\ &= 7.52 \times 10^{-16}\ \mathrm{A}^2\end{aligned}$$

This results in a sensitivity of −37.8 dBm. It is interesting to note that the collector current shot noise is insignificant in comparison with the base current shot noise. This is due to the front-end current being above the optimum value. From (6.24) this optimum current is 0.4 mA and, if we use this value, then $<i_n{}^2>_c$ is $4.8 \times 10^{-16}\ \mathrm{A}^2$ – a sensitivity of −38.8 dBm. (The change is not very dramatic because the R_f noise is dominant.) As noted previously, we can only obtain the optimum collector current by repeated calculation.

By way of comparison, an optimally biased BJT in the integrating front-end receiver previously considered would produce a sensitivity of −41.10 dBm (about 9 dB less than the FET receiver). However, at a data-rate of 1 Gbit/s, the difference in sensitivity reduces to 4.6 dB, which clearly shows that BJT receivers will have an advantage at high data-rates.

6.3 Common-collector front-end transimpedance designs

The major disadvantage of CE input designs is that, by virtue of the gain between the collector and base of the front-end transistor, the collector–base capacitance appears as a large capacitance at the input node. A high input capacitance implies a low feedback resistance (to obtain a high bandwidth) and so the thermal noise will be high. However, we can use a cascode input, which has a very low input capacitance, to get a large bandwidth. Unfortunately, cascode designs require a voltage reference to bias the CB stage correctly, and this can lead to a more complicated design.

One way of eliminating the input Miller capacitance is to use a common-collector, *CC*, input. As CC stages have a very high input resistance, preamplifiers using this input configuration will be a better approximation to the ideal amplifier than those using CE input stages. Unfortunately, CC

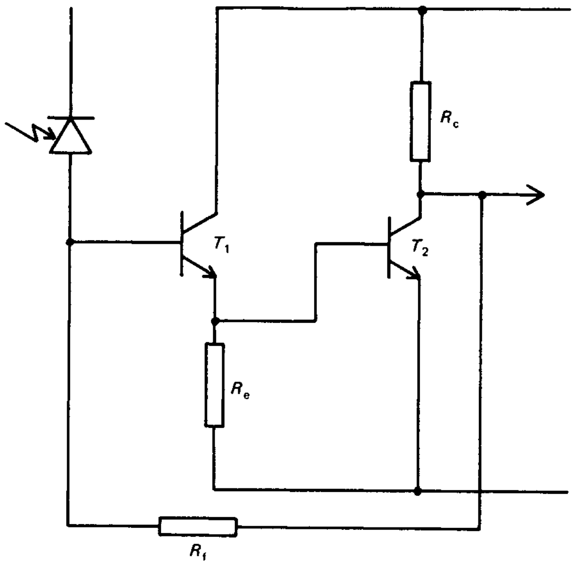


Figure 6.5 A simple common-collector/common-emitter, shunt feedback transimpedance receiver

stages do not exhibit voltage gain and so, as figure 6.5 shows, an amplifying stage has to follow the front-end.

6.3.1 Frequency response

The frequency response of CC input receivers is generally dominated by two poles: one is due to the front-end, while the other is due to the input time constant of the CE stage. By following a similar analysis to that used with the previous transimpedance design, we can write $A_0(s)$ as

$$A_0(s) = \frac{A_1}{(1 + s\tau_{in})} \times \frac{A_2}{(1 + s\tau_c)} \tag{6.27}$$

where τ_{in} is

$$\tau_{in} = R_f(C_d + C_s + C_f + C_{c1}) \tag{6.28}$$

and τ_c is given by

$$\tau_c = R_{C_{\pi 2}}(C_{\pi 2} + (1 - A_2)C_{c2}) \tag{6.29}$$

Here $R_{C_{\pi 2}}$ is the resistance in parallel with $C_{\pi 2}$, given by

$$R_{C\pi2} = \frac{(R_{o1} + r_{bb'2})r_{\pi2}}{R_{o1} + r_{bb'2} + r_{\pi2}} \tag{6.30}$$

where R_{o1} is the open-loop, output resistance of the front-end. Thus $Z_c(s)$ is given by

$$Z_c(s) = \frac{-A_0R_{eff}}{1 + sR_{eff}(C_d + C_s + C_{cl} + (1 - A_0)C_f + \tau_c/R_f) + s^2R_{eff}C_{in}\tau_c} \tag{6.31}$$

If the front-end pole is dominant, (6.31) simplifies to

$$Z_c(s) = \frac{-A_0R_{eff}}{1 + sR_{eff}(C_d + C_s + C_{cl} + (1 - A_0)C_f)} \tag{6.32}$$

Comparison with the equivalent equation for the CE design, (6.18), shows that the capacitive term is reduced. Therefore a CC input amplifier will have a greater R_f value than a CE design with the same bandwidth. Exactly the same conclusion applies to common-source FET input transimpedance preamplifiers.

We should note that $r_{bb'}$ affects the location of the second-stage pole – a high $r_{bb'}$ value results in a large $R_{C\pi2}$, and hence a large τ_c. A large τ_c results in s_2 being low in frequency, and this may cause the preamplifier transient response to exhibit undesirable over- and under-shoots. Hence a low $r_{bb'}$ will benefit both the receiver transfer function and noise.

6.3.2 Noise analysis

The noise performance of a CC stage is similar to that of a CE stage. However, because the voltage gain of a CC stage is unity or less, there will be some noise from the second-stage $r_{bb'}$. This additional noise term, $<i_n^2>_2$, is given by

$$<i_n^2>_2 = \frac{4kTr_{bb'2}}{A_1^2}\left[\frac{I_2B}{R_f^2} + (2\pi C_1)^2I_3B^3\right] \tag{6.33}$$

where A_1 is the voltage gain of the CC stage. Adding this to the terms in (6.22) gives the total equivalent input m.s. noise current. Again we see the importance of using low $r_{bb'}$ transistors.

In conclusion, although there is an extra noise term with CC input preamplifiers, these designs do allow for the use of a greater value of R_f, and this may produce a net reduction in noise in comparison with a CE input design.

6.3.3 Design example

The bipolar transimpedance receiver of the previous example is now designed to use a common-collector front-end. The same components and fabrication methods are employed as previous. Complete the design, and estimate the receiver sensitivity.

In order to make a fair comparison with the CE design examined previously, we will use a CC input receiver with the same voltage gain and transistor parameters as before. So, for a voltage gain of 20, we require the collector load of the second stage to be 250 Ω for a bias current of 2 mA. Now, the resistance in parallel with $C_{\pi 2}$ is, equation (6.30)

$$\begin{aligned} R_{C_{\pi 2}} &= \frac{(R_{o1} + r_{bb'2})r_{\pi 2}}{R_{o1} + r_{bb'2} + r_{\pi 2}} \\ &= \frac{(138 + 10)1.5 \times 10^3}{138 + 10 + 1.5 \times 10^3} \\ &= 135\,\Omega \end{aligned}$$

Thus τ_c is, equation (6.29)

$$\begin{aligned} \tau_c &= R_{C_{\pi 2}}(C_{\pi 2} + (1 - A_2)C_{c2}) \\ &= 135(2.9 + 6.3) \times 10^{-12} \\ &= 1.2\,\text{ns} \end{aligned}$$

We know that the receiver is likely to have a two pole response. In order to determine the value of the feedback resistor, we must use equation (6.31) repeatedly until we get the required bandwidth. If this is done, we find that $R_f = 18\,\text{k}\Omega$ for a 70 MHz bandwidth. So, as a check, equation (6.31) gives

$$\begin{aligned} Z_c(s) &= \frac{-A_0 R_{\text{eff}}}{1 + sR_{\text{eff}}(C_d + C_s + C_{c1} + (1 - A_0)C_f + \tau_c/R_f) + s^2 R_{\text{eff}} C_{\text{in}} \tau_c} \\ &= \frac{-18 \times 10^3}{1 + s857(0.8 + 0 + 0.3 + 2.1 + 0.07) \times 10^{-12} + s^2 1.2 \times 10^{-18}} \\ &= \frac{-18 \times 10^3}{1 + s2.7 \times 10^{-9} + s^2 1.2 \times 10^{-18}} \end{aligned}$$

We can find the bandwidth by solving the quadratic in the denominator of this equation. Thus the receiver has two real poles at 74 MHz and 284 MHz.

As regards the noise performance of this design, we use equation (6.22) together with equation (6.33) to give

$$
\begin{aligned}
<i_n{}^2>_c &= \frac{4kTI_2B}{R_f} + 2qI_bI_2B + \frac{2qI_{c1}}{g_{m1}^2}\left\{\frac{I_2B}{R_f{}^2} + (2\pi C_T)^2I_3B^3\right\} \\
&\quad + 4kTr_{bb'}\left\{\frac{I_2B}{R_f{}^2} + (2\pi C_1)^2I_3B^3\right\} \\
&\quad + \frac{4kTr_{bb'2}}{A_1{}^2}\left\{\frac{I_2B}{R_f{}^2} + (2\pi C_1)^2I_3B^3\right\} \\
&= 0.7 \times 10^{-16} + 4.2 \times 10^{-16} + 6.7 \times 10^{-22} \\
&\quad + 2.0 \times 10^{-18} + 2.2 \times 10^{-18} \\
&= 4.9 \times 10^{-16}\, A^2
\end{aligned}
$$

This results in a sensitivity of −38.75 dBm (an increase over the CE design of 0.77 dB). If we bias the front-end at the optimum collector current, then the increase in sensitivity is 1.7 dB, compared with an increase of only 1 dB for the CE input design. We can account for this difference by noting that the R_f noise is more dominant in the CE design, and this tends to mask the advantage. A further advantage of CC input preamplifiers is that, unlike CE input designs, they generally maintain a flat frequency response when optimally biased.

For further background reading, see references [1] to [6].

7 Current Systems and Future Trends

In previous chapters, we concentrated on the design and performance of individual components for use in optical links. What we have not yet examined is the overall design of practical links, and it is this that initially concerns us here.

When designing an optical link, system designers commonly use a *power budget* table which details the power losses encountered from source to receiver. This table enables the designer to implement system margins to account for ageing effects in the links. We will use the power budget to contrast two general cases: a low-speed data link using PCS fibre and LEDs, and a high-speed telecommunications link using all-glass fibre and lasers. We will then examine the design of two current optical communications links.

In the rest of this chapter we will examine some advanced components and systems that are being developed in the laboratory, and assess their likely impact on optical communications.

7.1 System design

The examples we consider here are an 850 nm wavelength, 10 Mbit/s link, operating over 500 m; and a long-haul, 1.55 μm wavelength, 1.2 Gbit/s link. As the length of the long-haul route has not been specified, we will use the power budget table to determine the repeater spacing. (Although these examples are tutorial in form, they will illustrate the basic principles behind link design.)

Table 7.1 shows the power budget for the two links. Because the short-haul link operates at a low date-rate, we can specify PCS fibre and an LED source. We will also assume that the link is made up of five, 100 m lengths of fibre, requiring four pairs of connectors. For the high-speed link, we will take laser diode sources and 1 km lengths of dispersion shifted, single-mode, all-glass fibre, connected together with fusion splices.

The last two parameters in the table, *headroom* and *operating margin*, represent excess power in the link. The headroom parameter is included to allow for the insertion of extra connectors, or splices, should a break in the fibre occur, as well as accounting for any power changes due to the effect of age. The headroom for the laser link is larger than for the LED link

Table 7.1 Link power budgets for a short-haul, low-data-rate link, and a long-haul, high-data-rate link.

	Short-haul link (10 Mbit/s)		*Long-haul link (1.2 Gbit/s)*	
Launch power	LED	–15 dBm	Laser	–3 dBm
Receiver sensitivity		–45 dBm		–34 dBm
Allowable loss		30 dB		31 dB
Source coupling loss		3 dB		2 dB
Fibre loss	(6 dB/km)	3 dB		0.2 dB/km
Joint loss	(2 dB/pair)	8 dB		0.2 dB/splice
Detector coupling loss		3 dB		2 dB
Headroom		5 dB		8 dB
Operating margin		8 dB		7 dB

because laser output power falls with age, whereas that of an LED remains relatively constant. The operating margin may be taken up by manufacturing variations in source power, receiver sensitivity and fibre loss. It will also allow for the addition of extra components such as power splitters and couplers.

We can find the distance between repeaters in a long-haul link by noting that the number of fibre–fibre splices is one less than the number of fibre sections. So, with an allowable fibre and splice loss of $31-19 = 12$ dB, it is a simple matter to show that the maximum number of fibre lengths is 30, with 29 fusion splices. Thus the maximum length between repeaters is 30 km. By reducing the headroom and operating margins, the maximum transmission distance in both links could be increased. However, this would not allow for the inclusion of power splitters or couplers at a later date.

Although the power budget gives an indication of the maximum link length, it does not tell us whether the links can transmit the required data-rate. From our previous discussions, the system bandwidth up to the input of the pre-detection filter, $f_{3\,\mathrm{dB}}$, should be at least half the data-rate, that is

$$f_{3\,\mathrm{dB}} \geq \frac{B}{2} \tag{7.1}$$

We can relate this bandwidth to the rise-time of the pulses, τ, at the input to the filter using

$$f_{3\,\mathrm{dB}} = \frac{0.35}{\tau} \tag{7.2}$$

(Although this equation only applies to a network with a single-pole response, the error involved in the general use of (7.2) is minimal.) If we combine (7.1) and (7.2), the minimum rise-time is given by

$$\tau \leq \frac{0.7}{B} \qquad (7.3)$$

We can find the system rise-time by adding the rise-times of individual components on a mean square basis, that is

$$\tau^2 = \Sigma\tau^2_{\mathrm{n}} \qquad (7.4)$$

(This equation results from convolving the impulse response of the individual components, to find the overall impulse response, and hence the rise-time.) Most sources are characterised by the rise-time of the optical pulses, while the receiver bandwidth is often quoted. However, as we saw in chapter 2, optical fibre is often characterised by the pulse dispersion, and the impulse response can take on several different shapes. If we assume a Gaussian shape impulse response, then the rise-time can be approximated by

$$\tau_{\mathrm{fibre}} \approx 2.3\sigma \qquad (7.5)$$

where σ is the total fibre dispersion. So, if we return to the short-haul link, a fibre bandwidth of 35 MHz km (optical) gives a dispersion of 2.7 ns for a 500 m length, resulting in a rise-time of 6.2 ns. (This assumes that the bandwidth is limited by modal dispersion. Hence we can neglect the linewidth of the LED.) A 10 ns LED rise-time and a receiver bandwidth of 10 MHz gives a total system rise-time of 37 ns. From (7.3), this results in a maximum data-rate of about 20 Mbit/s, and so the link is adequate for the 10 Mbit/s transmission speed that we require.

For the long-haul route, we will assume that the total fibre dispersion is 1.2 ps/nm/km. A laser linewidth of 1 nm yields a dispersion of 36 ps for the 30 km link length. This results in a rise-time of approximately 83 ps which, together with a laser rise-time of 150 ps and a receiver bandwidth of 800 MHz, yields a system rise-time of 470 ps. For transmission at 1.2 Gbit/s, the maximum rise-time should be 580 ps, and so the link will just transmit the required data-rate.

These results, together with the link budget, indicate that even if we cut the operating margin on the long-haul route, the link could not be extended very far, because of dispersion effects. Under these conditions, the link is said to be *dispersion limited.* However, the length of the short-haul route is determined by attenuation, that is the link is *attenuation limited,* and so the link could be extended by reducing the operating margin. We should note that, because of the approximations involved in the calculation of link capacity, the actual data-rates that can be carried are greater than those indicated. Hence the use of these formulae already allows an operating margin with both bit-rate and attenuation.

7.2 Current systems

In this section we will briefly examine the first optical transatlantic cable, *TAT8*, and, by way of contrast, a computer communications link operating at the Joint European Torus at Culham, UK. Although the technology employed in these links is very different, reliability is important in both cases.

TAT8 is the first optical transatlantic communications link. The cable has been laid in three sections, with a different manufacturer taking responsibility for each. The first section has been designed by the American Telephone and Telegraph Co., *AT&T*, and is a 5600 km length from the USA to a branching point on the continental shelf, to the west of Europe. At this point, the cable splits to France and the UK. The French company *Submarcom* designed the 300 km length link to France, while the British company Standard Telephones and Cables, *STC*, were responsible for the 500 km link to the UK.

In view of the link length, dispersion effects are highly important and so the operating wavelength is 1.3 μm. The use of single-mode laser diodes yields a total fibre dispersion of 2.8 ps/nm/km, and so the regenerator spacing is limited by fibre attenuation, not dispersion. The InGaAsP laser diode sources launch a minimum of approximately -6 dBm into single-mode fibre. With an average receiver sensitivity of -35 dBm for 10^{-9} error rate, the allowable loss over a repeater length is 29 dB. A typical operating margin of 10 dB, and a fibre attenuation of 0.48 dB/km, result in a repeater spacing of 40–50 km.

At each repeater, PIN photodiodes feed BJT transimpedance preamplifiers. The signals are then amplified further, prior to passing through pre-detection filters to produce raised-cosine spectrum pulses. Bipolar transistors are used throughout the regenerator because they are generally more reliable than GaAs MESFETs. As the preamplifier is a transimpedance design, front-end saturation does not occur, and so an mBnB line-code does not have to be used. Instead, the TAT8 system uses an even-parity code, with a parity bit being inserted for every 24 transmission bits (a *24B1P* code). As such a code has low timing content, surface acoustic wave, *SAW*, filters with a *Q* of 800 are used in the clock extraction circuit. Should the timing circuit fail in a particular regenerator, provision is made for the data to be sent straight through to the output laser, so that the data can be re-timed by the next repeater.

The TAT8 cable comprises six individual fibres; two active pairs carry two-way traffic at a data rate of 295.6 Mbit/s on each fibre. As well as parity bits, some of the transmitted bits are used for system management purposes and so the total capacity is 7560 voice channels. This should be compared with the 4246 channels available on the TAT7 co-axial cable link. Control circuitry enables a spare cable to be switched in if one of the active cables fails. As well as having spare cables, provision is made to switch in stand-by lasers (a photodiode placed on the non-emitting facet provides a measure of laser health). To increase system reliability further, redundant circuits are included in each regenerator.

By way of contrast, engineers at the Joint European Torus, *JET*, fusion reaction experiment use an optical link to transmit computer data around the site. Although the environment is electrically noisy, the main reason for the use of an optical link is that of electrical isolation. This is because earth loop faults could cause a high difference in earth potential between the ends of the data link, and this would prove fatal to a hard-wired link.

The system is basically an optical local area network, *LAN*, with a host computer controlling remote equipment. The maximum distance between terminals is typically 600 m, which is determined by physical constraints. This maximum distance, together with the 10 Mbit/s data rate, means that 200 μm core, 10 MHz km, PCS fibre can be specified. The attenuation at the 820 nm operating wavelength is typically 8 dB/km.

Packaged surface emitting LEDs, with a typical output power of −12 dBm, are used as the sources. At the receiver, a package Si PIN photodiode supplies a signal current to a transimpedance preamplifer. The receiver sensitivity is approximately −30 dBm, which results in an allowable link loss of 18 dB. As the fibre attenuation is 8 dB/km, and each link uses two connectors with a typical loss of 3 dB, the operating margin over a 600 m length is approximately 11 dB. The excess of received optical power means that a pre-detection filter is not required. To ease dynamic range restrictions, provision is made to reduce the transmitted power by 3 dB. (As the sources are LEDs, the output power is directly proportional to the drive current. Hence a 3 dB drop in power can be achieved by halving the drive current.) This feature is also useful when commissioning new links; if the link functions satisfactorily at half power, it will function well with full power.

As the communications system is a LAN, data may have to be sent through a large number of terminals before it reaches the destination terminal. In view of this, each terminal extracts a clock from the data, and regenerates the signal. To ensure a strong clock signal, Manchester encoded data is used. The link controllers are housed in readily accessible locations, and so maintenance of the equipment is not a major problem. However, if the error-rate over a particular link increases owing to increased fibre attenuation, reduced LED power, or lower receiver sensitivity, the entire network could collapse. To maintain transmission, a back-up link is installed. By comparing the synchronisation code in the received data frame to that of the ideal code, the error-rate over the main link can be monitored. If errors are present, then data transmission can be automatically switched to the back-up link.

7.3 Future trends

In this section, we will consider some of the latest advances in optical communications. Most of our study will be descriptive in form, and we begin by examining optical fibres which exhibit very low loss at wavelengths above 1.55 μm.

7.3.1 Fluoride-based optical fibres

One of the latest advances in optical fibres is the development of single-mode optical fibres, which exhibit very low-loss in the mid infra-red region, above 2 μm. As we saw earlier, Rayleigh scattering reduces as wavelength to the fourth power, and so very low-loss transmission requires operation at long wavelengths. However, in silica fibres, the absorption increases rapidly for wavelengths above the 1.55 μm window, and so very low-loss fibres have to be made from different materials.

The most promising glasses for low-loss fibres are those based on fluoride compounds (France *et al.*, [1]). Of these zirconium fluoride, ZrF_4, and beryllium fluoride, BeF_2, glasses have projected attenuations at 2.5 μm of 0.02 dB/km and 0.005 dB/km respectively. Unfortunately $\mathrm{BeF_2}$ is highly toxic, and so most of the work has been concerned with $\mathrm{ZrF_4}$ glasses. Probably the most suitable composition for fibre drawing is $\mathrm{ZrF_4}$–$\mathrm{BaF_2}$–$\mathrm{LaF_3}$–$\mathrm{AlF_3}$–NaF, usually abbreviated to *ZBLAN*. In ZBLAN fibres, the 2.87 μm fundamental of the OH bond causes a high level of attenuation. However, there is a transmission window at 2.55 μm, in which a measured loss of 0.7 dB/km has been recorded. Investigations show that the major loss mechanism is scattering from imperfections formed in the fibre during manufacture. So, with a more refined process, attenuations close to the Rayleigh scattering limit should be achievable.

The dispersion of ZBLAN fibres is highly dependent on the fibre structure. With an index difference of 0.014 and a core diameter of 6 μm, the dispersion is about 1 ps/nm/km, whereas a fibre with an index difference of 0.008 and a core diameter of 12 μm has a dispersion of greater than 15 ps/nm/km. By themselves, these dispersion times are very low; however, these fibres are likely to be used to transmit signals over very long distances, and so the total dispersion could be significant.

The move to higher wavelengths is likely to see a new generation of lasers and detectors. The most promising semiconductor laser source is a double heterojunction SLD, based on *InAsSbP* matched to an InAs substrate. InGaAs photodiodes can operate at long wavelengths, but lead sulphide, *PbS*, detectors also show promise.

7.3.2 Optical solitons

Optical solitons are produced by the interaction of high-energy optical pulses and certain non-linear effects in optical fibre. Their special characteristic is that they retain their shape for many hundreds if not thousands, of kilometres. Thus an optical communications link that uses solitons will not suffer from dispersion. Obviously such a link would be very desirable, but before we examine soliton generation and propagation in optical fibre, let us examine the historical background of solitons in general.

In Victorian times, Scott Russell observed a 'solitary wave' travelling along the Union Canal in Scotland. Russell followed the wave for several miles, observing that the wave did not dissipate as normal waves should do. Unfortunately Victorian mathematics was not sufficiently advanced to explain the propagation of this wave. It was not until the late 20th Century that mathematics become advanced enough to solve the non-linear equations that governed Russell's wave. The solutions to these equations were termed *solitons*, and it was in 1972 that two Soviet theoretical physicists predicted the possibility of optical solitons (Zakharov and Shabat, [2]).

Optical solitons in fibre are produced by the interaction between the material dispersion and a non-linear variation in refractive index. As we have already seen, the material dispersion passes through zero at a wavelength of 1.3 μm. Above this wavelength, shorter wavelength signals tend to travel faster than longer wavelength ones, and so the longer wavelength components of an optical pulse appear in the trailing edge of the pulse and the pulse is dispersed. In a soliton, this broadening of the pulse is exactly balanced by the effects of a non-linear refractive index.

Let us consider a bell-shaped pulse, of the form shown in figure 7.1, propagating through a length of optical fibre. If the wavelength of operation is greater than 1.3 μm, the trailing edge of the pulse contains the slower

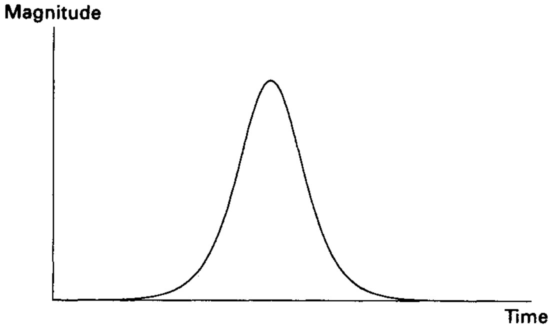


Figure 7.1 Soliton pulse in an optical fibre

long-wavelength pulse components. Taking the non-linear refractive index to be given by

$$n = n_0 + n_2 I \tag{7.6}$$

where I is the intensity of the pulse, we find that a wave of constant amplitude undergoes a phase shift per unit length of

$$\delta\phi = \frac{2\pi}{\lambda_0} n_2 I \tag{7.7}$$

We can differentiate (7.7) with respect to time to give

$$\frac{\mathrm{d}\delta\phi}{\mathrm{d}t} = \frac{2\pi}{\lambda_0} n_2 \frac{\mathrm{d}I}{\mathrm{d}t} \tag{7.8}$$

Now, on the leading edge of the pulse, $\mathrm{d}I/\mathrm{d}t$ is positive which results in the phase lag increasing as time goes by. Thus the leading edge is effectively slowed down. However, on the trailing edge, $\mathrm{d}I/\mathrm{d}t$ is negative and this results in a decreasing phase lag. Thus the trailing edge is speeded up. We can therefore see that the pulse is effectively shrunk by the non-linearity in refractive index (this is known as the *Kerr effect*). If this pulse shrinkage balances out the pulse dispersion, then we have a soliton pulse which is able to propagate for very large distances.

In order to send data using soliton pulses, the pulse shape should be sech^2 in form (shown in figure 7.1). However, any reasonably shaped pulse that has an area, R, that satisfies the inequality $R_0/2 < R < 3R_0/2$ will eventually evolve into a soliton with area R_0, and so specialised pulse shaping is not required. Any optical power not propagating as a soliton will eventually die away as a result of dispersion effects.

To propagate successfully over long distances, the amplitude of the soliton must be maintained so that the Kerr effect is continually present. However, optical signals suffer from attenuation, and so the solitons have to be regenerated after a certain distance. As dispersion is not a problem with soliton links, there is no need to regenerate the pulses fully. Instead, fibre amplifiers (refer to section 3.8) can be used, which means that soliton links can be capable of tremendous transmission distances (>500 km).

7.3.3 Advanced semiconductor lasers

In chapter 3, we briefly examined single-mode lasers based on ridge waveguide structures. Unfortunately, owing to manufacturing tolerances and temperature effects, the emission frequency of these lasers cannot be accurately controlled. In long-haul routes, using dispersion-shifted optical fibre, this could lead to undesirable dispersion effects. In single-mode *long external cavity, LEC,* lasers, a diffraction grating provides one of the laser facets (Mellis *et al.*, [3]). Such gratings act as wavelength selective filters with the reflected wavelength being dependent on the angle the grating makes to the incident light. So, if such a grating provides the optical feedback in a semiconductor laser, the wavelength of emission can be altered by varying the angle of the grating. For 1.55 μm lasers, such a scheme results in coarse tuning, by mechanical means, over a 50 nm range. Fine tuning over a 0.4 nm range is achieved by applying a voltage to a piezoelectric transducer, incorporated in the grating mounting. The linewidth of such lasers is typically 50 kHz and, as we shall see in section 7.3.5, such lasers are required for use in coherent detection receivers.

In chapter 3 we encountered semiconductor lasers based on a double heterojunction structure, with an active region thickness of typically 2 μm. As the threshold current density is directly proportional to the thickness of the active layer, we could reduce J_{th} by using thin active layers. However, if the active region is less than 100 Å (10 nm) thick, we have to take account of quantum mechanical effects. With such a thin active layer, the kinetic energy of the injected carriers becomes quantised. This effect is similar to the quantum mechanical problem of the one-dimensional potential well, hence these devices are known as *quantum-well* lasers (van der Ziel *et al.*, [4]). Such lasers exhibit a very low threshold current, typically less than 5 mA, and so they are very efficient devices.

Figure 7.2a shows the schematic of a typical GaAlAs/GaAs quantum-well laser. This structure is identical to that of a normal DH laser diode, except that the thickness of the GaAs active region is less than 100 Å. Under these conditions, quantum mechanics predicts that the effective band-gap of the active region is given by

$$E_{\text{active}} = E_g + \left[\frac{h^2}{8d^2}\right]\left[\frac{1}{m_e} + \frac{1}{m_{hh}}\right] \tag{7.9}$$

where d is the thickness of the active region, m_e is the effective mass of electrons in the CB, and m_{hh} is the effective mass of heavy holes in the VB. Thus we can see that the effective band-gap can be increased by reducing the thickness of the active region.

Rather than fabricate a diode with only one quantum-well, the so-called *single quantum-well* or *SQW* laser, quantum-wells can be cascaded in the same package to produce *multiple quantum-well* or *MQW* lasers. The band-gap distribution of such a device is shown in figure 7.2b. The obvious advantage of such a structure is that we effectively have a laser array, and so the light output can be very large. Indeed, GaAs MQW lasers have been fabricated into laser arrays with continuous operation output powers of 5 W

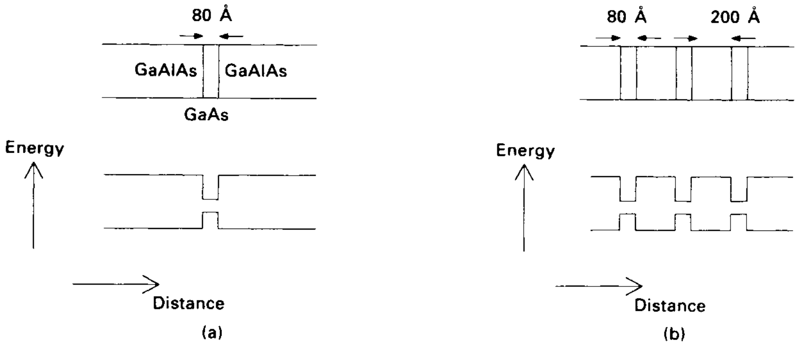


Figure 7.2 (a) Schematic of a typical GaAlAs/GaAs quantum-well laser and (b) band-gap distribution of a MQW laser

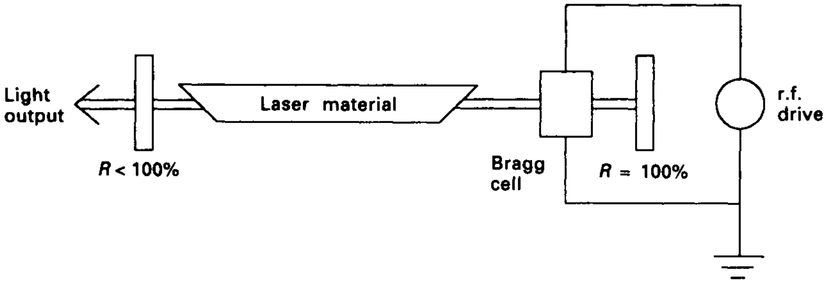


Figure 7.3 Mode-locking of a solid-state/gas laser

or more, and a conversion efficiency of 50 per cent. Such high powers are needed for pumping the fibre amplifiers that we encountered in section 3.8.

7.3.4 Mode-locking

In the modulation schemes we considered in section 3.7, we saw that the light output of a laser can be modulated by either varying the drive current, for a SLD, or by using some form of external modulator (usable with all types of lasers). One problem with these methods of modulation is that of speed – the maximum modulating speed is limited by the pulsing circuitry. In the method known as *mode-locking*, pulses can be generated which are typically less than 500 ps wide.

Figure 7.3 shows a possible arrangement for mode-locking a gas or solid-state laser. As can be seen, a Bragg cell is incorporated in the cavity. This Bragg cell is operated in continuous wave mode, and so the light that passes through the cell is shifted by the r.f. carrier frequency. If this carrier frequency is equal to the mode spacing of the cavity, the light output will have a very narrow linewidth, and consist of a series of pulses with spacing equal to the round trip transit time of photons in the cavity ($2L/c$). The width of these pulses is approximately equal to the inverse of the laser line-width, and so the pulses can be very short in duration (30 fs is the current minimum). One further advantage of a mode-locked laser is that the peak pulse power is N times the average laser power, where N is the number of allowed cavity modes (see section 3.5.2). This gain is because all the power that would have been in the individual modes is concentrated in just *one* propagating mode.

Given these properties, a mode-locked laser would appear to be an ideal source for use in high data-rate systems. However, we must still turn the train of pulses on and off, and this can cause some difficulty. One solution, when using a semiconductor laser, is to use a long cavity laser (approximately 2 cm long) incorporating a laser section, a gain section and a modulation section (Hansen *et al.*, [5]). The laser is mode-locked by modulating the gain section with the output of a c.w. oscillator. (This is similar to using a Bragg cell to modulate the gain in the laser cavity.) The mode-locked pulses then pass through the modulation section, which turns the pulses on and off. With such a scheme, the mode-locked pulses are produced at typically 2 Gbit/s, which is one of the data-rates being considered for use on long-haul routes. (For further information on mode-locking, interested readers are referred to Yariv [6]).

7.3.5 Coherent detection systems

So far we have only been concerned with optical signals whose amplitude varies in sympathy with a digital signal – amplitude shift keying, or *ASK*. However, ASK is not the only signalling format that can be used. Phase shift keying, *PSK*, or frequency shift keying, *FSK*, of an optical carrier are alternatives. (PSK can be easily generated by using an external phase modulator. FSK can be generated by modulating the drive current of a laser operating in saturation. This has the effect of varying the refractive index of the gain region in the active layer, so altering the laser frequency.) As the amplitude of PSK and FSK signals remains constant, we cannot use the direct detection receivers already considered. Instead, we must use *homodyne* or *heterodyne* receivers, which are collectively known as *coherent* receivers (Hodgkinson *et al.*, [7]). Heterodyning is used in most modern radio receivers, and so we will only examine this technique.

Figure 7.4 shows the schematic diagram of a heterodyne optical receiver suitable for the demodulation of ASK, FSK, or PSK. As can be seen, a fibre coupler is used to combine the output of a local oscillator, *l.o.*, laser and the received optical field. The resultant field, that is the sum of the

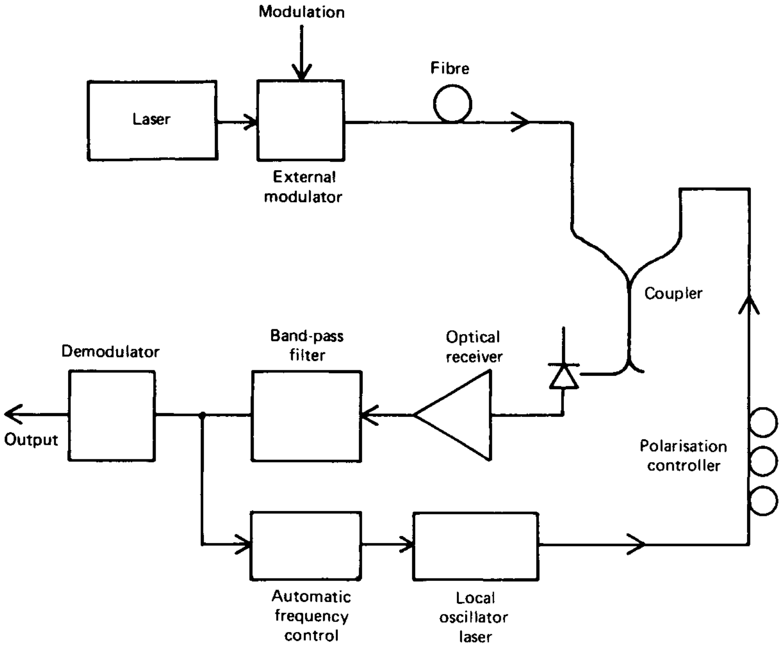


Figure 7.4 Schematic of an optical heterodyne detection system

individual fields, is then applied to an optical receiver. To analyse the receiver performance, let us consider a general case, where the received electric field, E_r, is

$$E_r = e_r\cos(\omega_r t + \phi) \tag{7.10}$$

and the local oscillator field, E_L, is

$$E_L = e_L\cos\omega_L t \tag{7.11}$$

So, we can write the resultant field, E_i, as

$$\begin{aligned} E_i &= E_r + E_L \\ &= e_r\cos(\omega_r t + \phi) + e_L\cos\omega_L t \end{aligned} \tag{7.12}$$

Now, the photodiode current is proportional to the incident power which, in turn, is proportional to the *square* of E_i. If we square E_i, we get

$$\begin{aligned} E_i^2 = {} & e_r^2\cos^2(\omega_r t + \phi) + e_L^2\cos^2\omega_L t \\ & + 2e_r e_L\cos(\omega_r t + \phi)\cos\omega_L t \end{aligned}$$

or in terms of optical power

$$P_i = P_r\cos^2(\omega_r t + \phi) + P_L\cos^2\omega_L t$$
$$+ 2\sqrt{P_r P_L}\cos(\omega_r t + \phi)\cos\omega_L t \tag{7.13}$$

Expansion of the cosine terms in (7.13) yields frequency components at d.c., $\omega_L - \omega_r$, $\omega_L + \omega_r$, $2\omega_L$ and $2\omega_r$. As we are considering light, the last three terms are very high in frequency and will not pass through the receiver. If the d.c. term is filtered out by coupling capacitors, then the only frequency component amplified by the receiver is the difference frequency, ω_{if}, given by

$$\omega_{if} = \omega_L - \omega_r \tag{7.14}$$

where ω_{if}, is the *intermediate frequency.* A demodulator can then recover the baseband signal. In *homodyne* receivers, the frequency and phase of the l.o. laser are the same as the received optical signal, and so $\omega_{if} = 0$.

Although the demodulated signal resembles the received signal, the amplitude has been increased by the local oscillator power, and this serves to increase receiver sensitivity. Unfortunately we cannot increase the receiver sensitivity indefinitely; an increase in l.o. power increases the diode shot noise by the same amount as the signal power (refer to equation 4.26). Hence the S/N reaches a limit known as the quantum limit for coherent detection.

Table 7.2 compares the quantum limit for heterodyne and homodyne detection of ASK, FSK and PSK signals. We can obtain this table by following a similar analysis to that presented in chapter 5. In the calculation of the results, we have assumed that the diode shot noise has a Gaussian probability density function, with mean square value identical to that of the photon arrival Poisson distribution.

As can be seen from the table, the detection of PSK, by either a heterodyne or homodyne receiver, results in a sensitivity greater than the direct detection quantum limit. Although the gain in sensitivity is not very great, we should remember that, because of the receiver noise, a direct detection receiver will not approach the quantum limit. So, in general, the use of coherent detection will result in greater receiver sensitivity (typically 16 dB more).

There are several practical difficulties associated with coherent detection. The most obvious is the requirement for very stable, tunable lasers that have a narrow linewidth. LEC lasers are ideally suited to this application. As the frequency difference between the l.o. and source lasers must be kept constant, an automatic frequency control, *AFC*, loop sets the frequency of the local oscillator laser.

Another difficulty is that the power in the demodulated signal depends on

Table 7.2 Comparison of quantum limits for heterodyne, homodyne and direct detection receivers (assuming a quantum efficiency of unity)

Receiver type	*Modulation format*	*Probability of error* (P_e)	*Number of photons per bit for* $P_e = 10^{-9}$
Heterodyne	ASK	$\frac{1}{2}\operatorname{erfc}\left(\frac{P_L\lambda_0}{4hcB_{eq}}\right)^{\frac{1}{2}}$	72
	FSK	$\frac{1}{2}\operatorname{erfc}\left(\frac{P_L\lambda_0}{2hcB_{eq}}\right)^{\frac{1}{2}}$	36
	PSK	$\frac{1}{2}\operatorname{erfc}\left(\frac{P_L\lambda_0}{hcB_{eq}}\right)^{\frac{1}{2}}$	18
Homodyne	ASK	$\frac{1}{2}\operatorname{erfc}\left(\frac{P_L\lambda_0}{2hcB_{eq}}\right)^{\frac{1}{2}}$	36
	PSK	$\frac{1}{2}\operatorname{erfc}\left(\frac{2P_L\lambda_0}{hcB_{eq}}\right)^{\frac{1}{2}}$	9
Direct detection	ASK	See chapter 4	21

the state of polarisation of the received and local oscillator fields; the maximum signal occurs when the two fields have the same polarisation state. In the laboratory this can be achieved by stressing the fibre. However, as the polarisation state at the end of a length of SM fibre can change with time, some form of automatic control is required for installed links. Such control can be achieved by winding a length of *polarisation preserving* fibre around a piezoelectric cylinder (Walker and Walker, [8]). Any voltage applied to the cylinder will stress the fibre, so altering the state of polarisation. If an automatic control loop supplies the cylinder voltage, then any changes in the polarisation of the received field can be tracked.

Until recently, coherent detection could not be demonstrated outside the laboratory, because of the difficulties we have just outlined. However, in October 1988, researchers at British Telecom Research Laboratories, Martlesham Heath, UK, were the first to demonstrate coherent detection over 176 km of installed glass fibre, without the use of intermediate repeaters (Creaner *et al.* [9]). At the transmitter, the output of a single-mode laser was passed through an external phase modulator. The signalling format used was differential phase shift keying, *DPSK* with a modulation speed of 565 Mbit/s. Prior to coupling into the fibre, the signal was amplified to 1 dBm by a travelling wave laser amplifier. At the end of the link, a heterodyne receiver demodulated the carrier, yielding a sensitivity of −47.6 dBm. For the experiment, LEC lasers and a piezoelectric polarisation controller were used.

Although coherent detection over an installed link was demonstrated for the first time, further development of the laser and polarisation control packages is required before the technique can become commonplace.

7.3.6 Optical broadcasting

With time division multiplexing, *TDM*, techniques, any increase in channel capacity results in an increase in transmission speed. This can place a strain on the digital processing circuitry; even if GaAs digital ICs are employed, the maximum data-rate is likely to be limited to less than 10 Gbit/s. An alternative approach is to use wavelength division multiplexing, *WDM*, in which different channels transmit on slightly different wavelengths. At the receiver, diffraction gratings filter out the required wavelength, prior to detection by an optical receiver. Such a scheme might be attractive for long-haul, high-capacity routes where the cost can be shared by many users. However, the requirement for individual diffraction gratings, receivers and associated processing circuitry makes this scheme unattractive for use in local-loop telecommunications.

In radio broadcast systems, frequency division multiplexing, *FDM*, techniques are commonplace, with each broadcasting station transmitting on a particular frequency. At the radio receiver, a local oscillator selects the required station using heterodyne techniques. A similar principle can also be applied to optical links (Brain [10]). At the optical receiver, the use of coherent detection, with variable frequency lasers, means that the required channel can be selected from those available.

The capacity of such a scheme could be very high. If we consider individual channel capacities of 565 Mbit/s operating with a spacing of 10 GHz (0.08 nm), we could transmit 400 channels within the 32 nm bandwidth of a wideband fibre amplifier. Although each channel would require a laser at the transmitter, the cost would be shared by a large number of consumers. All of the channels could be carried by a single optical fibre ring, with individual fibres going to the customer's premises. An alternative scheme, which does not require the use of expensive lasers or diffraction gratings, is *sub-carrier multiplexing, SCM.*

Figure 7.5 shows a 120-channel satellite broadcast scheme, together with the equivalent SCM system. As can be seen, in the SCM scheme a laser transmitter directly replaces the satellite. Rather than transmit the signals using free-space, as in the satellite system, the SCM scheme relies on optical fibre into the consumer's home. At the receiving end, the satellite dish has been replaced by a photodiode.

In common with the satellite receiver, the optical receiver has a frequency response that matches that of the received signal (2.7–7.5 GHz). Such an optical receiver is known as a *tuned front-end* receiver (Greaves and Unwin [11]). There are a number of advantages with such a receiver: standard mi-

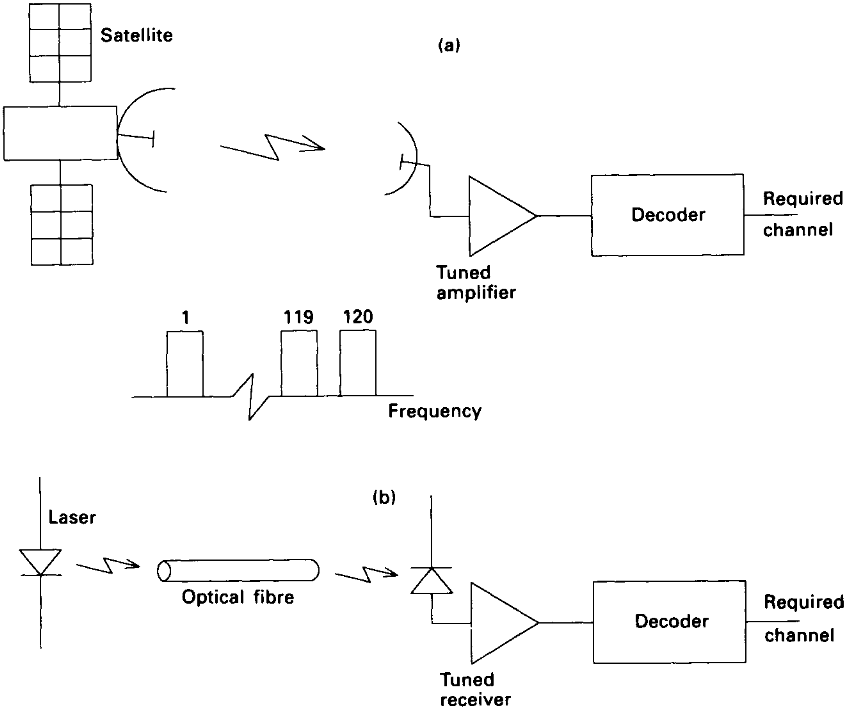


Figure 7.5 Schematic of (a) a satellite broadcast system and (b) an equivalent sub-carrier multiplex system

crowave design and fabrication techniques can be used to produce a microwave monolithic IC, *MMIC*, so keeping costs down; and, by carefully designing the input tuning circuit, the noise performance of the receiver can be improved over that of a more conventional wideband receiver.

At the output of both receivers, a standard satellite decoder can be used to demultiplex the required signal. Although the SCM scheme relies on the installation of optical fibre into the home, or close enough so that co-axial cable can be used, one obvious advantage of such a scheme is that the consumer would have access to a large number of television channels without the need for satellite receiver dishes.

Bibliography and References

In selecting the references for this text, I have included several books, and many technical papers. Of the books, some are quite specialised, and these appear under the relevant chapter headings. More general reference books appear at the end of the Bibliography. Of the technical papers, I have only listed one or two for each individual topic area. To obtain further information in these areas, the interested reader should examine the references that these papers mention. Although some of the papers listed here have not been referenced in the text, their relevance to a particular subject area should be obvious from their titles.

References

Chapter 1

1. Maimon, T. H. (1960). 'Stimulated optical radiation in ruby', *Nature*, **187**, 493–494.
2. Kao, C. K. and Hockman, G. A. (1966). 'Dielectric-fibre surface waveguides for optical frequencies', *Proc. of the IEE*, **113**, 1151–1158.
3. Personick, S. D. (1973). 'Receiver design for digital fiber optic communication systems, Parts I and II', *Bell System Tech. J.*, **52**, 843–886.
4. Smith, D. R., Hooper, R. C. and Garrett, I. (1978). 'Receivers for optical communications: a comparison of avalanche photodiodes with PIN–FET hybrids', *Optical and Quantum Electronics*, **10**, 293–300.

Chapter 2

1. Parton, J. E., Owen, S. J. T. and Raven, M. S. (1986). *Applied Electromagnetics*, 2nd edn, Macmillan, London.
2. Lorrain, P., Corson, D. P. and Lorrain, F. (1988). *Electromagnetic Fields and Waves*, 3rd edn, Freeman, New York.
3. Cheo, P. K. (1985). *Fiber Optics, Devices and Systems*, Prentice-Hall, Englewood Cliffs, New Jersey.
4. Gloge, D. (1971). 'Weakly guiding fibres', *Applied Optics*, **10**, 2252–2258.

5. Ainslie, B. J. *et al.* (1982). 'Monomode fibre with ultralow loss and minimum dispersion at 1.55 μm', *Electronics Letters*, **18**, 843–844.
6. Personick, S. D. (1973). 'Receiver design for digital fiber optic communication systems, Parts I and II', *Bell System Tech. J.*, **52**, 843–886.

Chapter 3

1. Kressel, H. and Butler, J. K. (1977). *Semiconductor Lasers and Heterojunction LEDs*, Academic Press, New York.
2. Kressel, H. (Ed.) (1980). *Semiconductor Devices for Optical Communications, Vol. 39, Topics in Applied Physics*, Springer-Verlag, New York.
3. Casey, H. C. and Panish, M. B. (1978). *Heterostructure Lasers, Part A: Fundamental Principles* and *Part B: Materials and Operating Characteristics*, Academic Press, New York.
4. Yariv, A. (1991). *Optical Electronics*, 4th edn, HRW Ltd, Orlando, Florida.
5. Nayar, B. K. and Booth, R. C. (1986). 'An introduction to integrated optics', *British Telecom Technology Journal*, **3**, 5–15.
6. Brierley, M. C. and France, P. W. (1987). 'Neodynium doped fluorozirconate fibre laser', *Electronics Letters*, **23**, 815–817.

Chapter 4

1. McIntyre, R. J. (1966). 'Multiplication noise in uniform avalanche diodes', *IEEE Transactions: Electronic Devices*, **ED-13**, 164–168.
2. McIntyre, R. J. and Conradi, J. (1972). 'The distribution of gains in uniformly multiplying avalanche photodiodes', *IEEE Transactions: Electronic Devices*, **ED-19**, 713–718.
3. Stillman, G. E. *et al.* (1983). 'InGaAsP photodiodes', *IEEE Transactions: Electronic Devices*, **ED-30**, 364–381.
4. Kressel, H. (Ed.) (1980). *Semiconductor Devices for Optical Communications, Vol. 39, Topics in Applied Physics*, Springer-Verlag, New York.

Chapter 5

1. Personick, S. D. (1973). 'Receiver design for digital fiber optic communication systems, Parts I and II', *Bell System Tech. J.*, **52**, 843–886.
2. Smith, D. R. and Garrett, I. (1978). 'A simplified approach to digital receiver design', *Optical and Quantum Electronics*, **10**, 211–221.

Chapter 6

1. Hooper, R. C. *et al.* (1980). 'PIN–FET hybrid optical receivers for longer wavelength optical communications systems', in *Proceedings of the 6th European Conference on Optical Communications, York*, 222–225.

2. Smith, D. R. *et al.* (1980). 'PIN–FET hybrid optical receiver for 1.1–1.6 μm optical communication systems', *Electronics Letters*, **16**, 750–751.
3. Hullett, J. L., Muoi, T. V. and Moustakas, S. (1977). 'High-speed optical preamplifiers', *Electronics Letters*, **13**, 668–690.
4. Sibley, M. J. N., Unwin, R. T. and Smith, D. R. (1985). 'The design of PIN–bipolar transimpedance preamplifiers for optical receivers', *J. Inst. of Electrical and Electronic Engineers*, **55**, 104–110.
5. Moustakas, S. and Hullett, J. L. (1981). 'Noise modelling for broadband amplifier design', *IEE Proceedings Part G: Electronic Circuits and Systems*, **128**, 67–76.
6. Millman, J. (1979). *Microelectronics: Digital and Analog Circuits and Systems*, McGraw-Hill, New York, chapters 11–14.

Chapter 7

1. France, P. W. *et al.* (1987). 'Progress in fluoride fibres for optical telecommunications', *British Telecom Technology Journal*, **5**, 28–44.
2. Zakharov, V. E. and Shabat, A. V. (1972). 'Exact theory of 2-dimensional self focusing and 1-dimensional self modulation of nonlinear waves in nonlinear media', *Sov. Phys. JTEP*, **34**, 62–69.
3. Mellis, J. *et al.* (1988). 'Miniature packaged external-cavity semiconductor laser with 50 GHz continuous electrical tuning range', *Electronics Letters*, **24**, 988–989.
4. van der Ziel, J. P. *et al.* (1975). 'Laser oscillation from quantum states in very thin GaAs–$Al_{0.2}Ga_{0.8}As$ multilayer structures', *Applied Physics Letters*, **26**, 463–466.
5. Hansen, P. B. *et al.* (1993). '2 cm long monolithic multi-section laser for active mode-locking at 2.2 GHz', *Electronics Letters*, **29**, 739–741.
6. Yariv, A. (1991). *Optical Electronics*, 4th edn, HRW Ltd, Orlando, Florida.
7. Hodgkinson, T. G. *et al.* (1985). 'Coherent optical transmission systems', *British Telecom Technology Journal*, **3**, 5–18.
8. Walker, G. R. and Walker, N. G. (1988). 'A rugged all-fibre endless polarisation controller', *Electronics Letters*, **24**, 1353–1354.
9. Creaner, M. J. *et al.* (1988). 'Field demonstration of 565 Mbit/s DPSK coherent transmission system over 176 km of installed fibre', *Electronics Letters*, **24**, 1354–1356.
10. Brain, M. (1989). 'Coherent optical networks', *British Telecom Technology Journal*, **7**, 50–57.
11. Greaves, S. G. and Unwin, R. T. (1993). 'The variation of short gate-length GaAs MESFET intrinsic noise parameters with bias', *Microwave & Optical Technology Letters*, **6**, 892–895.

Useful reference books

Barnoski, M. K. (Ed.) (1981). *Fundamentals of Optical Fiber Communications*, Academic Press, New York.

Basch, E. E. (Ed.) (1987). *Optical-Fiber Transmission*, Sams, New York.

Kressel, H. (Ed.) (1980). *Semiconductor Devices for Optical Communications, Vol. 39, Topics in Applied Physics*, Springer-Verlag, New York.

Senior, J. (1985). *Optical Fiber Communications: Principles and Practice*, Prentice-Hall, Englewood Cliffs, New Jersey.

Yariv, A. (1991). *Optical Electronics*, 4th edn, HRW Ltd, Orlando, Florida.

Index