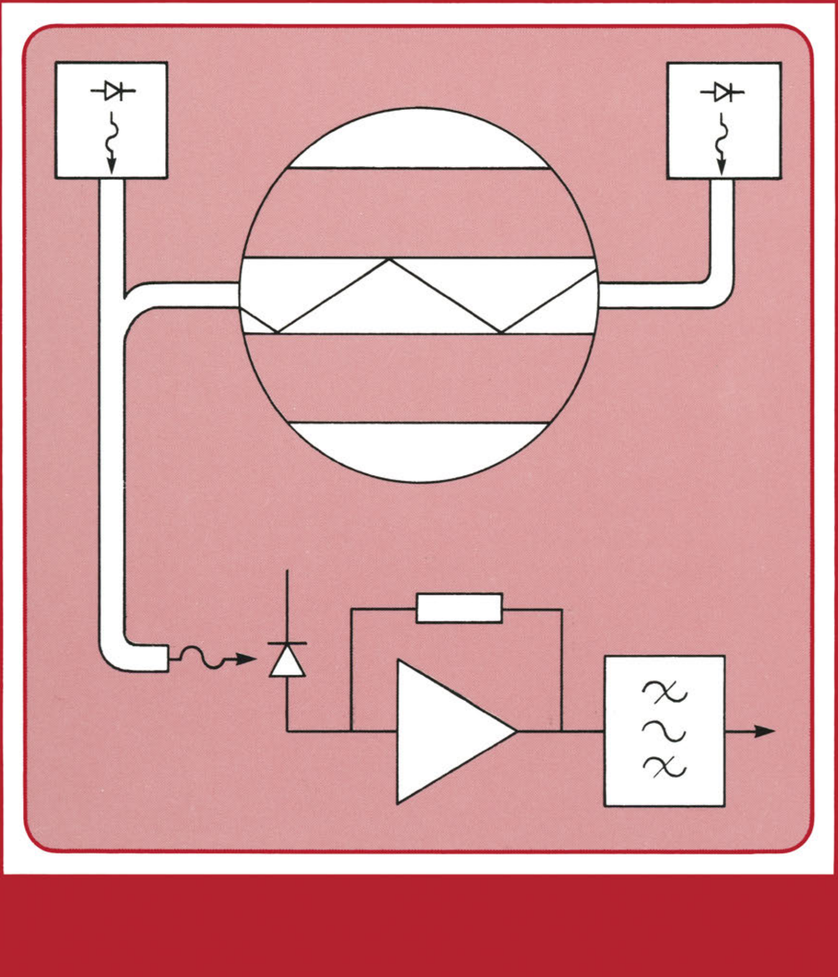
OPTICAL COMMUNICATIONS



M. J. N. SIBLEY

MACMILLAN NEW ELECTRONICS
INTRODUCTIONS TO ADVANCED TOPICS

Optical Communications

Macmillan New Electronics Series
Series Editor: Paul A. Lynn

Rodney F. W. Coates, *Underwater Acoustic Systems*
Paul A. Lynn, *Radar Systems*
A. F. Murray and H. M. Reekie, *Integrated Circuit Design*
Dennis N. Pim, *Television and Teletext*
M. J. N. Sibley, *Optical Communications*
Martin S. Smith, *Introduction to Antennas*
P. M. Taylor, *Robotic Control*

Optical Communications

M. J. N. Sibley

Department of Electrical and Electronic Engineering
The Polytechnic of Huddersfield

Macmillan New Electronics
Introductions to Advanced Topics

M
MACMILLAN

First published 1990

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

Typeset by TecSet Ltd, Wallington, Surrey.

British Library Cataloguing in Publication Data
Sibley, M. J. N.
Optical communications.
1. Optical communication systems
I. Title
621.38′0414

ISBN 978-0-333-47513-3 ISBN 978-1-349-20718-3 (eBook)
DOI 10.1007/978-1-349-20718-3

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 polytechnic 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

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 Mr K. Fullard of the Joint European Torus project at Culham, England, 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 Education 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.

List of Symbols

α	attenuation constant/absorption coefficient
α_{e}	electron ionisation coefficient
α_{h}	hole ionisation coefficient
$\boldsymbol{a}_x$, $\boldsymbol{a}_y$	unit vectors
A_0	total preamplifier voltage gain
$A(\omega)$	total voltage gain of receiver system
β	phase constant
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 following amplifier
C_{j}	junction capacitance
C_{s}	stray input capacitance
C_{T}	total receiver input capacitance
δ_n	refractive index change
δE_{c}	conduction band step
δE_{v}	valence band step
ϵ_0	permittivity of free-space (8.854×10^{-12} F/m)
ϵ_{r}	relative permittivity
D_{mat}	material dispersion coefficient
D_{wg}	waveguide dispersion coefficient
E_{g}	band-gap difference
$F(M)$	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

$h_p(t)$	input pulse shape
$H_{eq}(\omega)$	equalising network transfer function
$H_f(\omega)$	pre-detection filter transfer function
$H_{out}(\omega)$	Fourier transform (FT) of output pulse
$H_p(\omega)$	FT of received pulse
$H_T(\omega)$	normalised transimpedance
$<i_n^2>_0$	mean square (m.s.) noise current for logic 0 signal
$<i_n^2>_1$	m.s. noise current for logic 1 signal
$<i_n^2>_c$	m.s. equivalent input noise current of preamplifier
$<i_n^2>_{DB}$	m.s. photodiode bulk leakage noise current
$<i_n^2>_{DS}$	m.s. photodiode surface leakage noise current
$<i_n^2>_{pd}$	m.s. photodiode noise current
$<i_n^2>_Q$	quantum noise
$<i_n^2>_T$	total signal-independent m.s. noise current
$<i_s^2>$	m.s. photodiode signal current
$i_s(t)$	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
$<I_s>$	average signal current
$<I_s>_0$	average signal current for a logic 0
$<I_s>_1$	average signal current for a logic 1
I_{th}	threshold current
I_{DB}	photodiode bulk leakage
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)
λ	wavelength
L_n	diffusion length in p-type
m	modulation depth
M	multiplication factor
M_{opt}	optimum avalanche gain

η	quantum efficiency
n	refractive index
$<n^2>_{\mathrm{T}}$	total m.s. output noise voltage
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	average received power
P_{e}	total probability of error
q	electron charge (1.6×10^{-19} C)
r_{π}	base emitter resistance
$r_{\mathrm{bb'}}$	base-spreading resistance
r_{e}	reflection coefficient
R_1, R_2	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_0	photodiode responsivity
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	surface recombination velocity
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_{nr}	non-radiative recombination time
τ_{r}	radiative recombination time
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_0	impedance of free space
$Z_0(s)$	open loop transimpedance
$Z_{\mathrm{c}}(s)$	closed loop transimpedance
$Z_{\mathrm{f}}(s)$	feedback network transimpedance
Z_{in}	total input impedance
$Z_{\mathrm{T}}(\omega)$	transimpedance

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.

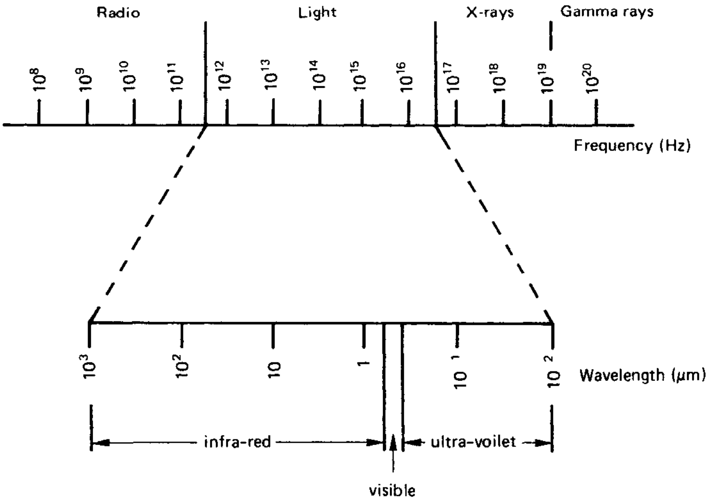


Figure 1.1 The electromagnetic spectrum

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 interfere with light-wave propagation in the atomosphere: 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 modern 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 easist 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 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.

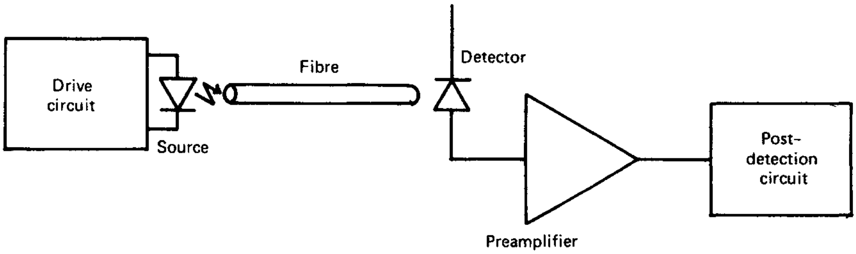


Figure 1.2 A basic optical comunications link

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 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 fibre lasers, 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

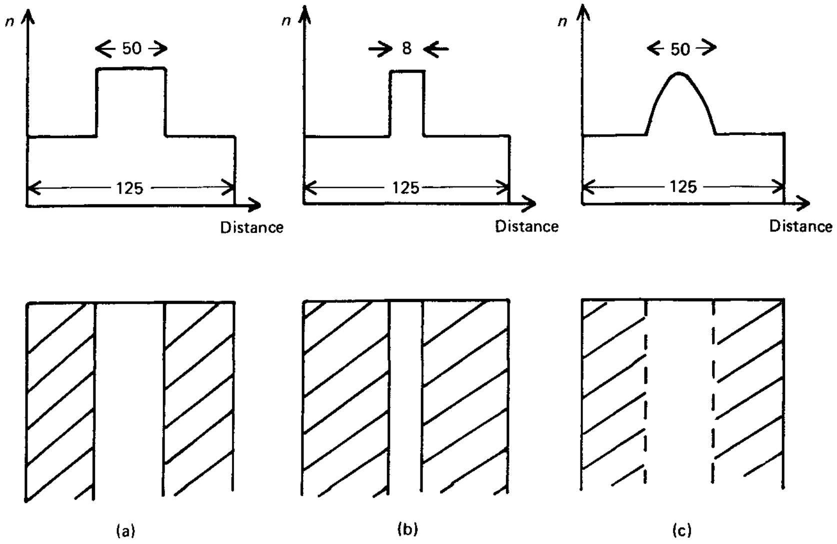


Figure 2.1 Typical refractive index profiles of (a) step-index multimode, (b) step-index single mode, and (c) graded-index multimode fibres (all dimensions are in μm)

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 500 μm for large-core MM fibres.

Optical fibres can be made from two different materials: silica glass (SiO_2) and plastic. The addition of certain dopants to the glass will vary the refractive index; 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 all-glass fibres and much cheaper to produce. However, the very high attenuation and low bandwidth of these fibres tend to limit their uses. 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 GHzkm, or MHzkm. Attenuation is dependent 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 the largest bandwidth.

In this chapter we shall examine the properties and design of optical fibres. We could study the propagation of light in optical fibres by solving Maxwell's equations applied to a cylindrical waveguide. However, this involves some rather complicated mathematics. Instead, we will solve Maxwell's equations in an infinite block of dielectric, and then apply them to a planar optical waveguide. When we come to examine optical fibres, we shall make use of the results from these analyses to draw some general conclusions; we will quote relevant formulae and important results directly. (The work presented here 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:

$$\nabla \times \boldsymbol{E} = -\mu\frac{\partial \boldsymbol{H}}{\partial t}\,(2.1\text{a}) \quad \text{and} \quad \nabla \times \boldsymbol{H} = \epsilon\frac{\partial \boldsymbol{E}}{\partial t} + \sigma\boldsymbol{E} \qquad (2.1\text{b})$$

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 $\boldsymbol{H}$. So

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

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

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

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

$$\frac{\partial^2 \boldsymbol{E}}{\partial z^2} = -\omega^2\mu\epsilon\boldsymbol{E} + \text{j}\omega\mu\sigma\boldsymbol{E} \qquad (2.3)$$

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

$$\frac{\partial^2 \boldsymbol{H}}{\partial z^2} = -\omega^2\mu\epsilon\boldsymbol{H} + \text{j}\omega\mu\sigma\boldsymbol{H} \qquad (2.4)$$

If we now let $\gamma^2 = -\omega^2\mu\epsilon + \text{j}\omega\mu\sigma$, the solutions to these equations are:

$$\boldsymbol{E} = E_{x\text{o}}\text{exp}(\text{j}\omega t)\text{exp}(-\gamma z) \qquad (2.5)$$

and

$$\boldsymbol{H} = H_{\text{yo}}\text{exp}(\text{j}\omega t)\ \text{exp}(-\gamma z) \qquad (2.6)$$

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 $\gamma = \alpha + j\beta$, where α and β are the *attenuation* and *phase constants* respectively, we get

$$\boldsymbol{E} = E_{xo}\exp(-\alpha z)\cos(\omega t - \beta z) \tag{2.7}$$

and

$$\boldsymbol{H} = H_{yo}\exp(-\alpha z)\cos(\omega t - \beta z) \tag{2.8}$$

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 will find it useful to derive some propagation parameters.

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

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

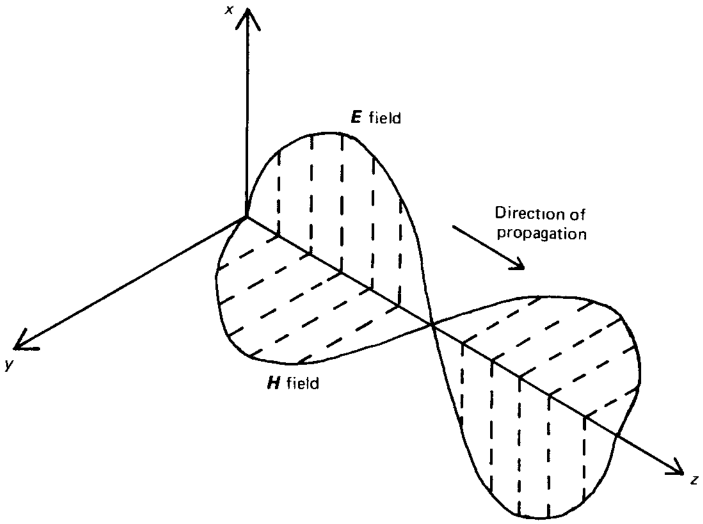


Figure 2.2 Variation of E and H for a TEM wave propagating in the z-direction

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 and the relative permeability is approximately unity, that is $\sigma << 0$, and $\mu_r \approx 1$. Over a distance of a few wavelengths, this results in $\alpha \approx 0$, and so we can write the E and H fields as

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

and

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

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

Figure 2.3 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

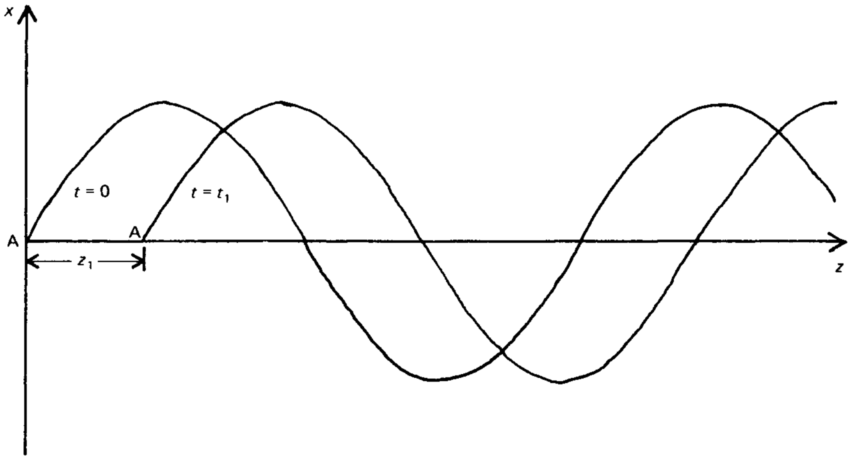


Figure 2.3 Illustrative of the phase velocity of a constant phase point, A, on the E field of a TEM wave

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

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:

$$\beta = \frac{2\pi}{\lambda} \tag{2.17}$$

(It is interesting to note that if ϵ_r, and hence n, varies with frequency, then β and v_p will also vary. Thus, if we have two light-waves of slightly different frequencies, then the two waves will travel at different velocities. 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

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

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 $\boldsymbol{H}$, then 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, $\boldsymbol{S}$, and its *instantaneous* value is given by

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

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

$$S_{av} = R_e\{\tfrac{1}{2}E \times H^*\} \tag{2.20}$$

where R_e denotes 'the real part of. .', $H^* = H\exp(-j[\,\omega t - \phi])$, 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 E and H. Thus the ray-path has the direction of power flow.)

2.1.3 *Material dispersion and group velocity*

As we have seen, the velocity of light in a dielectric depends upon the refractive index. However, the 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 the E field.

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

$$\begin{aligned} e_{AM} &= E_{x0}(1 + m\cos\omega_m t)\cos\omega_c t \\ &= E_{x0}\{\cos\omega_c t + \frac{m}{2}[\cos(\omega_c + \omega_m)t + \cos(\omega_c - \omega_m)t]\} \end{aligned} \tag{2.21}$$

where m is the depth of modulation. As $\beta = 2\pi/\lambda$, each frequency component will have its own value of β. If we take the variation of β with ω to be linear around the region of ω_c (that is, $d\beta/d\omega$ is a constant) then the change in β, $\delta\beta$, will be proportional to $\delta\omega$ where $\delta\omega = \omega_m$. Therefore, e_{AM} will be given by

$$\begin{aligned} e_{AM} = E_{x0}(\cos\omega_c t - \beta z) &+ E_{x0}\frac{m}{2}\cos[(\omega_c + \delta\omega)t - (\beta + \delta\beta)z] \\ &+ E_{x0}\frac{m}{2}\cos[(\omega_c - \delta\omega)t - (\beta - \delta\beta)z] \end{aligned} \tag{2.22}$$

Now, the sideband components can be written as

$$\boldsymbol{E}_{x0}\frac{m}{2}\cos[(\omega_c t - \beta z) + (\delta\omega t - \delta\beta z)] +$$

$$\boldsymbol{E}_{x0}\frac{m}{2}\cos[(\omega_c t - \beta z) - (\delta\omega t - \delta\beta z)]$$

$$= \boldsymbol{E}_{x0} m\cos(\omega_c t - \beta z)\cos(\delta\omega t - \delta\beta z) \tag{2.23}$$

and so the total wave can be given by

$$\boldsymbol{E}_{x0}(\cos\omega_c t - \beta z) + \boldsymbol{E}_{x0} m\cos(\omega_c t - \beta z)\cos(\delta\omega t - \delta\beta z) \tag{2.24}$$

From this equation we can see that the carrier wave travels at the phase velocity, while the modulation envelope travels at a velocity of $\delta\omega/\delta\beta$. This is the *group velocity*, v_g. Thus we can write

$$v_p = \frac{\omega}{\beta} \tag{2.25}$$

and

$$v_g = \frac{d\omega}{d\beta} \tag{2.26}$$

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.

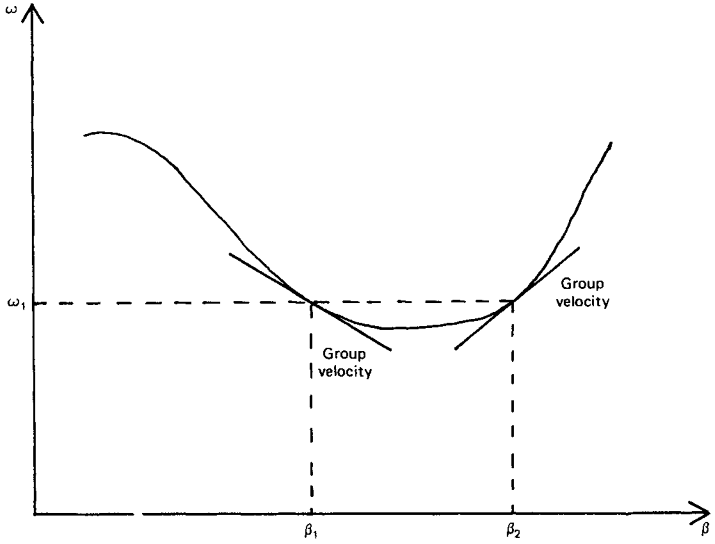


Figure 2.4 Showing the relationship between the phase and group velocity of a modulated TEM wave

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, and this will affect 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. Now, we can express $\delta\tau$ as

$$\delta\tau = \frac{\mathrm{d}\tau}{\mathrm{d}\lambda} \times \delta\tau \tag{2.27}$$

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

$$D_{\mathrm{mat}} = \frac{\mathrm{d}\tau}{\mathrm{d}\lambda} = \frac{\mathrm{d}}{\mathrm{d}\lambda} \times \frac{1}{v_g} \tag{2.28}$$

Now

$$\frac{1}{v_g} = \frac{\mathrm{d}\beta}{\mathrm{d}\omega} = \frac{\mathrm{d}\lambda}{\mathrm{d}\omega} \times \frac{\mathrm{d}\beta}{\mathrm{d}\lambda} \quad = -\frac{\lambda^2}{2\pi c} \times \frac{\mathrm{d}\beta}{\mathrm{d}\lambda} = -\frac{\lambda^2}{2\pi c} \times \frac{\mathrm{d}}{\mathrm{d}\lambda} \times \frac{\omega n}{c}$$

$$= -\frac{\lambda^2}{2\pi c} \times \frac{\mathrm{d}}{\mathrm{d}\lambda} \times \frac{2\pi c n}{\lambda c} = -\frac{\lambda^2}{c} \times \frac{\mathrm{d}}{\mathrm{d}\lambda} \times \frac{n}{\lambda}$$

$$= \frac{1}{c} \quad \left[n - \lambda \frac{dn}{d\lambda}\right] \frac{N_g}{c} \tag{2.29}$$

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

$$D_{\mathrm{mat}} = \frac{\mathrm{d}\tau}{\mathrm{d}\lambda} = \frac{\mathrm{d}}{\mathrm{d}\lambda} \times \frac{N_g}{c} = \frac{1}{c} \left[\frac{\mathrm{d}n}{\mathrm{d}\lambda} - \tau \frac{\mathrm{d}^2 n}{\mathrm{d}\lambda^2} - \frac{\mathrm{d}n}{\mathrm{d}\lambda}\right] = -\frac{\lambda}{c} \times \frac{\mathrm{d}^2 n}{\mathrm{d}\lambda^2} \tag{2.30}$$

(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 fibre). 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}.

As an example, if we consider amplitude modulation of a 600 nm wavelength light source by a sinewave of frequency 100 MHz, then $\delta\lambda = 2.4 \times 10^{-4}$ nm and, if $D_{mat} = 500$ ps/nm/km, then $\sigma_{mat} = 0.12$ ps/km. Thus the difference in arrival time between the sidebands is insignificant. However, this assumes that the spectral width, or *line-width*, of the source is zero, that is, the light emission is a single sine-wave of frequency 5×10^{14} Hz. In practice, this is not the case. If the line-width of the source is 10 nm (that is, the output consists of a range of frequencies from 4.96×10^{14} to 5.04×10^{14} Hz) then $\delta\lambda \approx 10$ nm, and $\sigma_{mat} \approx 5$ ns/km. From these calculations, we can see that the spectral purity of the source can dominate σ_{mat}, if the source line-width is large. So, for high data-rate or long-haul applications, it is important to use narrow line-width 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 later, in which the zero dispersion point is at 1.55 μm — a transmission window which offers lower attenuation.

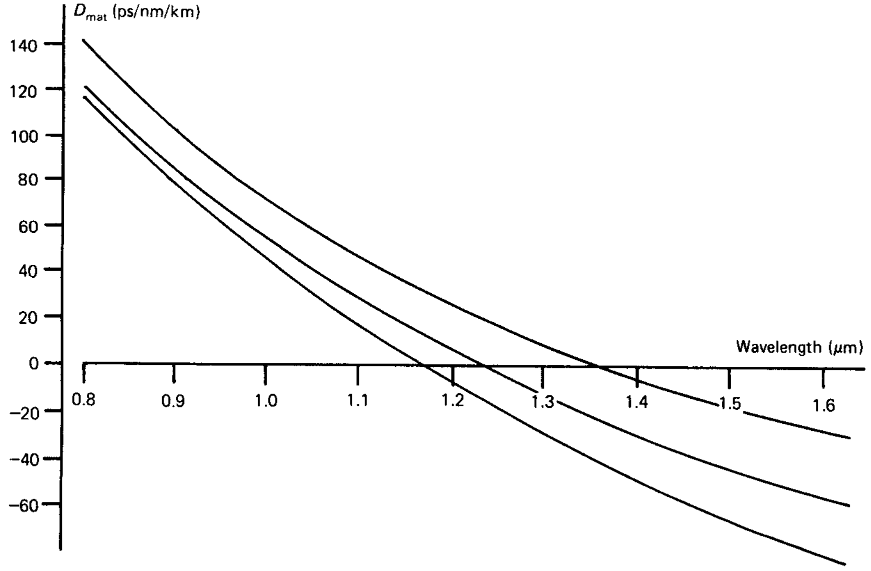


Figure 2.5 Variation of material dispersion, D_{mat}, with wavelength for three different glass fibres

We have seen that the compositon of the fibre causes dispersion of the signal. 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.

2.2.1 Reflection and refraction at boundaries

Figure 2.6 shows a transverse electric, *TE*, wave, $\boldsymbol{E}_i$ and $\boldsymbol{H}_i$, incident on a boundary between two dissimilar, non-conducting, dielectrics (the wave-

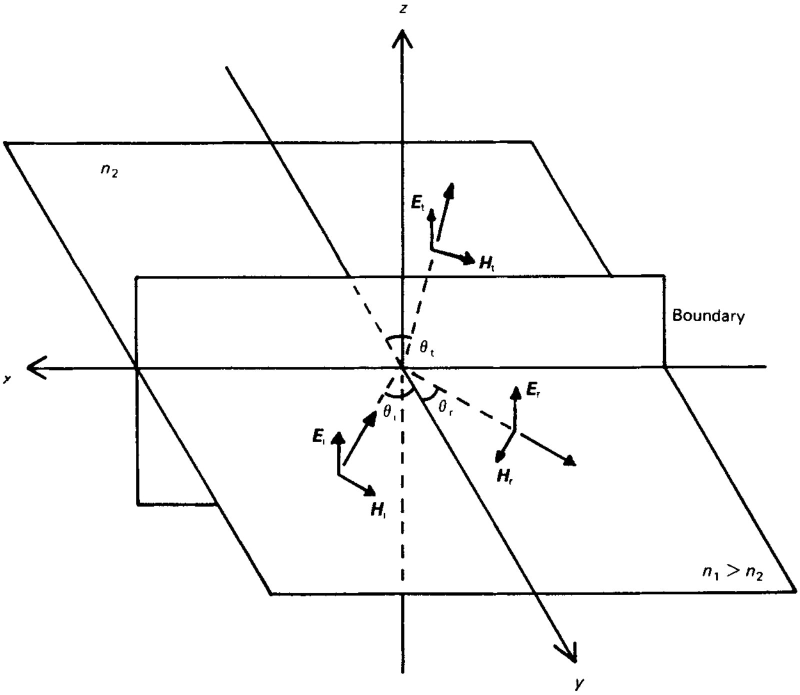


Figure 2.6 Reflection and refraction of a TEM wave, at the boundary between two dielectric materials

guide boundary). As can be seen, some of the wave undergoes reflection, $\boldsymbol{E}_r$ and $\boldsymbol{H}_r$, while the rest is transmitted (or *refracted*), $\boldsymbol{E}_t$ and $\boldsymbol{H}_t$. In order to determine the optical power in both waves, we initially resolve the waves into their x, y, and z components. We can express the incident E field as the combination of a field propagating in the negative x-direction, and another field travelling in the negative y-direction. Thus, $\boldsymbol{E}_i$ can be written as

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

(Here x and y refer to the distances travelled along the respective axes.)

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

$$E_i + E_r = E_t \tag{2.32}$$

Dividing by E_i yields

$$1 + r_e = t_e \tag{2.33}$$

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

$$\boldsymbol{E}_r = \boldsymbol{a}_z r_e E_0 \exp(j\beta_1[x\sin\theta_r - y\cos\theta_r]) \tag{2.34}$$

and

$$\boldsymbol{E}_t = \boldsymbol{a}_z t_e E_0 \exp(j\beta_2[x\sin\theta_t + y\cos\theta_t]) \tag{2.35}$$

We can substitute the expressions for the E fields at the interface, $y = 0$, back into (2.32) to give

$$E_0\exp(j\beta_1 x\sin\theta_i) + r_e E_0 \exp(j\beta_1 x\sin\theta_r)$$
$$= t_e E_0 \exp(j\beta_2 x\sin\theta_t) \tag{2.36}$$

In order to satisfy (2.33), the exponential terms in (2.36) 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

$$\theta_i = \theta_r \tag{2.37}$$

which is *Snell's law of reflection*, while the second gives

$$\sin\theta_t = \frac{\beta_1}{\beta_2}\sin\theta_i = \frac{n_1}{n_2}\sin\theta_i \tag{2.38}$$

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

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

$$\begin{aligned} \boldsymbol{H}_i &= (-\boldsymbol{a}_x\cos\theta_i + \boldsymbol{a}_y\sin\theta_i)\frac{E_0}{Z_1}\exp(j\beta_1[x\sin\theta_i + y\cos\theta_i]) \\ &= (-\boldsymbol{a}_x\cos\theta_i + \boldsymbol{a}_y\sin\theta_i)\frac{E_i}{Z_1} \end{aligned} \tag{2.39}$$

$$\boldsymbol{H}_r = (\boldsymbol{a}_x\cos\theta_r + \boldsymbol{a}_y\sin\theta_r)\frac{E_r}{Z_1} \tag{2.40}$$

$$\boldsymbol{H}_t = (-\boldsymbol{a}_x\cos\theta_t + \boldsymbol{a}_y\sin\theta_t)\frac{E_t}{Z_2} \tag{2.41}$$

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

$$\begin{aligned} &-\frac{E_0}{Z_1}\cos\theta_i\exp(j\beta_1 x\sin\theta_i) + r_e\frac{E_0}{Z_1}\cos\theta_r\exp(j\beta_1 x\sin\theta_r) \\ &= -t_e\frac{E_0}{Z_2}\cos\theta_t\exp(j\beta_2 x\sin\theta_t) \end{aligned} \tag{2.42}$$

where we have substituted for E_r and E_t. Because the exponential terms are all equal, (2.42) becomes

$$\frac{\cos\theta_i}{Z_1}(1 - r_e) = t_e\frac{\cos\theta_t}{Z_2} \tag{2.43}$$

Since $1 + r_e = t_e$, we can eliminate t_e from (2.43) to give

$$r_e = \frac{Z_2\cos\theta_i - Z_1\cos\theta_t}{Z_2\cos\theta_i + Z_1\cos\theta_t} = \frac{n_1\cos\theta_i - n_2\cos\theta_t}{n_1\cos\theta_i + n_2\cos\theta_t} \tag{2.44}$$

and, by using Snell's Law, we can eliminate θ_t from (2.44). Therefore

$$r_e = \frac{\cos\theta_i - \surd[(n_2/n_1)^2 - \sin^2\theta_i]}{\cos\theta_i + \surd[(n_2/n_1)^2 - \sin^2\theta_i]} \tag{2.45}$$

Close examination of (2.45) reveals that r_e is unity if the term under the square root is zero, that is, if $\sin^2\theta_i = (n_2/n_1)^2$. 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

$$\sin^2\theta_c = \frac{n_2^2}{n_1^2} \text{ or, } \sin\theta_c = \frac{n_2}{n_1} \tag{2.46}$$

Substitution of this result into Snell's Law gives the angle of refraction to be 90°, and so the 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. This implies that total internal reflection takes place. However, there will still 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 (2.47):

$$\boldsymbol{E}_t = \boldsymbol{a}_z t_e E_0 \exp(j\beta_2[x\sin\theta_t + y\cos\theta_t]) \tag{2.47}$$

To evaluate $\boldsymbol{E}_t$ we require to find $\sin\theta_t$ and $\cos\theta_t$. If $\theta_i > \theta_c$, then $\sin\theta_i > n_2/n_1$. If we substitute this into Snell's Law, then 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

$$\cos\theta_t = \surd(1 - \sin^2\theta_t) = jA \tag{2.48}$$

where A is a *real* number given by

$$A = \surd[(n_1/n_2)^2\sin^2\theta_i - 1] \tag{2.49}$$

Thus the refracted wave can be written as

$$\begin{aligned}\boldsymbol{E}_t &= \boldsymbol{a}_z t_e E_0 \exp(j\beta_2[x\sin\theta_t + yjA]) \\ &= \boldsymbol{a}_z t_e E_0 \exp(-\beta_2 Ay)\exp(j\beta_2 x\sin\theta_t)\end{aligned} \tag{2.50}$$

So, although total internal reflection takes place, an E field exists in the lower refractive index material. This field propagates without loss in the negative x-direction, but undergoes attenuation as $\exp(-\beta_2 Ay)$ 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 (2.41):

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

Substitution for $\cos\theta_t$ yields

$$\boldsymbol{H}_t = (-\boldsymbol{a}_x jA + \boldsymbol{a}_y \sin\theta_t)\frac{E_t}{Z_z} \tag{2.51}$$

The presence of j in the x-component of $\boldsymbol{H}_t$ shows that it is 90° out of phase, time-wise, with the E field. We can now find the power transmitted across the interface. From section 2.1.2, the average Poynting vector is the real part of

$$\boldsymbol{S}_{av} = \frac{1}{2}\boldsymbol{E} \times \boldsymbol{H}^* \tag{2.52}$$

where $\boldsymbol{H}^* = \boldsymbol{H}\exp(-j[\omega t - \phi])$, and ϕ is the temporal phase angle between the individual E and H field components. So, with the fields given by (2.50) and (2.51), we have

$$\begin{aligned}\boldsymbol{S}_{av} &= -\,\boldsymbol{a}_x\frac{1}{2} \times \frac{E_t^2}{Z_2}\sin\theta_t - \boldsymbol{a}_y\frac{1}{2} \times \frac{AE_t^2}{Z_2}\exp(j\pi/2)\\ &= -\boldsymbol{a}_x\frac{1}{2} \times \frac{E_t^2}{Z_2}\sin\theta_t - \boldsymbol{a}_y j\frac{1}{2} \times \frac{AE_t^2}{Z_2}\end{aligned} \tag{2.53}$$

Hence the real part of the transmitted power is

$$\boldsymbol{S}_{av} = \boldsymbol{a}_x\frac{1}{2} \times \frac{E_t^2}{Z_2}\sin\theta_t \tag{2.54}$$

Thus although total internal reflection takes place, there is still a TEM wave propagating along the boundary — the *evanescent* wave.

As an example, consider a TE wave with a free-space wavelength of 600 nm, propagating in a dielectric of refractive index 1.5, incident on a boundary with a second dielectric of refractive index 1.4. We will take the angle of incidence to be 75° to a normal drawn perpendicular to the dielectric boundary. With these parameters, $\theta_c = 69°$, $\beta_1 = 1.57 \times 10^7\ \text{m}^{-1}$, $\beta_2 = 1.47 \times 10^7\ \text{m}^{-1}$, $Z_1 = 251\ \Omega$, $Z_2 = 269\ \Omega$, $\sin\theta_t = 1.04$ and $\cos\theta_t = j0.27$. Now, r_e will be

$$\begin{aligned}r_e &= \frac{0.27 - \sqrt{(0.87 - 0.93)}}{0.27 + \sqrt{(0.87 - 0.93)}}\\ &= \frac{0.27 - j0.25}{0.27 + j0.25} = 1\angle{-2\phi}\end{aligned} \tag{2.55}$$

where $\phi = \tan^{-1}(0.25/0.27) = 0.75$ rads. (This angle is the spatial phase-change experienced by the E field on reflection.) Since $t_e = 1 + r_e$:

$$t_e = \frac{2 \times 0.27}{0.27 + j0.25} = 0.54\angle{-\phi} \tag{2.56}$$

Thus, $\boldsymbol{E}_{\rm t}$ (equation 2.50) will be

$$\boldsymbol{E}_{\rm t} = \boldsymbol{a}_z 0.54 E_0 \exp(-3.93 \times 10^6\, y)\exp(\mathrm{j}1.57 \times 10^7 x) \underline{| -\phi} \tag{2.57}$$

Hence the average transmitted power is

$$S_{\rm av} = 5.4 \times 10^{-4} E_0{}^2 \exp(-7.86 \times 10^6 y) \tag{2.58}$$

By following a similar procedure, it can be shown that the modulus of the incident and reflected optical powers are equal. (We should expect this because the magnitude of $r_{\rm e}$ is unity, that is, total internal reflection occurs.) So, we can say that the evanescent wave couples with, but takes no power from, the waves travelling in the dielectric. It can, however, deliver power and most SM couplers rely on this property.

It is interesting to calculate the attenuation that the evanescent wave experiences away from the boundary. If we take y equal to one wavelength in the second dielectric, then the power at this distance is

$$\begin{aligned} S_{\rm av} &= 5.4 \times 10^{-4} E_0{}^2 \exp(-3.38) \text{ for } y = 430 \text{ nm} \\ &= 18.4 \times 10^{-6} E_0{}^2 \end{aligned} \tag{2.59}$$

So, the power in the surface wave is 5.4×10^4 times less than the power at the interface — an attenuation of roughly 15 dB.

2.2.2 Propagation modes

In the previous preview 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 $\theta_{\rm c}$, then total internal reflection would occur. It might be thought that any ray satisfying this requirement must propagate without loss. However, if there are two rays propagating, then the E fields will interfere with each other and, if the waves are out-of-phase, they will cancel each other out. In order to determine the criteria for successful propagation, let us consider two rays, of the same amplitude and frequency, propagating in a dielectric slab.

Figure 2.7 shows the situation we will analyse. As can be seen, the E field components of ray1 and ray2 are in-phase at A. This means that the interference is constructive, and so the rays should propagate. However, at point B the ray1 wavefront has undergone a phase change, because of travelling distance b and reflection off the boundary, while the wavefront due to ray2 has undergone a phase change corresponding to the distance a. In order to maintain constructive interference at point B, the difference in

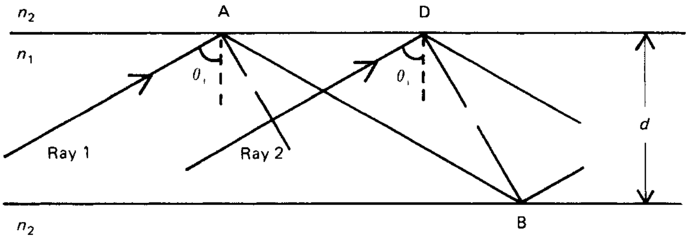


Figure 2.7 Illustrative of the requirement for successful propagation of two TEM waves in a planar optical waveguide (CD = a, AB = b)

phase between the two rays should be zero, or differ by an integral number of cycles.

We can find the phase change due to reflection off the boundary, from the reflection coefficient. From (2.44), r_e is

$$r_e = \frac{\cos\theta_i - \sqrt{(n_2/n_1)^2 - \sin^2\theta_i]}}{\cos\theta_i + \sqrt{(n_2/n_1)^2 - \sin^2\theta_i}} \tag{2.60}$$

$$\text{or, } r_e = 1 \underline{|-2\phi} \text{ where } \phi = \tan^{-1}\frac{\sqrt{[(n_1^2\sin^2\theta_i - n_2^2)]}}{n_1\cos\theta_i} \tag{2.61}$$

Therefore, successful propagation occurs provided

$$\frac{2\pi n_1}{\lambda_0} \times b - 2\phi = \frac{2\pi n_1}{\lambda_0} \times a + 2\pi N \tag{2.62}$$

where 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

$$b = \frac{d}{\cos\theta_i}, \text{ and } a = d\sin\theta_i\left[\tan\theta_i - \frac{1}{\tan\theta_i}\right]$$

$$= \frac{d(1 - 2\cos^2\theta_i)}{\cos\theta_i}$$

Therefore, (2.62) becomes

$$\frac{4\pi n_1}{\lambda_0} d\cos\theta_i - 2\phi = 2\pi N$$

or, by substituting for ϕ

$$\tan\left[\frac{2\pi n_1}{\lambda_0} \times d\cos\theta_i - N\pi\right] = \frac{\surd[n_1^2\sin^2\theta_i - n_2^2}{n_1\cos\theta_i} \qquad (2.63)$$

In (2.63) we have the x-axis component of β in the waveguide (note that $\beta_1 = 2\pi n_1/\lambda_0$) and so we can write

$$\tan(\beta_{1x}d - N\pi) = \frac{2\pi\surd[n_1^2\sin^2\theta_i - n_2^2]}{\beta_{1x}\lambda_0} \qquad (2.64)$$

where

$$\beta_{1x} = \frac{2\pi n_1}{\lambda_0} \times \cos\theta_i \qquad (2.65)$$

From our previous discussions, the evanescent field undergoes attenuation as $\exp(-\alpha_2 y)$. If we had solved Maxwell's equations in the waveguide, then we would have found that

$$\alpha_2 = \frac{2\pi\surd[n_1^2\sin^2\theta_i - n_2^2]}{\lambda_0} \qquad (2.66)$$

Therefore we can write (2.63) as

$$\tan(\beta_{1x}d - N\pi) = \frac{\alpha_2}{\beta_{1x}} \qquad (2.67)$$

Both (2.63) and (2.67) are known as *eigenvalue* equations. Solution of (2.67) will yield the values of β_{1x}, the *eigenvalues*, for which light rays will propagate, while solution of (2.63) will yield the permitted values of θ_i. Unfortunately these equations can only be solved by graphical or numerical methods. As an example, consider 1.3 μm wavelength light propagating in a planar waveguide of width 200 μm and depth 10 μm, with core and cladding refractive indices of 1.46 and 1.44 respectively. We can find the angle of incidence for each mode by substituting these parameters into (2.63), 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 (2.63), for varying angles of incidence, θ_i.

This graph shows that approximate values of θ_i are 89°, 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 88.83°, 86.48°, 84.15° and 81.88°.

We can estimate the number of propagating 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 (2.63), to give

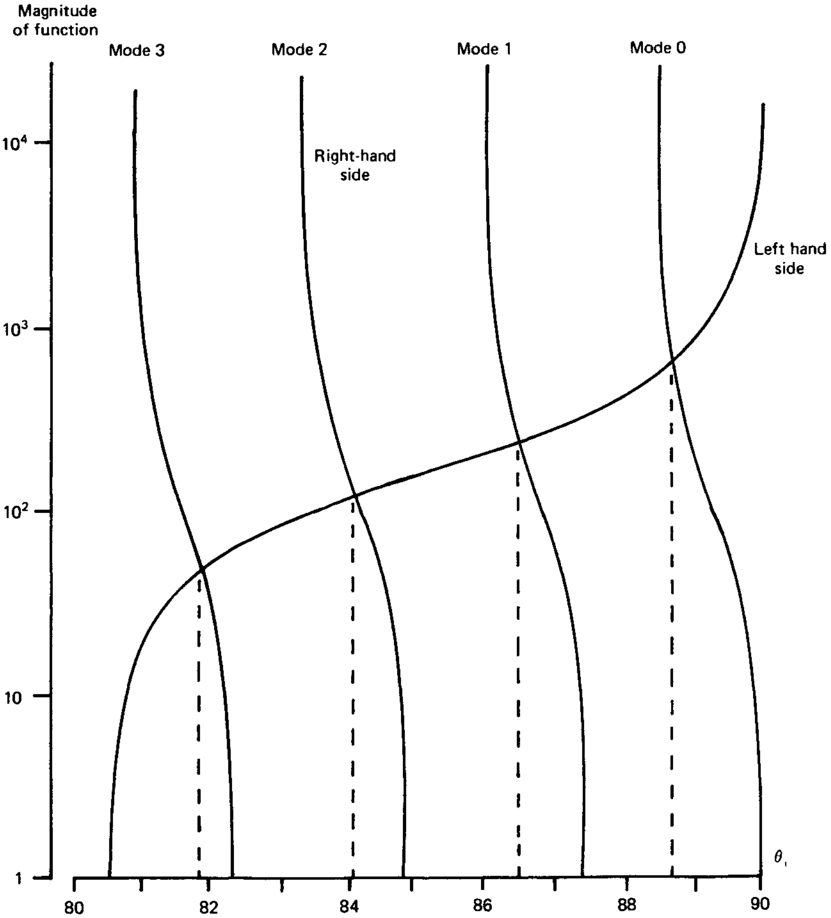


Figure 2.8 Eigenvalue graphs for a planar dielectric waveguide

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

or

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

If we define a normalised frequency variable, V, as

$$V = \frac{2\pi d(n_1{}^2 - n_2{}^2)^{\frac{1}{2}}}{\lambda_0} \tag{2.68}$$

then the maximum number of modes will be

$$N_{\max} = \frac{V}{\pi} = \frac{2d(n_1^2 - n_2^2)^{\frac{1}{2}}}{\lambda_0} \tag{2.69}$$

Equation (2.69) 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 $V = 11.6$ and so the number of propagating modes is 4. Note that we can find the maximum propagating frequency, or wavelength, from V.

It should be apparent that we can reduce the number of propagating modes by decreasing the waveguide thickness. In particular, if $V \leqslant \pi$ 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.

2.2.3 Modal dispersion

We have just 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

$$\delta t = t_{\max} - t_{\min} \tag{2.70}$$

where $t_{\max}$ and $t_{\min}$ are the propagation times of the highest and lowest order modes, respectively. We can find these times 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

$$\delta t = \frac{LN_{g1}}{cN_{g2}}(N_{g1} - N_{g2}) \tag{2.71}$$

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

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

$$= \frac{N_{g1}}{cn_2}\delta n$$

or

$$\sigma_{\mathrm{mod}} = \frac{N_{g1}}{cn_2}\delta n \tag{2.72}$$

where δn is the refractive index difference, and σ_{mod} is the dispersion per unit length.

As an example, if we consider 850 nm wavelength light propagating through 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 $N_{g2} = 1.53$, then $\sigma_{\mathrm{mod}} = 0.39$ ns/km. If we use modal analysis, then there are thirteen modes and the angles of incidence for the zero and thirteenth-order modes are 89.21° and 69.96° respectively. Thus σ_{mod} using this method is 0.35 ns/km, and the error in using (2.72) is small. If the waveguide is single-mode, then σ_{mod} reduces to zero.

2.2.4 Waveguide dispersion

As well as suffering from modal and material dispersion, a propagating signal will also undergo *waveguide dispersion*. In common with material dispersion, waveguide dispersion results from the variation of the group velocity with wavelength.

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

$$\tau = \frac{\mathrm{d}\beta_{1x}}{\mathrm{d}\omega} = -\frac{\lambda_0^2}{2\pi c} \times \frac{\mathrm{d}}{d\lambda}\beta_{1x}$$

$$= -\frac{\lambda_0^2}{2\pi c} \times \frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\frac{2\pi n_1 \cos\theta_i}{\lambda_0}\right]$$

$$= \frac{\cos\theta_i}{c}\left[n_1 - \lambda_0\frac{\mathrm{d}n_1}{\mathrm{d}\lambda}\right] - \frac{n_1\lambda_0}{c} \times \frac{\mathrm{d}}{\mathrm{d}\lambda}\cos\theta_i \tag{2.73}$$

The first term in (2.73) is simply the material dispersion resolved onto the x-axis. However, the second term is the waveguide dispersion, σ_{wg}. This is due to the angle of incidence for each mode, and hence the propagation time, varying with wavelength. In multi-mode (MM) waveguides, this source of dispersion is insignificant in comparison with the modal dispersion. However with SM waveguides, the waveguide dispersion can be comparable to the material dispersion. (We will be returning to this in section 2.3.2.)

2.2.5 *Numerical aperture*

Figure 2.9 shows two light rays entering a planar waveguide. Refraction of both rays occurs on entry; however ray1 fails to propagate in the guide 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

$$n_0 \sin\theta_1 = n_1 \sin(90^\circ - \theta_c) = n_1 \cos\theta_c$$
$$= n_1 \left[1 - \frac{n_2^2}{n_1^2}\right]^{1/2}$$

Therefore

$$\sin\theta_1 = \frac{1}{n_0}(n_1^2 - n_2^2)^{\frac{1}{2}} \tag{2.74}$$

If the guide is in air, then

$$NA = \sin\theta_1 = (n_1^2 - n_2^2)^{1/2} \tag{2.75}$$

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

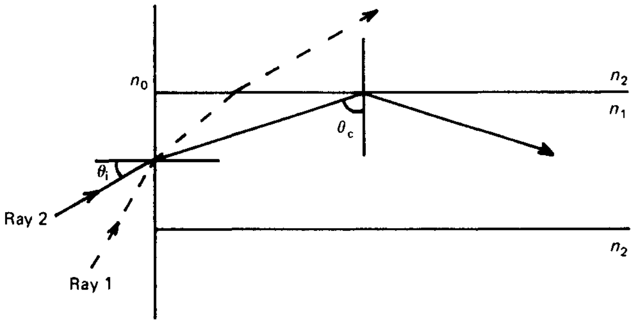


Figure 2.9 Construction for the determination of the numerical aperture

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. The solution of Maxwell's equations applied to an optical fibre involve us in some rather tedious mathematics. Instead, we shall confine ourselves to a general discussion, using the points raised from our examination of the planar waveguide. Important results and formulae will be quoted wherever necessary.

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 2.10a 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 2.10b shows, these rays do not fully utilise the area of the core. As skew rays travel significantly further than meridional rays, they generally undergo higher attenuation.

2.3.1 Step-index multimode fibre

In step-index, MM fibres, light travels down the fibre core in a series of modes which can be either transverse electric, TE, or transverse magnetic, TH. In addition, there can also be modes which have neither *E* or *H*

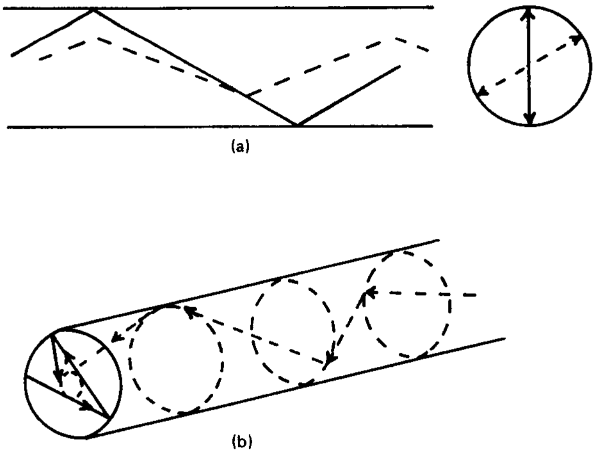


Figure 2.10 Propagation of (a) meridional, and (b) skew rays in the core of a step-index, MM, optical fibre

transverse to the fibre axis. Those are known as *hybrid* modes, designated as EH or HE depending on the relative magnitude of the E and H field components transverse to the fibre axis.

In common with the planar waveguide, the solution of the cylindrical co-ordinate form of Maxwell's equations results in an eigenvalue equation which yields the possible waveguide modes. If $(n_1 - n_2)/n_2 \leqslant 1$, then the fibre is known as *weakly guiding* [4]. With this restriction, the eigenvalue equation for a fibre with radius a, reduces to

$$\frac{J_{\mathrm{v}-1}(ua)}{J_{\mathrm{v}}(ua)} = -\frac{w}{u} \times \frac{K_{\mathrm{v}-1}(wa)}{K_{\mathrm{v}}(wa)} \tag{2.76}$$

where $J_{\mathrm{v}}(ua)$ and $K_{\mathrm{v}}(wa)$ are Bessel functions of the first and second kind (both widely tabulated), and v is the Bessel function order. The parameters u and w are defined by

$$u^2 = \left[\frac{2\pi n_1}{\lambda_0}\right]^2 - \beta^2 \text{ and } w^2 = \beta^2 - \left[\frac{2\pi n_2}{\lambda_0}\right]^2 \tag{2.77}$$

where β is the phase constant *along the fibre axis*. If we define a normalised frequency variable as

$$V = \frac{2\pi a(n_1{}^2 - n_2{}^2)^{1/2}}{\lambda_0} \tag{2.78}$$

then it is easy to show that

$$V^2 = (ua)^2 + (wa)^2 \tag{2.79}$$

As with the eigenvalue equation we encountered in section 2.2.2, the solution of equation (2.76) involves graphical techniques. The solutions can be obtained by plotting a graph of the left- and right-hand sides of (2.76) against ua, and then finding the points of intersection – the solutions to (2.76). 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 TE_{vm}.

Figure 2.11 shows a plot of both sides of (2.76), for a fibre with a normalised frequency of 12.5. The curves that follow a tangent function are the left-hand side of (2.76). 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 $K_{-1}(x) = K_1(x)$.) An interesting feature of these plots is that there are no eigenvalues for $ua > V$. Thus V is known as the *normalised cut-off frequency*. From figure 2.11, we can see that, provided the argument ua is less than V, the number of eigenvalues, that is, the number of modes, is one greater than the number of zeros for the particular Bessel function order.

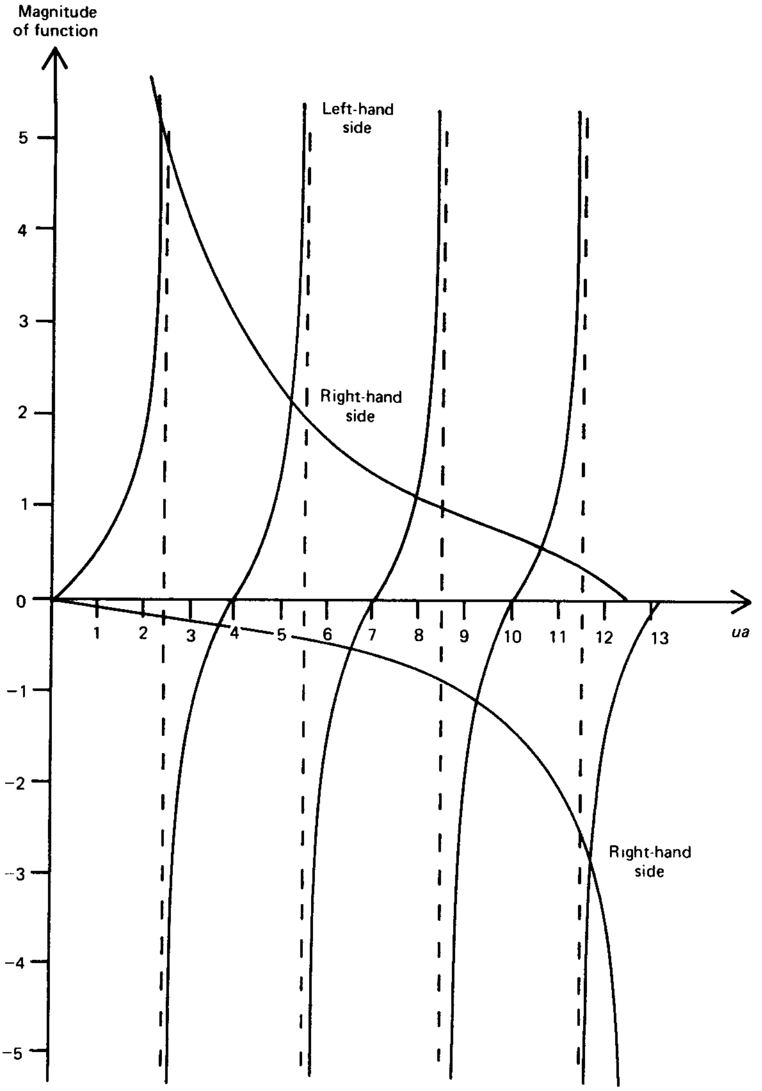


Figure 2.11 Eigenvalue graphs for the zero- and first order modes in a cylindrical waveguide

Thus for the zero-order function, the number of zeros with $ua < V$ is 4, and so the number of modes is 5. For $\nu = 1$, 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 (2.75). 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

$$\Omega = \pi\Theta_i^2 \tag{2.80}$$

Now, if Θ_i is small, then $\Theta_i \approx \sin\Theta_i$ — the numerical aperture. So, Ω will be given by

$$\Omega = \pi NA^2 = \pi(n_1^2 - n_2^2) \tag{2.81}$$

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

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

or

$$N_{\text{max}} = \frac{V^2}{2} \tag{2.82}$$

For a large number of modes, we can approximate σ_{mod} by the expression derived for the planar waveguide, equation (2.72), with d replacing a. So, if we consider a 125 μm radius, PCS fibre with $n_1 = 1.5$ and $n_2 = 1.4$, then the number of 850 nm wavelength modes is about 2×10^5. If we take N_{g1} to be 1.64, then σ_{mod} will be 390 ns/km, which will tend to dominate the dispersion characteristic. Because of this, the bandwidth–length product for step-index, MM fibres varies from less than 1 MHz km to 100 MHz km.

2.3.2 Step-index single-mode fibre

In the previous section, we saw that the propagating modes had to satisfy the following eigenvalue equation:

$$\frac{J_{v-1}(ua)}{J_v(ua)} = -\frac{w}{u} \times \frac{K_{v-1}(wa)}{K_v(wa)} \tag{2.83}$$

In a SM fibre, only the lowest order mode can propagate. This corresponds to $v = 0$ and so (2.83) becomes

$$\frac{J_1(ua)}{J_0(ua)} = \frac{w}{u} \times \frac{K_1(wa)}{K_0(wa)} \tag{2.84}$$

As we saw earlier, there will be m possible modes for $\nu = 0$. So, for SM operation, m must be equal to one 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 2.11 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 2.78) then the condition for SM operation is

$$\lambda_0 > 2.6a(n_1^2 - n_2^2)^{\frac{1}{2}} \tag{2.85}$$

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.

The dispersion in SM fibres is due to σ_{mat} and σ_{wg}. We briefly considered waveguide dispersion when we examined the planar waveguide. After some rather complex mathematics, we can approximate the waveguide dispersion coefficient, D_{wg}, by

$$D_{\mathrm{wg}} = \frac{2N_{g2}\lambda_0}{(2\pi a)^2 2cn_2^2} \tag{2.86}$$

In common with the material dispersion, the units of D_{wg} are usually ns/nm/km, and so we can reduce σ_{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 1.3 μm. Thus the material and waveguide dispersion will cancel each other out at a certain wavelength. Figure 2.12 shows the theoretical variation with wavelength of D_{wg}, D_{mat} and total dispersion ($D_{\mathrm{mat}} + D_{\mathrm{wg}}$) 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 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.

2.3.3 *Graded-index fibres*

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

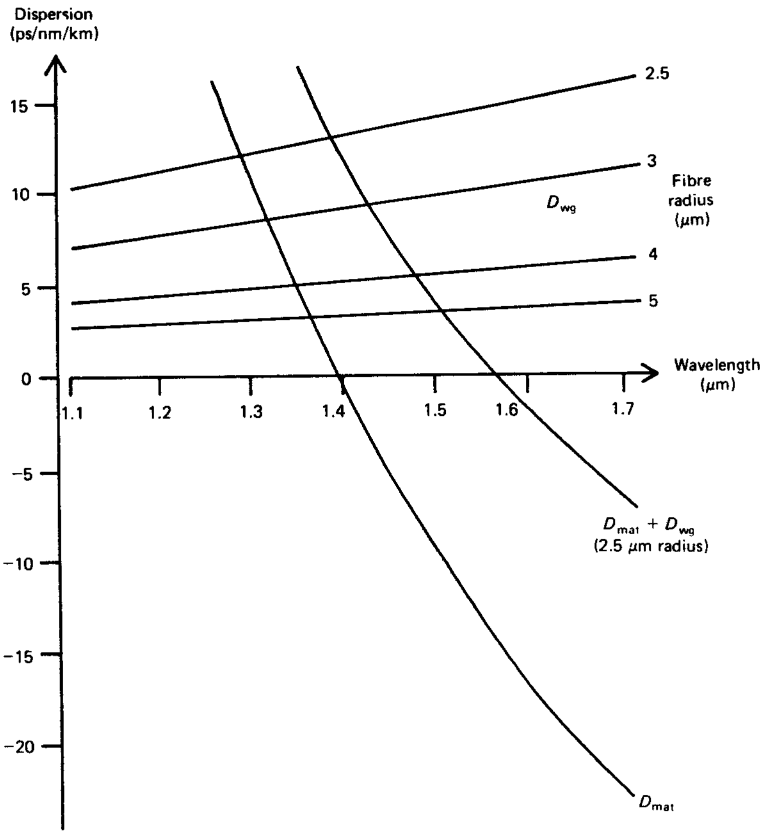


Figure 2.12 Showing the shift in the zero dispersion point, obtained by balancing D_{mat} with D_{wg}

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

$$\begin{aligned} n(r) &= n_1[1 - 2(\delta n/n_2)(r/a)^{\alpha}]^{\frac{1}{2}} \quad 0 \leqslant r \leqslant a \\ &= n_2 \qquad\qquad\qquad\qquad\quad r \geqslant a \end{aligned} \tag{2.87}$$

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 of 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, then 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, then the output pulse shape, h_{out}, will be

$$h_{\mathrm{out}}(t) = \frac{1}{\sigma\surd(2\pi)} \exp - t^2/(2\sigma^2) \tag{2.88}$$

and the pulse spectrum, H_{out}, will be

$$H_{\mathrm{out}}(\omega) = \exp-\omega^2\sigma^2/2 \tag{2.89}$$

where σ is the 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

$$\frac{H_{\mathrm{out}}(\omega)}{H_{\mathrm{out}}(0)} = \exp(-\omega^2\sigma^2/2) = \frac{1}{2} \tag{2.90}$$

and so the 3 dB bandwidth will be

$$\omega_{\mathrm{opt}} = 1.18/\sigma \tag{2.91}$$

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

$$\omega_{\text{elec}} = 0.83/\sigma \tag{2.92}$$

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, σ_{mat}, and the waveguide dispersion, σ_{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 S. D. Personick [6]). So, the total dispersion, σ, will be given by

$$\sigma^2 = {\sigma_{\text{mod}}}^2 + (\sigma_{\text{mat}} + \sigma_{\text{wg}})^2 \tag{2.93}$$

As an example, consider a 25 μm radius, MM fibre with $n_1 = 1.5$, $N_{g1} = 1.64$, $n_2 = 1.4$, $N_{g2} = 1.53$ and $D_{\text{mat}} = 500$ ps/nm/km. If the source is an LED with a line-width of 30 nm, and the emission is at 850 nm, then $\sigma_{\text{mod}} = 390$ ns/km (from equation 2.72) and $\sigma_{\text{mat}} = 15$ ns/km. (The waveguide dispersion is insignificant when compared with σ_{mat}.) Since we add these terms on a mean square basis, the modal dispersion is the dominant factor, and so $\sigma \approx \sigma_{\text{mod}} = 390$ ns/km. Thus the bandwidth of this fibre is only 2.6 MHzkm.

By way of contrast, consider a SM fibre with $n_1 = 1.48$, $N_{g1} = 1.64$, $n_2 = 1.47$, $N_{g2} = 1.63$ and $D_{\text{mat}} = -5$ ps/nm/km. If we use a 1.3 μm wavelength source, then the waveguide dispersion is −20 ps/nm/km (from equation 2.86). Thus the total dispersion is 25 ps/nm/km. With a 1 nm line-width laser source, this results in $\sigma = 25$ ps/km, and a bandwidth of 33 GHzkm. If we use a 30 nm linewidth LED source, then the bandwidth reduces to 1.1 GHzkm.

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 from 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 2.13. 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

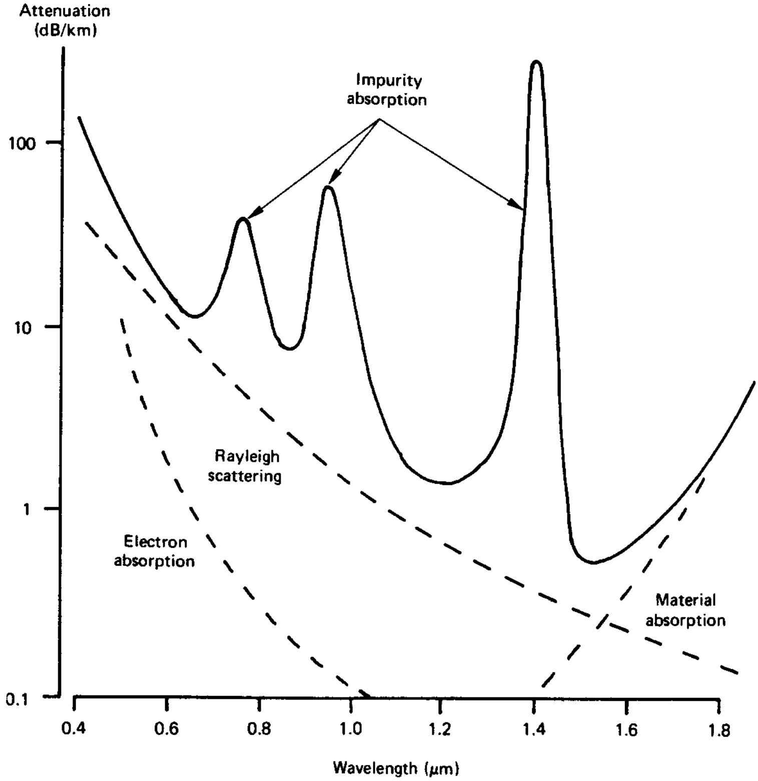


Figure 2.13 Attenuation/wavelength characteristic of a silica-based glass fibre

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 nm and 725 nm, as in figure 2.12. Between these peaks are regions of low attenuation — the transmission windows at 850 nm, 1.3 μm and 1.55 μm. As a large 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 (p.p.b.) 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 concentrations. 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.

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

2.5.2 Rayleigh scattering

Rayleigh scattering results from the scattering of light by 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 interest, 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 ZF_4, 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 0.14μm; however, the tail of this peak extends through to about 1 μm, so causing attenuation in the transmission windows.

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 $\mathrm{GeO_2}$ will increase the refractive index of $\mathrm{SiO_2}$ 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°. 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 2.14a, 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 2.14b). As the deposit forms, 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.

2.6.3 Fibre drawing from a preform

Having produced a preform, the fibre is drawn from it in a *fibre pulling tower*, shown schematically in figure 2.15. A clamp at the top of the tower holds the preform in place, and a circular drawing furnace softens the tip.

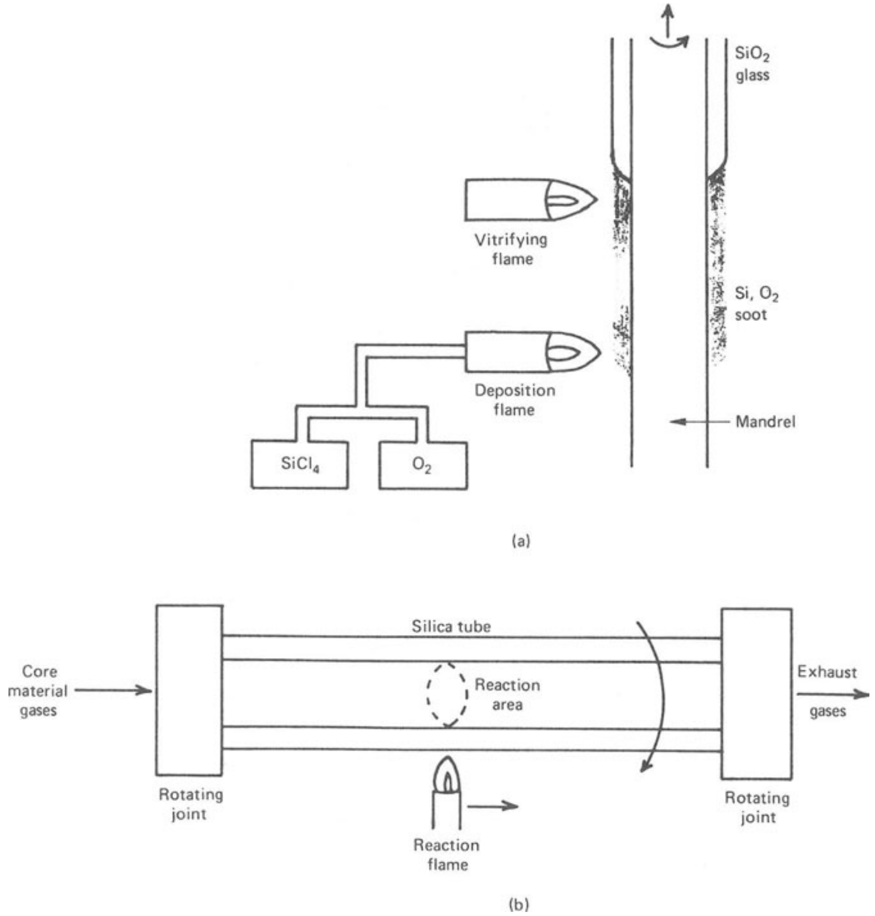


Figure 2.14 (a) Formation of silica cladding tube, and (b) deposition of core glasses

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.

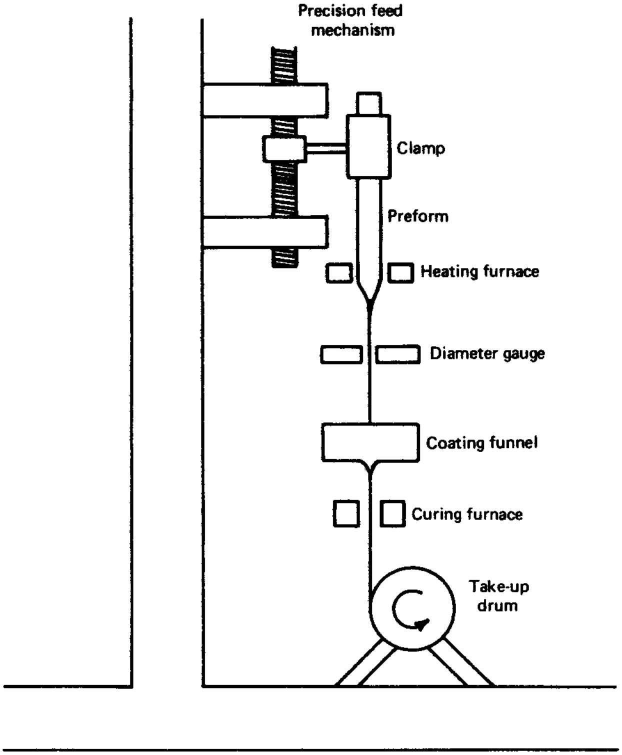


Figure 2.15 Schematic of a fibre pulling tower

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, replaces the preform. As figure 2.16 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

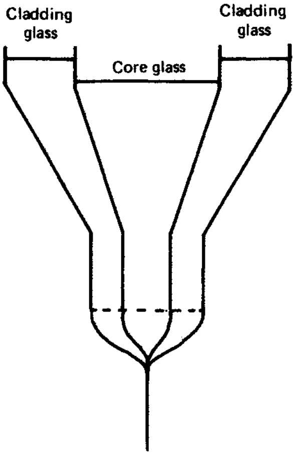


Figure 2.16 Double crucible method of optical fibre production

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 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 2.17.

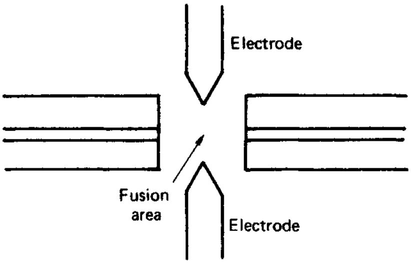


Figure 2.17 Schematic diagram of a fusion splicer

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 microscrope reveals whether the joint is satisfactory; if it is not, then the joint 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 2.18.

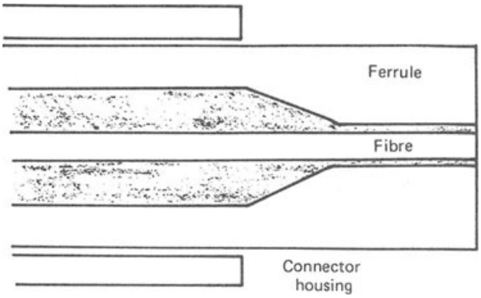


Figure 2.18 Basic construction of a ferrule type connector

Prior to insertion into the connector, the protective fibre coating is first 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. So polishing of the fibre end with successively finer abrasives is often used. The main body of the connector is then crimped onto 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, and then fusing them with the input fibre (figure 2.19). 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 2.19c. 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

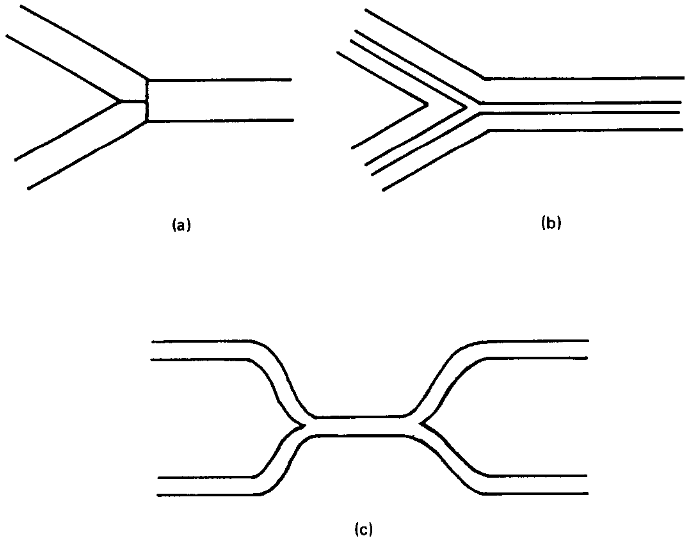


Figure 2.19 (a) Chamfered ends of output fibres for (b) a fused Y-coupler; (c) fused biconical taper coupler

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 2.20. 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*, and we shall deal with these in greater detail in the final chapter.

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.

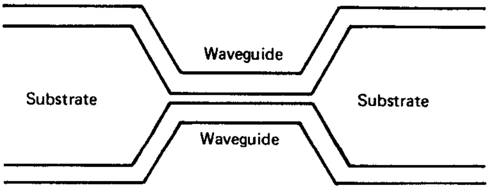


Figure 2.20 Schematic of an SM, evanescent field coupler/power splitter

3 Optical Transmitters

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

(1) It must be possible to operate the device continuously at room temperatures for many years.
(2) It must be possible to modulate the light output over a wide range of modulating frequencies.
(3) 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 the fibre, the emitting area should be small.
(5) To reduce material dispersion, the output spectrum should be narrow.

We shall examine two sources that satisfy these requirements—the light emitting diode, *LED*, and the semiconductor laser. As a complete mathematical analysis is rather involved, we will give a descriptive discussion of their various properties. Before we examine these sources in greater detail, it will be useful to discuss light emission in semiconductors. (Three very comprehensive references are Kressel and Butler, *Semiconductor Lasers and Heterojunction LEDs* [1]; Kressel *et al.*, chapter 2 in *Semiconductor Devices for Optical Communications* [2]; and Casey and Panish, *Heterostructure Lasers,* Parts A and B [3].)

3.1 Light emission in semiconductors

When forward biased, the barrier voltage of a p–n semiconductor junction diode reduces, so allowing electrons and holes to cross the depletion region (figure 3.1). The minority carriers, electrons in the p-type and holes in the n-type, recombine when electrons drop down from the conduction band, *CB*, to the valence band, *VB*. 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 vibra-

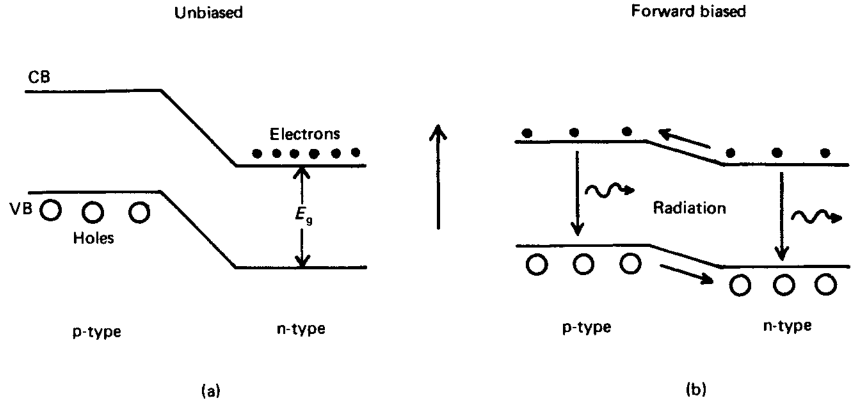


Figure 3.1 The p–n junction under (a) zero, and (b) forward bias

tions 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 we can regard as being proportional to the electron momentum.)

Figure 3.2a shows a simplified *E–k* diagram for an indirect band-gap material, such as silicon, *Si*, or germanium, *Ge*. As can be seen, the electron and hole momenta are different. So, if an electron drops down

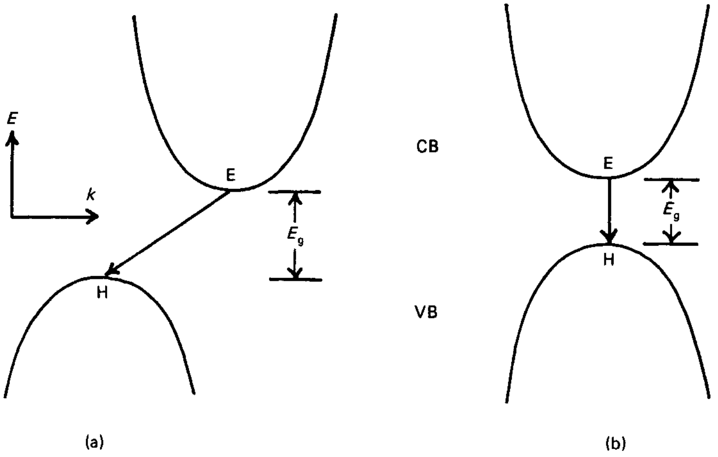


Figure 3.2 Energy/wave-number diagrams for (a) an indirect, and (b) a direct band-gap semiconductor

from region E in the CB to region H in the VB, then a change of momentum has to take place and a phonon is emitted.

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 3.2b. As can be seen, the electron and hole momenta are the same, that is 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

$$\lambda_0 = \frac{hc}{qE_g} = \frac{1244}{E_g} \text{ (nm)} \tag{3.1}$$

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 optical fibre light sources. However, the III–V semiconductors have a direct band-gap, and so are most 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 nm to 800 nm. The 100 nm spread in wavelength occurs because E_g depends upon the ratio of Ga to Al. (We shall return to this point presently.)

Table 3.1 Characteristics of various semiconductor materials (D—direct, and 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

3.2 Heterojunction semiconductor light sources

Light emission can occur on both sides of the p–n junction. However, if we concentrate the recombining carriers to a small active area, the light output will be increased, and we can launch more power into a fibre. Such confinement can be achieved by forming a junction between two dissimilar band-gap materials – 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 shall initially examine a single heterojunction, or *SH*, diode.

Figure 3.3 shows the energy diagram of a 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 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 3.3b. 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 structure will confine both holes and electrons to a narrow active layer. As figure 3.4 shows, the potential steps either side

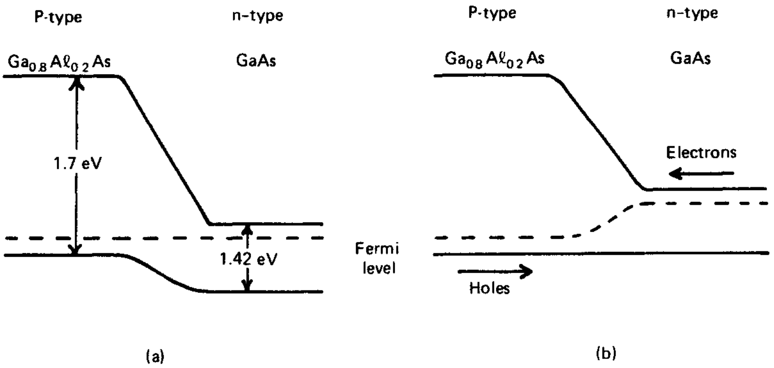


Figure 3.3 Energy diagram of a heterojunction under (a) zero bias, and (b) forward bias

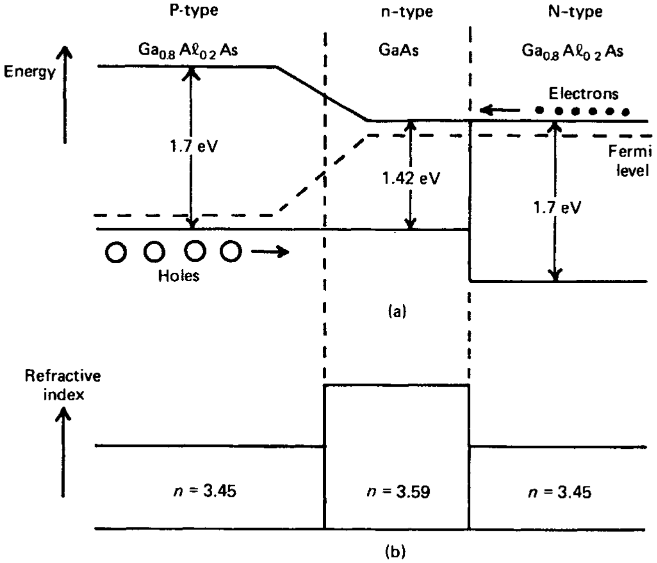


Figure 3.4 (a) Energy diagram, and (b) refractive index profile for a forward-biased P–n–N, double heterojunction diode

the active region, the GaAs, inhibit carrier movement. So there can be a large number of carriers in the active layer, which results in 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 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 $0 < x < 0.45$, then we can find E_g from the following empirical relationship:

$$E_g = 1.42 + 1.25x + 0.27x^2 \tag{3.2}$$

Because the active layer emits the light, the surrounding material can be an indirect band-gap semiconductor. As an example, a diode wiith $x = 0.03$ and $y = 0.2$ will emit light of wavelength 852 nm. We can find the refractive index of the material from

$$n = 3.59 - 0.71x \text{ for } 0 < x < 0.45 \tag{3.3}$$

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

$$E_g = 1.35 - 1.89x + 1.48x^2 - 0.56x^3 \tag{3.4}$$

with the refractive index being given by

$$n^2 = 9.6 + 4.52x - 37.62x^2 \tag{3.5}$$

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.

3.3 Light emitting diodes (LEDs)

There are currently 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

3.3.1 Surface emitting LEDs

Figure 3.5 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 by the LED. A micro-lens placed in the well at the top of the device will increase the coupled power. However, the efficiency of this arrangement is dependent upon 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.

3.3.2 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

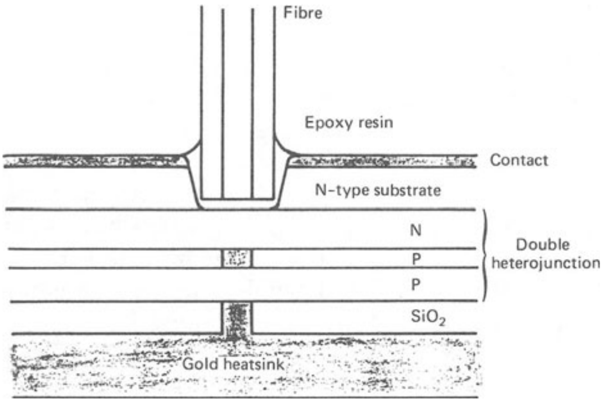


Figure 3.5 Cross-section through a typical surface emitting LED

LED. Such a device is known as an *edge emitting LED*, or *ELED*, and a typical structure is shown in figure 3.6.

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.

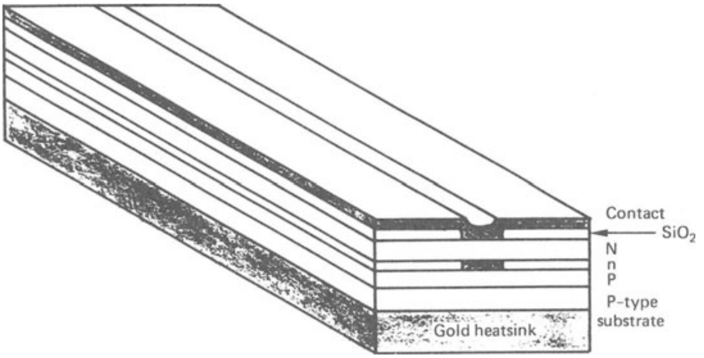


Figure 3.6 Structure of an edge-emitting, N–n–P, double heterojunction, stripe-contact LED

3.3.3 Spectral characteristics

As we saw at the start of this chapter, 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 3.7). Most of the recombinations take place over the average band-gap difference; however some recombination occurs between the higher and lower energy levels. Thus the energy lost through recombination has a mean value of E_g and a deviation of δE_g. This deviation is typically between kT and $2kT$, where k is Boltzmann's constant (1.38×10^{-23} J/K) and T is the absolute temperature of the junction. Although the actual deviation is dependent on the amount of impurity doping, the approximation will suit our purposes.

The spread in band-gap results in a spread of emitted wavelengths about a nominal peak. 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

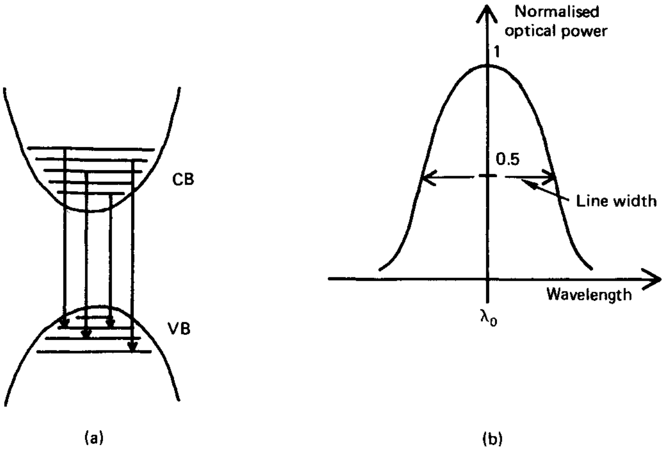


Figure 3.7 (a) Photon emission from conduction band energy levels, and (b) resultant spectral characteristic

large number of modes into the fibre, and so modal dispersion may be dominant. In most LEDs, the linewidth is typically 30 nm.

3.3.4 Modulation capabilities and conversion efficiency

The output power/drive current characteristic of an LED is approximately linear. If we superimpose an a.c. signal onto a d.c. bias level, we can write the output optical power, $p(\omega)$, as

$$p(\omega) = \frac{p(0)}{(1 + (\omega\tau)^2)^{\frac{1}{2}}} \tag{3.6}$$

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

$$\frac{1}{2}\pi\tau \text{ Hz}$$

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. When both radiative and non-radiative recombination is present, τ will be given by

$$\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}} \tag{3.7}$$

where τ_r and τ_{nr} are the radiative and non-radiative recombination times respectively. These time constants will also give us a measure of the diode conversion efficiency. It may be shown that the internal quantum efficiency, η_{int}, is given by

$$\eta_{int} = \frac{\tau_{nr}}{\tau_{nr} + \tau_r} \tag{3.8}$$

So in order to produce a fast device, both τ_r and τ_{nr} should be kept low, with the proviso that $\tau_r >> \tau_{nr}$ in order to keep the efficiency high. With a low level of doping, that is, lower than the injected carrier concentration,

$$\tau_r \propto \left(\frac{d}{J}\right)^{\frac{1}{2}} \tag{3.9}$$

where d is the distance between the heterojunctions and J is the current density. With such low doping levels, recombination at the heterojunction interfaces determines the value of τ_{nr}. The recombination is due to mismatches between the heterojunction crystal lattices, characterised by the *surface recombination velocity, S*. Thus, with low doping

$$\tau_{nr} \propto \frac{d}{S} \tag{3.10}$$

As τ_r and τ_{nr} and 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.

For high doping levels, > 10^{18}, 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.

3.3.5 LED drive circuits

For analogue modulation of an LED, we can use the simple class A amplifier shown in figure 3.8. Provided the modulation depth, m, is less than 100 per cent, no signal distortion will occur. The modulation depth is defined as

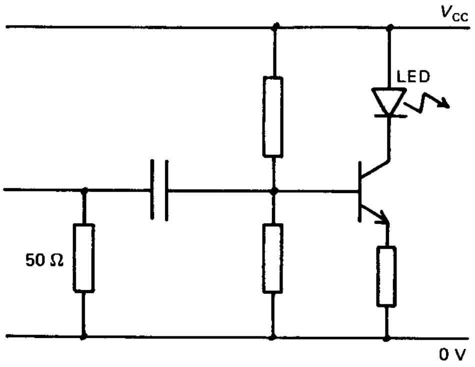


Figure 3.8 A simple analogue driver for LED sources

$$m = \frac{\delta I}{I_B} \tag{3.11}$$

where I_B is the LED bias current. With careful selection of the drive transistor, 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 3.9a. 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 3.9b). When the input is high, T_1 turns on and current is diverted away from the LED, so turning it 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.

3.4 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. Instead of being absorbed, the incident photon causes an electron to cross the band-gap, so generating

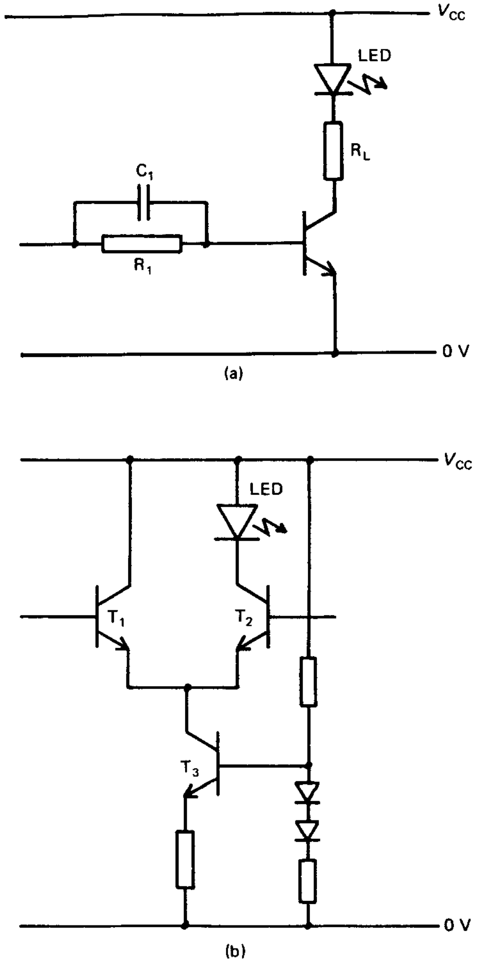


Figure 3.9 (a) A simple digital driver with speed-up capacitor, and (b) a basic emitter-coupled switch for LED light sources

another photon, figure 3.10. 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 original photon has been amplified; indeed, the acronym laser stands for *l*ight *a*mplification by the *s*timulated *e*mission of *r*adiation. Because 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

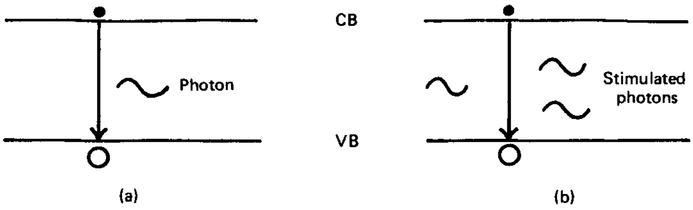


Figure 3.10 Light generation by (a) spontaneous emission, and (b) stimulated emission

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 3.11. The 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.

3.4.1 SLD characteristics

Laser diodes and LEDs differ in several ways: a laser diode requires the application of a constant current to maintain stimulated emission; the

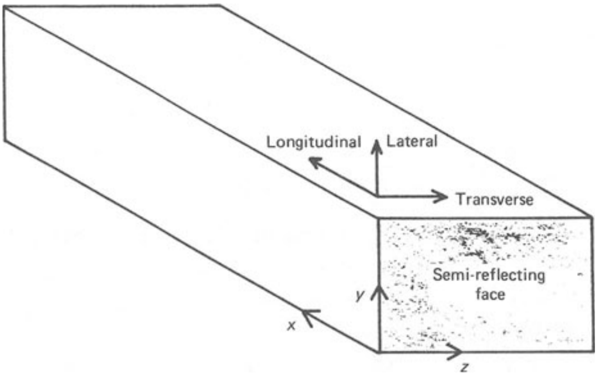


Figure 3.11 A basic Fabry–Perot cavity

output beam is more directional; and the response time is faster. Rather than perform a complete analysis of laser diode characteristics, we will use a simplified analysis based on our previous discussion of modes in a planar dielectric. We shall consider a stripe contact laser, which is similar in construction to a stripe contact ELED.

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. We will initially consider a longitudinal TE wave, $E(x,t)$. If we neglect the effects of the cavity side-walls, and assume that the cavity confines all of the E field, then $E(x,t)$ is given by

$$E(x, t) = |E|\exp(-\alpha x/2)\exp\mathrm{j}(\omega t - \beta_1 x) \qquad (3.12)$$

where α is the attenuation of the optical *power* per unit length (hence the factor $\frac{1}{2}$) 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

$$E(0, t) = |E|\exp(0)\exp\mathrm{j}\omega t \qquad (3.13)$$

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

$$E_\mathrm{r}(0, t) = \sqrt{(R_1R_2)}|E|\exp[(g - \alpha)L]\exp\mathrm{j}(\omega t - 2\beta_1 L) \qquad (3.14)$$

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

$$\sqrt{(R_1R_2)}|E|\exp[(g - \alpha)L] \geqslant |E|$$

Therefore the optical gain is given by

$$g \geqslant \alpha + \frac{1}{2L} \times \ln\left[\frac{1}{R_1R_2}\right] \qquad (3.15)$$

As we noted earlier, it is the current density in the active region, J , that produces the population inversion, and hence the cavity gain. After a rather lengthy derivation, the relationship between g and J in an AlGaAs laser diode can be approximated by

$$g = 0.045\left[\frac{J}{d} - 4222\right] \qquad (3.16)$$

where g is the optical gain per cm, J is in units of A cm^{-2}, and d is the thickness of the active layer in μm. If the equality in (3.15) is satisfied, the cavity will begin to lase. We can find the value of J required for lasing, by substituting g from (3.15) into (3.16). As an example, if the loss per cm is 10, $R_1 = R_2 = 0.3$, and the cavity length is 300 μm, then the required gain per cm is 50. If we take an active layer thickness of 2 μm, then lasing will begin at a current density of 10.7 kA/cm^2. Because it marks the boundary between operation as an ELED and operation as a laser, this value of J is known as the *threshold current density*, J_{th}.

We are more usually concerned with the current required for lasing to occur — the *threshold current*, I_{th}. We can find this by multiplying J_{th} by the area of the active region. So, if we assume an active region width of 20 μm, then I_{th} = 642 mA. Figure 3.12 shows the variation of output power with diode current, and it clearly shows the threshold point above which the light output increases dramatically for a small change in current.

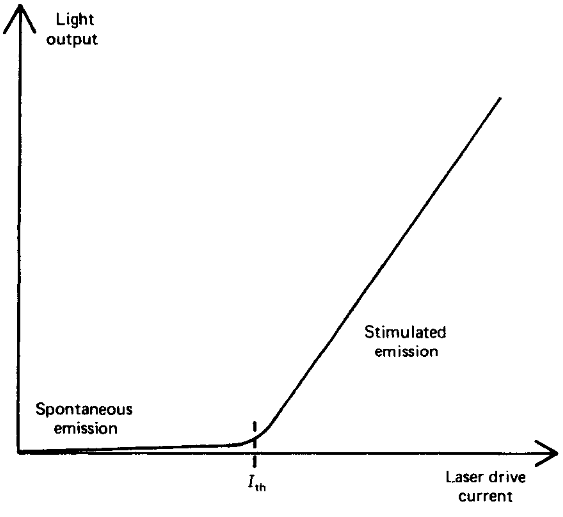


Figure 3.12 Variation of light output with drive current for a semiconductor laser diode

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. So, to return to (3.13) and (3.14), we can see that the phase of the two waves must be the same at $x = 0$, that is

$$\exp j(-2\beta_1 L) = 1 \tag{3.17}$$

Therefore

$$2\beta_1 L = 2\pi N \tag{3.18}$$

where N is an integer. Since $\beta_1 = 2\pi n_1/\lambda_0$, (3.18) becomes

$$\lambda_0 = \frac{2n_1 L}{N} \tag{3.19}$$

So, the laser will only amplify wavelengths that satisfy (3.19). 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 like that shown in figure 3.13a; solution of (3.19) will yield the mode spacing. If we consider an $Al_{0.03}Ga_{0.97}$ As diode, then the nominal wavelength is 852 nm and the refractive index is 3.57. Taking a cavity length of 600 μm gives a nominal mode number of 5028. The next mode corresponds to $N = 5029$, resulting in a mode spacing of 0.17 nm. With a line-width of 5 nm, this results in 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, an increase in bias current causes a rise in cavity gain, and this reduces the linewidth. The reduction occurs because the cavity exponentially amplifies the first mode to reach threshold, at the expense of all other modes. In practice, modes close to the fundamental also undergo amplification, and so the output consists of a range of modes following a *gain profile* (Figure 3.13b and c). We can approximate this profile to the Gaussian distribution

$$g(\lambda) = g(\lambda_0)\exp\left[\frac{-(\lambda - \lambda_0)^2}{2\sigma^2}\right] \tag{3.20}$$

where σ is the line-width of the laser output. This result, together with the line spectrum, causes the emission spectrum shown in figure 3.13c. The line-width of typical stripe contact SLDs can vary from 2 to 5 nm.

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

$$d < \frac{\lambda_0}{2(n_1^2 - n_2^2)^{\frac{1}{2}}} \tag{3.21}$$

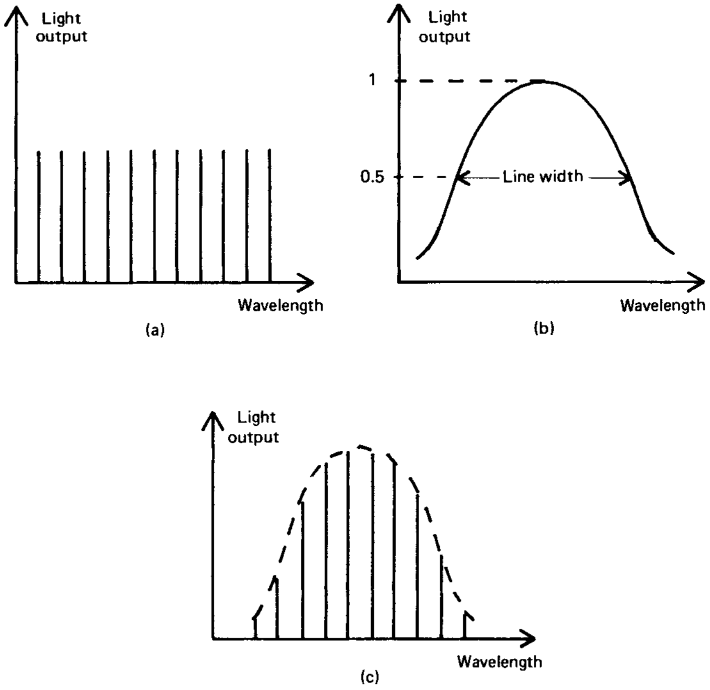


Figure 3.13 (a) Allowable modes in a Fabry–Perot cavity, (b) Gaussian gain function, and (c) resultant spectral characteristic

where n_2 is the refractive index of the surrounding material. In most laser diodes, the active region is typically less than 1 μm thick and (3.21) 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.

3.4.2 SLD structures

As lasers are normally used in long-haul, high-data-rate routes, which use SM fibres, it is generally desirable to minimise the line-width 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 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 3.15), 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 which 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 3.14. 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 optical waveguide, and so single lateral mode operation is often achievable. The threshold current of these devices is typically 30 mA, and they are often used in long-wavelength systems.

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 a BH laser can consist of a single line — a considerable advantage in long-haul routes. Unfortunately the gain profile is dependent on the junction temperature, and so the wavelength of emission can change during operation. If the laser is to operate at a zero dispersion wavelength, then a change in source wavelength will result in some 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 3.15). If the period of this grating is a multiple number of half wavelengths of the light in the active region, then only one mode will propagate; thus

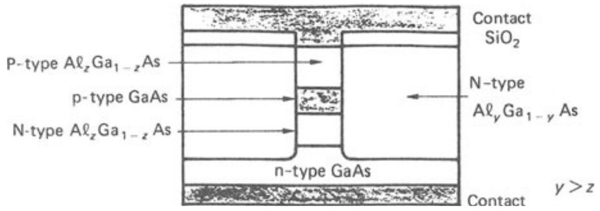

Contact
SiO$_2$
P-type Aℓ$_z$Ga$_{1-z}$As
p-type GaAs
N-type Aℓ$_z$Ga$_{1-z}$As
N-type Aℓ$_y$Ga$_{1-y}$As
n-type GaAs
Contact
$y > z$

Figure 3.14 Cross-section through a buried-heterojunction, semiconductor laser diode

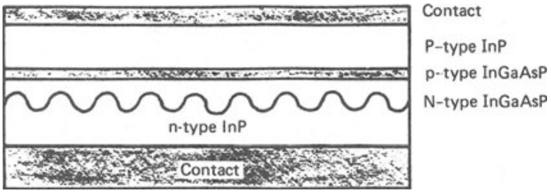


Figure 3.15 Cross-section through a distributed feedback semiconductor laser diode

the output consists of a single wavelength. It should be noted that the grating is not part of the active layer. This is because a grating in the active region will cause surface dislocations, which would decrease the efficiency. 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.

3.4.3 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 3.16, in which a monitor photodiode attached to the non-emitting laser facet provides the feedback signal. In order to alleviate heatsinking problems, some commercial laser packages incorporate semiconductor Peltier coolers, which also help to keep the threshold current low.

As previously mentioned, the light output of SLDs is due to stimulated emission. As this process is faster than spontaneous emission, the emitter-coupled circuit of figure 3.9b, shown previously, is often used. For high-speed operation, microwave bipolar transistors, or GaAs MESFETs, must be used. Although the rise-time of the laser optical pulse can be very fast, ≈ 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.

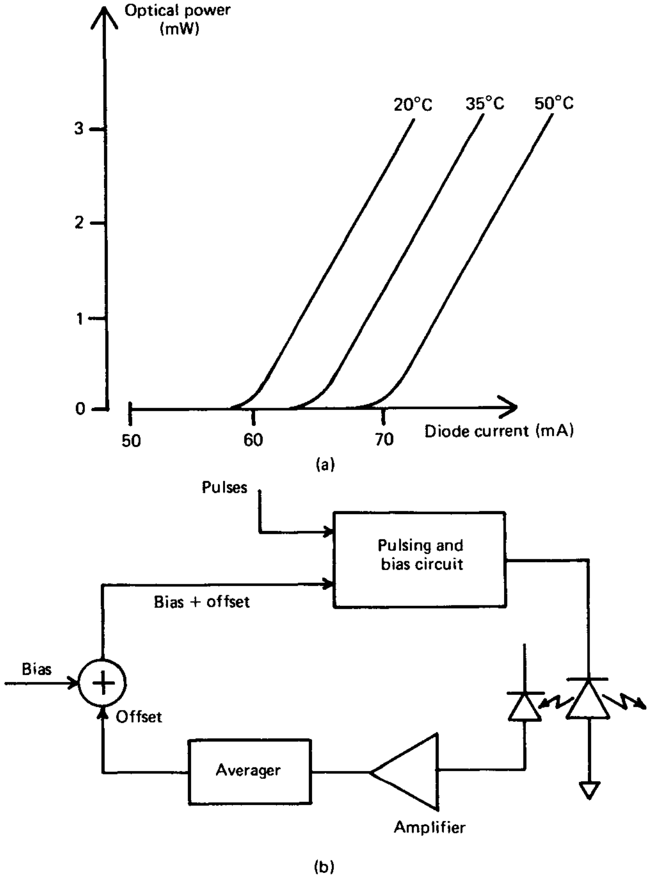


Figure 3.16 (a) Variation of I_{th} with temperature, and (b) a simple laser bias stabiliser

3.4.4 *Packaging and reliability*

SLDs for use in the laboratory are usually mounted in brass studs, similar to the one shown in figure 3.17. With this package, the body of the stud forms the anode of the SLD, and so the lead at the rear of the package must be connected to a negative voltage, current source. A thread on the back of the stud enables the diode to be bolted 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 3.17. 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

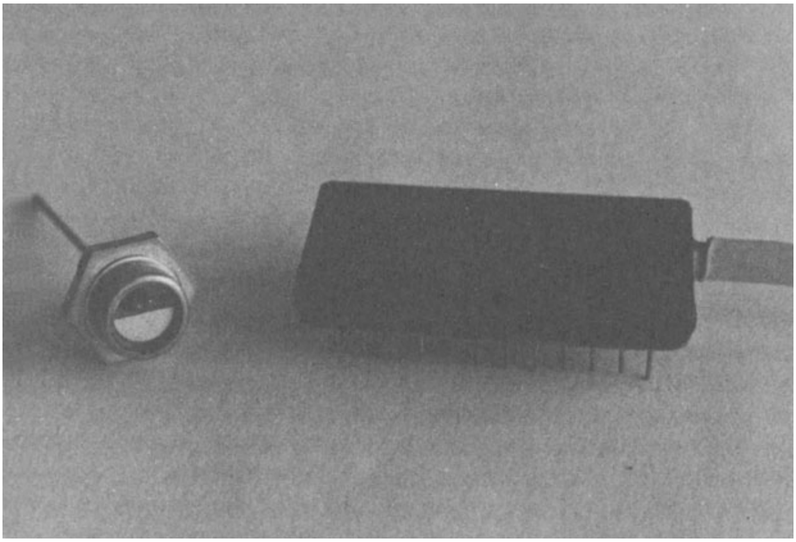


Figure 3.17 Stud mounted and dual-in-line laser diode packages

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.

When a SLD ages, the threshold current requirement tends to increase. The feedback loop described previously can compensate for this; however the threshold point tends to become less well defined as time passes. Thus the threshold current monitoring circuit 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.

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 an 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 a 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 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 last 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

$$I_{\text{diode}} = (I_{\text{d}} + I_{\text{s}})(\exp[Vq/\eta kT] - 1) \tag{4.1}$$

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 can be seen, 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*. When operating in this mode, the frequency response of the diode is poor, and so photovoltaic diodes are rarely used in optical links.

When reverse biased, 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

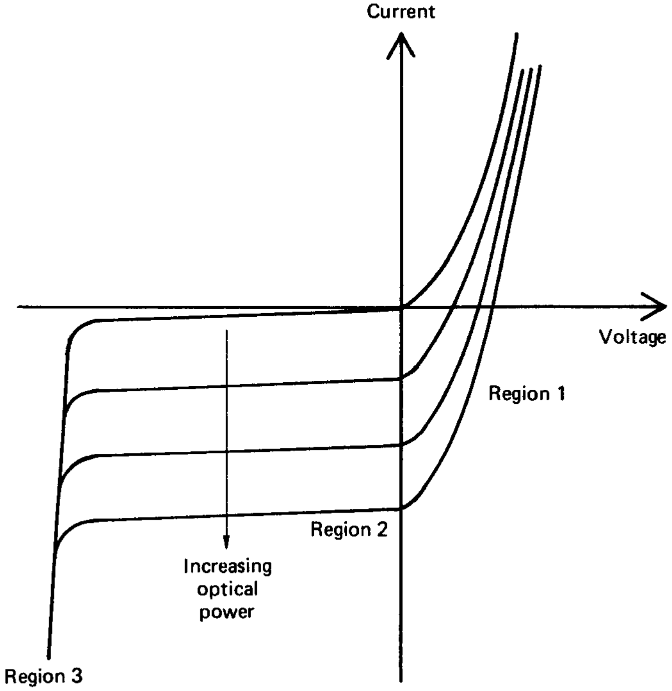


Figure 4.1 *V–I* characteristic of a photodiode, with varying amounts of incident optical power

$$I_{\text{diode}} = I_{\text{d}} + I_{\text{s}} \tag{4.2}$$

The *responsivity* of the photodiode, R_0, is defined as the change in reverse bias current per unit change in optical power, so efficient detectors have large responsivities. Detectors with a p–*i*–n structure, *PIN* photodiodes, have been produced with responsivities close to unity.

Avalanche photodiodes, APDs, operate in region 3 of the *V–I* characteristic. When biased in this region, a photo-generated electron–hole pair 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 characteristic is very steep in this region, and 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, the same basic arguments can be applied 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, 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:

$$\frac{\delta I}{I} = -\alpha\delta x \tag{4.4}$$

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, $x = 0$, then

$$\ln\frac{I}{I_0} = -\alpha x$$

giving

$$I = I_0\exp(-\alpha x) \tag{4.5}$$

From our discussion of photo-emission, we can see that α will vary with wavelength. Figure 4.2 shows this variation for several semiconductor materials. The plots clearly indicate 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 the differing density of energy levels in the

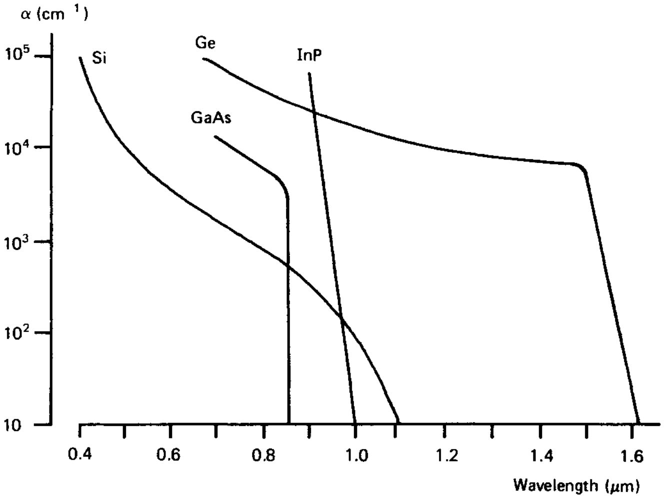


Figure 4.2 Variation of absorption coefficient with wavelength, for a number of semiconductor materials

VB and CB of the material — the same mechanism 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. As an example, a 40 μm thick Si detector operating at 850 nm has an α value of 6.3×10^4 m^{-1} which results in absorption of 92 per cent of the light. 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

In the introduction to this chapter we saw that a photon of light causes the excitation of an electron to the CB, where it takes part in the conduction of current. We expressed the reverse biased diode current as

$$I_{\text{diode}} = I_{\text{d}} + I_{\text{s}} \tag{4.6}$$

Now, I_s is directly dependent on the rate of generation of electron–hole pairs 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

$$N_{\mathrm{ph}} = \frac{P}{E_{\mathrm{ph}}}$$

$$= \frac{P\lambda_0}{hc} \tag{4.7}$$

Only some of these photons generate electron–hole pairs. Specifically, the number of carrier pairs generated per second, N is given by

$$N = \eta N_{\mathrm{ph}}$$

$$= \frac{\eta P\lambda_0}{hc} \tag{4.8}$$

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

$$I_{\mathrm{s}} = qN$$

$$= \frac{q\eta P\lambda_0}{hc} \tag{4.9}$$

We can rearrange this equation to give the change in current per unit change in optical power, the responsivity. Hence

$$R_0 = \frac{I_{\mathrm{s}}}{P} = \frac{q\eta\lambda_0}{hc} \tag{4.10}$$

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, and light absorption in, 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). In general, the light intensity in the n-type is very low, and we can ignore absorption in this region. However, absorption in the p-type has a dramatic effect on the quantum efficiency.

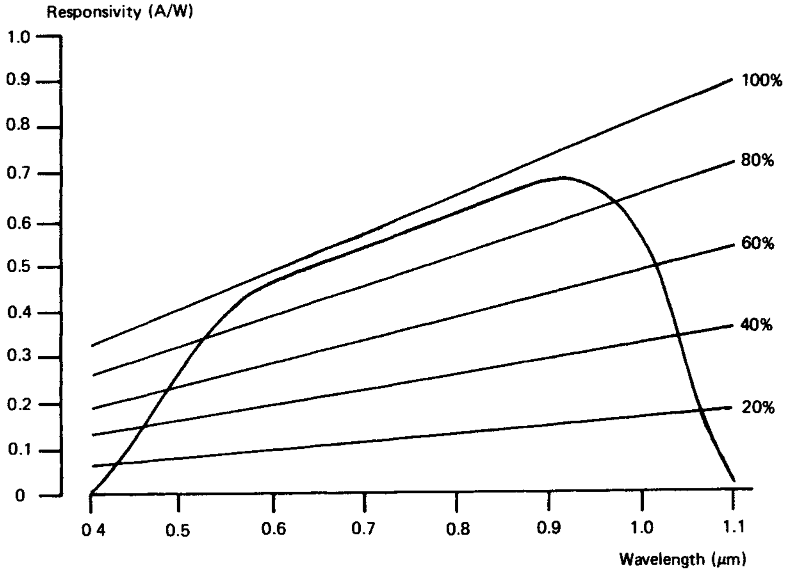


Figure 4.3 Variation of responsivity with wavelength for a typical Si PIN photodiode. Also shown is the theoretical variation of R_0 for a range of quantum efficiencies

Incident light of a low wavelength gives a low penetration depth ($1/\alpha$), resulting in electron–hole generation in the p–type layer. Unless carrier generation occurs within a diffusion length of the depletion layer boundary, the electrons recombine and the diode current does not change. However, if photon absorption occurs within a diffusion length of the depletion layer boundary, then the electrons 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

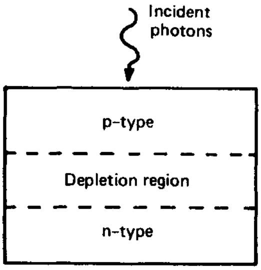


Figure 4.4 Schematic of a reverse-biased p–n junction photodiode

p-type only occurs within a diffusion length of the depletion layer, which explains why the quantum efficiency reduces with low wavelength.

Photon absorption in the depletion region causes the generated electrons and holes 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 clearly 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.

A wide depletion region implies a high reverse bias voltage. However, the maximum depletion layer thickness is unlikely to reach the 40 μm depth required for efficient absorption. The solution is to insert an intrinsic semiconductor layer between the p-and n-type, so that the depletion layer depth is effectively increased. The resulting diode is known as a *PIN* diode.

4.2 PIN photodiodes

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

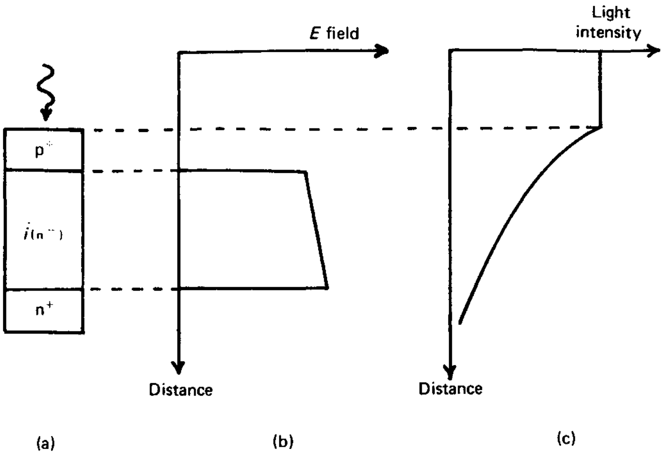


Figure 4.5 (a) PIN photodiode schematic, (b) electric field intensity, and (c) light intensity across the photodiode

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 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 the problem, some photodiode packages have a hemispherical lens which collects light from a large area and focuses it onto 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, then the depletion region will extend beyond the contact rim. Since the SiO_2 layer is transparent to light, photons can be absorbed without passing 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, if the doping in the p-type is higher than in the

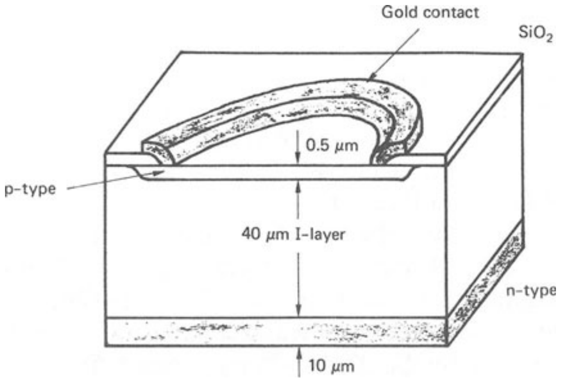


Figure 4.6 Cross-section through a typical silicon PIN photodiode

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^- junction, most of the depletion region exists in the lightly doped intrinsic layer. If the doping level in the I-layer is N_D, and V_t is the barrier potential ($\approx$ 0.75 V for Si) then, under a reverse bias of V_b, the depletion layer depth, d, will be

$$d = \left(\frac{2\epsilon_0\epsilon_r(V_b + V_t)}{qN_D}\right)^{\frac{1}{2}} \tag{4.11}$$

If the I-layer is completely depleted, that is, 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 (4.11) to yield V_b. As an example, a 40 μm thick I-layer with N_D equal to 10^{19} m^{-3}, gives a punch-through voltage, for a Si PIN diode, of around 12 V.

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 the example used. 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 transit time across the I-layer; carrier transit time from the p^+ and n^+ regions; and the junction capacitance interacting with an external load resistance.

The transit time across the I-layer depends upon the depth of the depletion region and the magnitude of the E field. If we take the previous example, then E = 0.3 V μm^{-1} at a bias of 12 V. The mobility of electrons in intrinsic silicon is 1350 cm^2 V^{-1} s^{-1} and so the electron velocity is 4×10^4 m s^{-1} giving a transit time of approximately 1 ns. If the E field is sufficiently high, then the electrons would reach their saturation velocity of 10^5 m s^{-1}, resulting in a transit time of 0.4 ns.

If the transit time were the only time constant in the diode, then the bandwidth would be of the order of one gigahertz. However, if electron–hole pair generation occurs within a diffusion length of the depletion region, L_n, the electrons will diffuse into the I-layer. If the thickness of the p^+ layer is greater than L_n, the maximum transit time to the depletion region is the lifetime of the electrons. This is usually of the order of 10 ns and so, for high-speed operation, the p^+ layer must be made as thin as possible. A similar argument applies to carrier pair generation in the n^+ layer but, because the light is generally of such low intensity, the diffusion effects are normally negligible.

If we initially neglect the package capacitance, then the junction capacitance will tend to dominate over the diffusion capacitance. The junction capacitance, C_j, is given by

$$C_j = \frac{\epsilon_0 \epsilon_r A}{W} \tag{4.12}$$

where A is the cross-sectional area of the diode and W is the depletion region thickness. Taking the previous example, A is approximately 2×10^3 μm^2 and W is 40 μm. Thus C_j is about 5 fF, which is low enough for the package capacitance to dominate. (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 range from less than 0.8 pF for high-speed detectors, to 150 pF for low-speed, large area detectors.

4.2.4 Photodiode circuit model

Figure 4.7 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 of the bulk semiconductor and the contact resistance. The load resistor, R_L, shunts the *total* diode capacitance, C_d, and this time constant usually limits the speed of response.

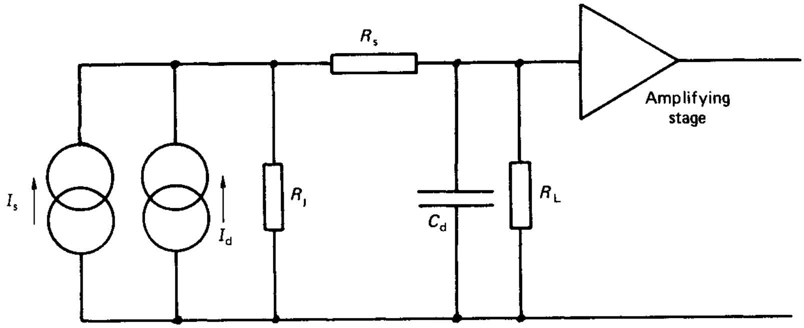


Figure 4.7 A circuit model for a typical photodiode

In general, we can ignore R_j and R_s, and so the bandwidth of the detector is given by

$$f = \frac{1}{2\pi R_L C_d} \tag{4.13}$$

We should note that any following amplifier will have a time constant, and this may prove to be the limiting factor.

4.3 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 to 1.7 μm. Thus detectors made of a similar material should respond to 1.3 or 1.55 μm light. In practice, an alloy of In, Ga and As is used, where the proportions of In and Ga alter the band-gap. 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 × 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 4.8 shows the structure of a typical InGaAs photodiode.

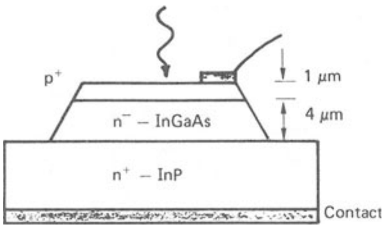


Figure 4.8 Cross-section through a typical long-wavelength PIN photodiode

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 μm^{-1} which is well above the 1 V μm^{-1} required for the 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 0.06 pF, which is considerably less than that of an equivalent Si diode. However if the detector is packaged, the capacitance may well rise to 0.8 pF or more. Thus for high-speed operation, hybrid thick-film receivers use unpackaged photodiodes.

4.4 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 4.9 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. It should be noted that an APD biased close to breakdown could breakdown because of the reverse leakage current. Thus APDs are usually biased just below breakdown, with the bias voltage being tightly controlled.

4.4.1 APD structures

Figure 4.10 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

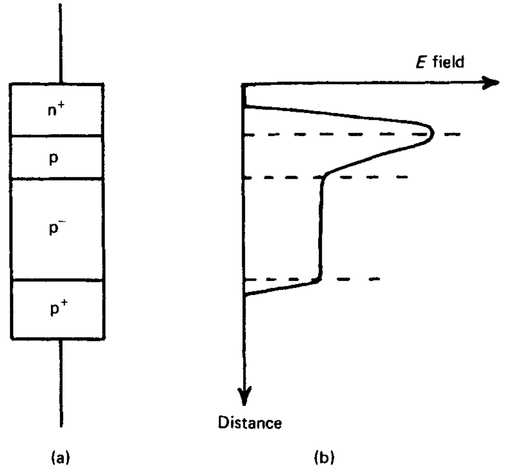


Figure 4.9 (a) APD schematic, and (b) variation of electric field intensity across the diode

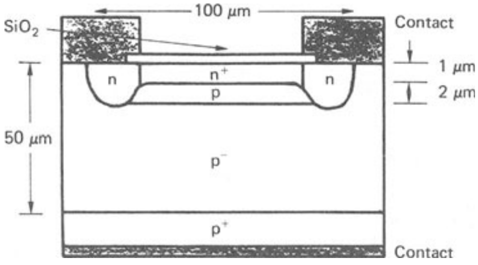


Figure 4.10 Cross-section through a typical silicon APD

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, germanium APDs can be used. However, as noted before, these diodes exhibit a high dark 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 4.11 shows a typical structure. In this particular design, the InGaAs 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 photogenerated carriers. However, because of the n–N heterojunction, only holes cause breakdown in the N-type region.

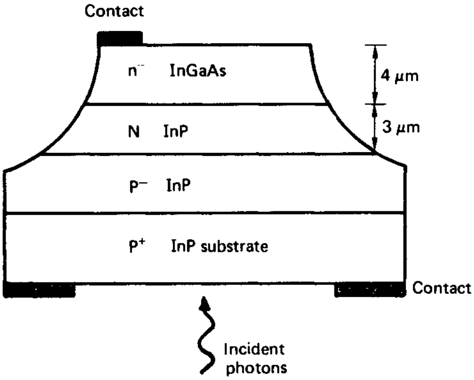


Figure 4.11 Cross-section through a typical long-wavelength heterojunction APD

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.

4.4.2 *Current multiplication*

In an APD, avalanche multiplication increases the primary current, that is, the unmultiplied photocurrent given by (4.9). Thus we can write the responsivity as

$$R_0 = \frac{Mq\eta\lambda_0}{hc} \qquad (4.14)$$

where M is the multiplication factor. It therefore follows that M is given by

$$M = \frac{I_m}{I_s} \qquad (4.15)$$

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

$$M = \frac{1 - k}{\exp(-(1 - k)\alpha_e W) - k} \tag{4.16}$$

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

$$M = \frac{1}{1 - (V/V_{br})^n} \tag{4.17}$$

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

$$V_{br}(T) = \mathrm{V}_{br}(T_0) + a(T - T_0) \tag{4.18}$$

and

$$n(T) = n(T_0) + b(T - T_0)$$

where a and b are empirical constants, <1. So, as shown in figure 4.12, M depends upon both the bias voltage and the temperature.

4.4.3 *Speed of response*

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 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

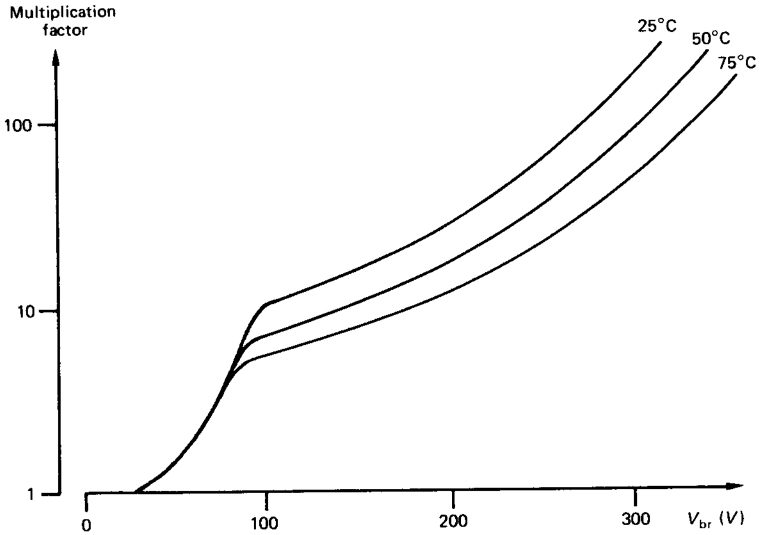


Figure 4.12 Theoretical variation of multiplication factor, M, with reverse bias voltage, V_{br}, for three different temperatures. Unity gain has been taken at $V_{br} = 30$ V

$$\tau = t_d + t_a + t_s$$

$$= \frac{W_i}{v_{se}} + \frac{MkW_a}{v_{se}} + \frac{1}{v_{sh}}(W_a + W_i) \tag{4.19}$$

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.

4.5 Photodiode noise

The signal at the end of an 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

$$\frac{S}{N}=\frac{<I_s^2>}{<i_n^2>_{pd}+<i_n^2>_c} \tag{4.20}$$

where $<I_s^2>$ is the mean square, *m.s.*, value of the photodiode signal current, $<i_n^2>_{pd}$ is the m.s. value of the photodiode noise, and $<i_n^2>_c$ 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*, $<i_n^2>_Q$, due to quanta of light generating packets of electron–hole pairs; thermally generated dark current, $<i_n^2>_{DB}$, occurring in the photodiode bulk material; and surface leakage current, $<i_n^2>_{DS}$. (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 $<i_n^2>_{pd}$ as

$$<i_n^2>_{pd}=<i_n^2>_Q+<i_n^2>_{DB}+<i_n^2>_{DS} \tag{4.21}$$

4.5.1 PIN photodiode noise

No current multiplication occurs in a PIN detector and so, with a receiver noise equivalent bandwidth B_{eq}, we can write the detector S/N as

$$\frac{S}{N}=\frac{<I_s^2>}{<i_n^2>_Q+2qI_{DB}B_{eq}+2qI_{DS}B_{eq}} \tag{4.22}$$

where I_{DB} and I_{DS} are the bulk and surface leakage currents respectively. However, if the leakage currents are negligible

$$\frac{S}{N}=\frac{<I_s^2>}{<i_n^2>_Q} \tag{4.23}$$

With this condition, we must take account of the quantum nature of light and so the S/N defined by (4.23) is known as the *quantum limit.* We can determine $<i_n^2>_Q$ 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

$$p(n)=\mathrm{Pos}[n,m]=\frac{m^n e^{-\mathrm{m}}}{n!} \tag{4.24}$$

Taking a quantum efficiency of unity, the number of carrier pairs 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 electron–hole pair 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

$$\begin{aligned} P_e &= \tfrac{1}{2}(P(0|1) + P(1|0)) \\ &= \tfrac{1}{2}\frac{(m^0 e^{-m} + 0)}{(0!)} \\ &= \tfrac{1}{2}\, e^{-m} \end{aligned} \tag{4.25}$$

So, for a typical error rate of 1 bit in 10^9, we require an average of $m = 21$ electron–hole pairs. 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

$$P = \frac{1}{2} \times \frac{21}{\eta} \times \frac{hc}{\lambda_0} \times B \tag{4.26}$$

As an example, if λ_0 is 850 nm, η is taken to be unity and B is 34 Mbit/s, then the quantum limit for a 10^{-9} error rate is 80 pW or −71 dB m. In reality, the noise from the photodiode dark current and following amplifier stages will 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, that is

$$<i_n^2>_Q = 2q<I_s> \quad \text{A}^2/\text{Hz} \tag{4.27}$$

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

$$\frac{S}{N} = \frac{<I_s^2>}{2q<I_s>B_{eq}} \tag{4.28}$$

4.5.2 APD noise

In an APD, the *primary current* (that is, the current produced by a unity gain photodiode) is multiplied by the avalanche gain. 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, photon arrival and hence signal current are described by a Poisson process. 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

$$F(M) = M^x \tag{4.29}$$

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 upon the value of k and the type of carrier undergoing multiplication. Detailed analysis (R. J. McIntyre, [1]) shows that $F(M)$ can be approximated by

$$F_e(M) = kM_e + \frac{(1 - k)(2M_e - 1)}{M_e} \tag{4.30}$$

for the electron avalanche, and

$$F_h(M) = \frac{M_h}{k} + \frac{(1 - 1/k)(2M_h - 1)}{M_h} \tag{4.31}$$

for hole avalanche. The equations clearly show the need to fabricate devices out of materials with low values of k. As an example, a Si SPD with an M value of 100 and $k = 0.02$ has $F_e(M)$ equal to ≈ 4, whereas a Ge APD with $M = 20$ and $k = 0.5$ gives $F_e(M)$ equal to ≈ 11. So, Si APDs are far superior to Ge APDs.

We can express the S/N for an APD as

$$\frac{S}{N} = \frac{<I_s^2>M^2}{2q<I_s>M^2F(M)B_{eq} + 2qI_{DB}M^2F(M)B_{eq} + 2qI_{DS}B_{eq} + <i_n^2>_c} \tag{4.32}$$

and, if we ignore the leakage currents and we assume a noiseless receiver, we can write (4,.31) as

$$\frac{S}{N} = \frac{<I_s^2>M^2}{2q<I_s>M^2F(M)B} \tag{4.33}$$

Comparison with the PIN equation (4.28) reveals that, because of the excess noise factor, an APD receiver cannot approach the quantum limit. However, 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

$$\frac{S}{N} = \frac{<I_s^2>M^2}{<i_n^2>_c} \tag{4.34}$$

Thus the *S*/*N* for an APD receiver can be greater than that for a PIN receiver. In most APD receivers, the sensitivity advantage reduces because we cannot ignore the detector noise.

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, exactly the same pre- and post-amplifier combination can be used to detect analogue or digital signals. The difference between the two receivers arises from the way the signals are processed 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 significantly distort the received signal, 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.

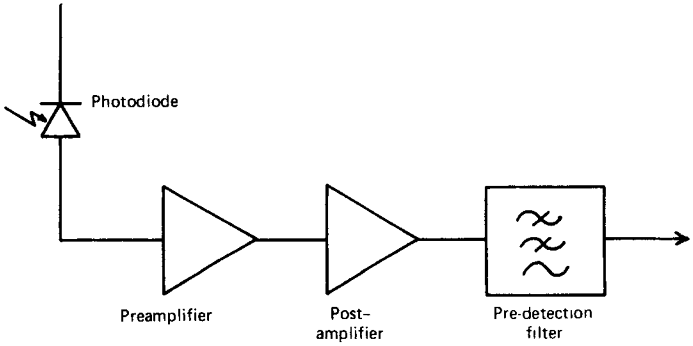


Figure 5.1 The basic structure of an optical receiver

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 voltage gain of the pre- and post-amplifier has been modelled as a 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 $(C_d{}^+C_{in})$, and $H_f(\omega)$ is the pre-detection filter transfer function. Thus we can express $Z_T(\omega)$ as

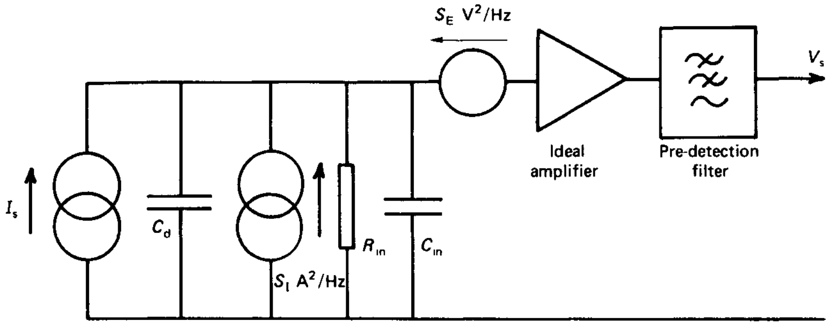


Figure 5.2 a.c. equivalent circuit of an optical receiver

$$Z_T(\omega) = Z_{in}A(\omega)H_f(\omega) \tag{5.3}$$

If we now turn to the noise sources, we can see that the series noise generator produces an m.s. *input* noise current of

$$\frac{S_E}{|Z_{in}|^2} \text{ or } S_E|Y_{in}|^2 \text{ A}^2/\text{Hz}$$

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, $S_{eq}(f)$, will be given by

$$S_{eq}(f) = S_I + S_E|Y_{in}|^2$$

Noting that

$$Y_{in} = \frac{1}{R_{in}} + j\omega C_T$$

where C_T is $C_d + C_{in}$, we can write $S_{eq}(f)$ as

$$S_{eq}(f) = S_I + S_E\left(\frac{1}{R_{in}{}^2} + (2\pi C_T)^2 f^2\right) \tag{5.4}$$

So, the total m.s. output noise voltage, $<n^2>_T$, is

$$\begin{aligned} <n^2>_T &= \int_0^\infty S_{eq}(f)|Z_T(\omega)|^2 df \\ &= \left(S_I + \frac{S_E}{R_{in}{}^2}\right)\int_0^\infty |Z_T(\omega)|^2 df \end{aligned} \tag{5.5}$$

Except for $Z_T(\omega)$, which depends upon 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 $Z_T(\omega)$ 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 upon 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:

$$H_p'(y) = \frac{1}{T} \times H_p(2\pi y/T) \text{ and}$$

$$H_{out}'(y) = \frac{1}{T} \times H_{out}(2\pi y/T)$$

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 upon the relative shapes of the input and output pulses. So, to return to (5.5), we can write

$$<n^2>_T = \left(S_I + \frac{S_E}{R_{in}^2} \right) R_T^2 B I_2 + (2\pi C_T)^2 S_E R_T^2 B^3 I_3 \tag{5.10}$$

where $I_2 = \int_0^\infty [H_T'(y)]^2 \, dy$, and $I_3 = \int_0^\infty [H_T'(y)]^2 y^2 \, dy$

(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

$$<i_n^2>_c = \left(S_I + \frac{S_E}{R_{in}^2} \right) BI_2 + (2\pi C_T)^2 S_E B^3 I_3 \tag{5.11}$$

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 upon 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, the pulses are sampled to determine whether a 1 or a 0 has been received. 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 sinx/x pulse shape will satisfy the ISI requirement. Figure 5.3 shows a sequence of sinx/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 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, a number of difficulties with such an output pulse shape:

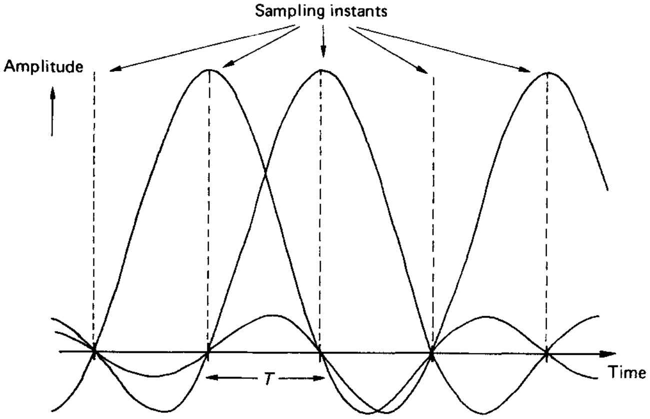


Figure 5.3 A sequence of sinx/x pulses

(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 predetection 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. (Jitter can be minimised 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, the ISI either side of the sampling instant will be reduced, leading to more jitter tolerance. As we will see presently, this advantage can only be obtained at the cost of a reduction in *S*/*N* ratio.

Let us consider a pulse shape, $h_{\text{out}}(t)$, given by (5.12):

$$h_{\text{out}}(t) = \left(\frac{\sin \pi Bt}{\pi Bt}\right) \times \frac{\cos \pi Bt}{1 - (2Bt)^2} \tag{5.12}$$

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, $H_{\text{out}}(f)$, is given by

$$\begin{aligned} H_{\text{out}}(f) &= 1 & |f|<(1+r)\frac{B}{2} \\ &= \tfrac{1}{2}[1 + \cos\{(\pi|f| - \pi f_1)/rB\}] & (1-r)\frac{B}{2} < |f| < (1+r)\frac{B}{2} \\ &= 0 \text{ elsewhere} & (5.13) \end{aligned}$$

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 considered the more

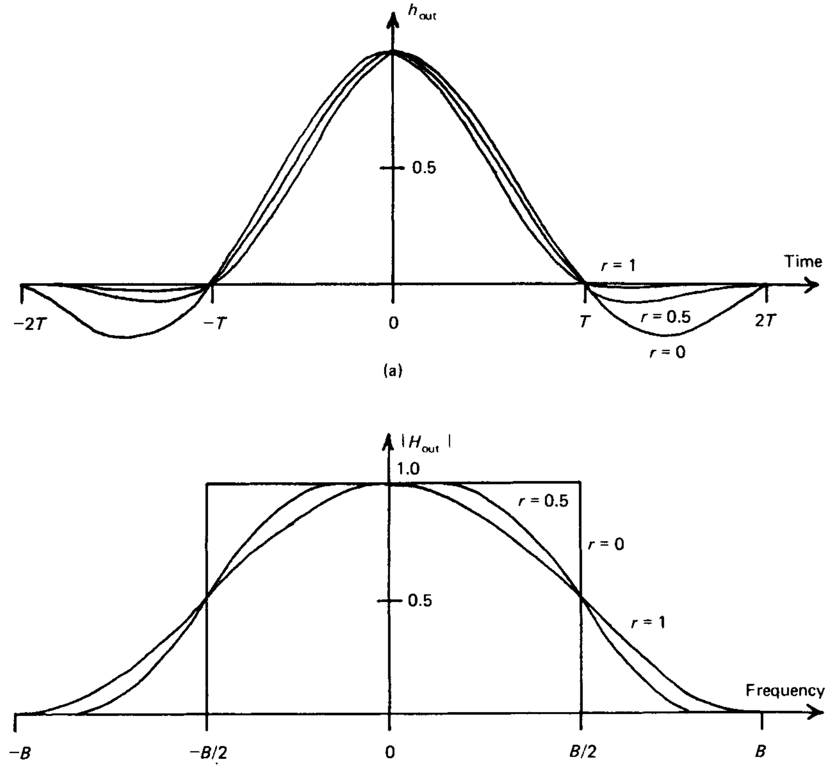


Figure 5.4 (a) A selection of raised cosine spectrum pulses, and (b) corresponding spectra

important parameter, and so the output pulses of the preamplfier– filter combination are designed to have a *full-raised cosine spectrum*, that is, $r = 1$. Hence the normalised spectrum of the output pulses is

$$\begin{aligned} H_{\text{out}}'(y) &= \frac{1}{2T} \times \{1 + \cos\pi y\} && 0 < |y| < 1 \\ &= 0 \text{ elsewhere} && (5.14) \end{aligned}$$

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, S.D. 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

$$H_p'(y) = \frac{1}{2} \times \frac{\sin\alpha\pi y}{\alpha\pi y} \text{ for rectangular pulses,} \tag{5.15}$$

and

$$H_p'(y) = \frac{1}{T} \times \exp\{-(2\pi\beta y)^2/2\} \text{ for Gaussian pulses} \tag{5.16}$$

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

$$\gamma = \int_{-\frac{T}{2}}^{\frac{T}{2}} h_p(t)dt \tag{5.17}$$

These pulses are shown in figure 5.5.

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.

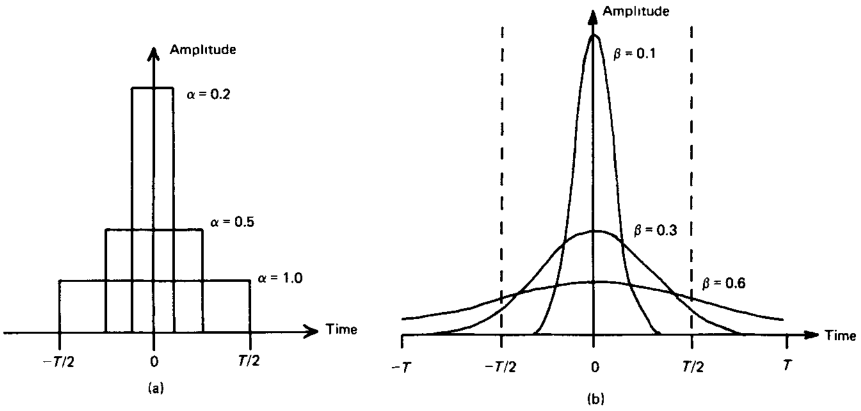


Figure 5.5 (a) Rectangular, and (b) Gaussian shape pulses with various pulse widths

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

We should note that, for Gaussian input pulses, the values of I_2 and I_3 increase rapidly for $\beta > 0.5$. This is to be expected 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 by the input pulse shape. For full-width rectangular input pulses, and full-raised cosine spectrum output 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, such as 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 datastream 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 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, the 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.

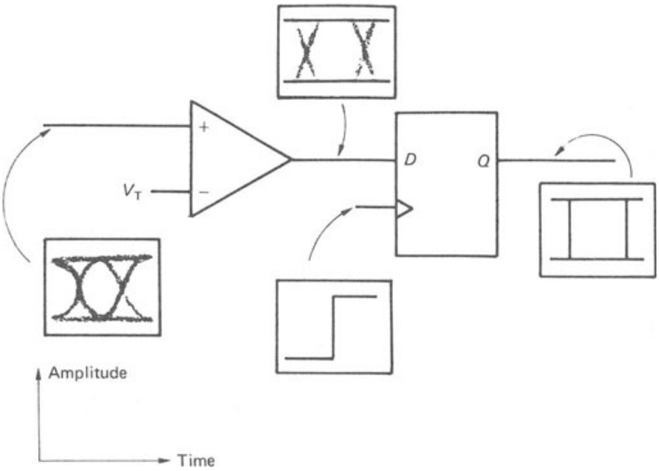


Figure 5.6 Eye diagrams and schematic diagram of a threshold-crossing detector and central decision gate

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, $P_{\mathrm{e}0|1}$. Similarly, the area to the left of V_{T} is the probability that a logic 1 becomes a logic 0, $P_{\mathrm{e}1|0}$. If the noise has a Gaussian distribution,

$$P_{\mathrm{e}0|1} = \frac{1}{\surd(2\pi{\sigma_{\mathrm{off}}}^2)} \int_{-V_{\mathrm{T}}/2}^{\infty} \exp\{-(v - v_{\min})^2/2{\sigma_{\mathrm{off}}}^2\}\,\mathrm{d}v \tag{5.18}$$

and

$$P_{\mathrm{e}1|0} = \frac{1}{\surd(2\pi{\sigma_{\mathrm{on}}}^2)} \int_{-\infty}^{V_{\mathrm{T}}/2} \exp\{-(v_{\max} - v)^2/2{\sigma_{\mathrm{on}}}^2\}\,\mathrm{d}v \tag{5.19}$$

where σ_{off} and σ_{on} are the r.m.s. noise voltages, 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, these terms are included for reasons of brevity.) As the probilities of sending a logic 1 or logic 0 are identical and equal to $\frac{1}{2}$, the total error probability, P_{e}, is

$$P_{\mathrm{e}} = 0.5(P_{\mathrm{e}0|1} + P_{\mathrm{e}1|0}) \tag{5.20}$$

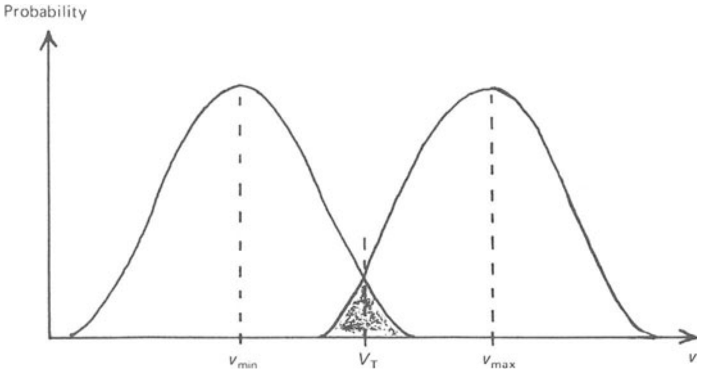


Figure 5.7 Probability density function plot for logic 1 and logic 0 pulses in the presence of additive Gaussian noise

If we neglect signal-dependent shot noise, $\sigma_{\text{off}} = \sigma_{\text{on}} = \sigma$. In such circumstances, the optimum threshold voltage lies mid-way between v_{max} and v_{min}. The reason for this is that if V_T is biased to the left, $P_{e0|1}$ will increase at the expense of $P_{e1|0}$, whereas the opposite is true if V_T is biased slightly to the right. So, with these assumptions, $P_{e0|1} = P_{e1|0}$ and

$$P_e = P_{e0|1}$$

$$= \frac{1}{\sqrt{(2\pi\sigma^2)}V_T}\int_0^{\infty} \exp\{-(v - v_{\text{min}})^2/2\sigma^2\}\,\mathrm{d}v \qquad (5.21)$$

If we change variables by letting

$$x = \frac{v - v_{\text{min}}}{\sigma}\text{, we get}$$

$$P_e = \frac{1}{\sqrt{(2\pi)}} \times \int_Q^{\infty} \exp(-x^2/2)\,\mathrm{d}x \qquad (5.22)$$

where

$$Q = \frac{V_T - v_{\text{min}}}{\sigma} \qquad (5.23)$$

Since V_T lies mid-way between v_{min} and v_{max}:

$$V_T = \frac{v_{\text{min}} + v_{\text{max}}}{2}$$

and

$$Q = \frac{v_{\text{max}} - v_{\text{min}}}{2\sigma} \qquad (5.24)$$

So, provided we know the signal voltage levels and the rms noise voltage at the input to the comparator, we can determine the error probability from (5.22). Although this integral can be evaluated by numerical methods, we can express it in terms of the widely tabulated *complementary error function, erfc*, as

$$P_e = \tfrac{1}{2}\,\text{erfc}(Q/\sqrt{2}) \qquad (5.25)$$

where

$$\text{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^{\infty} \exp(-y^2)\,\mathrm{d}y$$

For $Q > 2$, we can approximate (5.25) to

$$P_e = \frac{1}{Q\sqrt{2\pi}}\left(1 - \frac{0.7}{Q^2}\right)\exp(-Q^2/2) \qquad (5.26)$$

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

$$Q = \frac{I_{\text{max}} - I_{\text{min}}}{2\sqrt{<i_n^2>_c}}$$

or

$$\frac{I_{\text{max}} - I_{\text{min}}}{2} = Q\sqrt{<i_n^2>_c} \qquad (5.27)$$

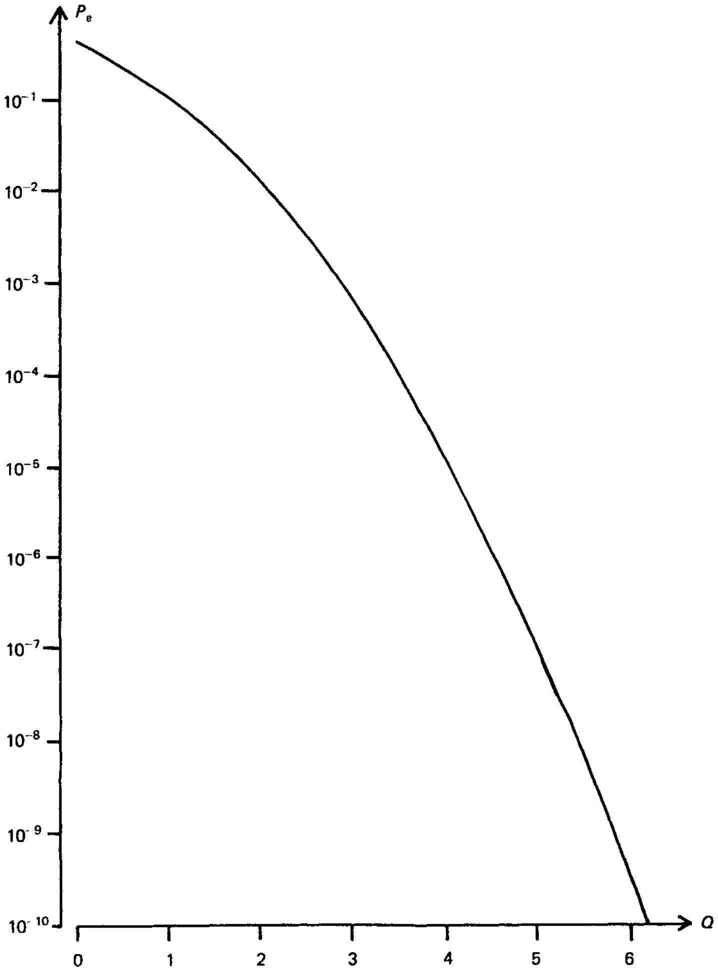


Figure 5.8 Graph of error probability, P_e, against signal-to-noise ratio parameter, Q, for a threshold-crossing detector

where $<i_n^2>_c$ 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 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

$$P = \frac{I_{max}}{2R_0} = \frac{I_{max} - I_{min}}{2R_0} \times \frac{I_{max}}{I_{max} - I_{min}}$$

$$= \frac{Q\surd<i_n^2>_0}{R_0} \times \frac{1}{1 - \epsilon} \qquad (5.28)$$

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 upon the *difference* between the two light levels.

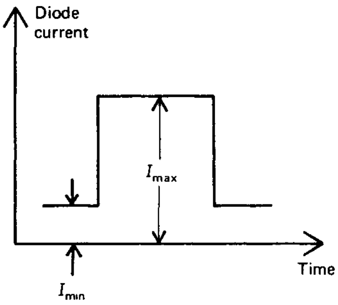


Figure 5.9 Illustrative of a non-zero extinction ratio

Most modern light sources have a very small extinction ratio, that is, 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

$$P = \frac{1}{R_0} Q\surd<i_n^2>_c \qquad (5.29)$$

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.

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, signal noise cannot be ignored. 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 performed 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

$$S_{\mathrm{Id}} = 2qI_{\mathrm{DB}}M^2F(M) + 2qI_{\mathrm{DS}} \tag{5.30}$$

and

$$S_{\mathrm{Is}} = 2q\langle I_{\mathrm{s}}\rangle M^2F(M) \tag{5.31}$$

where $\langle I_{\mathrm{s}}\rangle$ 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

$$\langle i_{\mathrm{n}}^2\rangle_{\mathrm{pd}} = (S_{\mathrm{Id}} + S_{\mathrm{Is}})BI_2 \tag{5.32}$$

Now, $\langle I_{\mathrm{s}}\rangle$ is dependent on the presence of an optical pulse in a particular time-slot. If we take full-width rectangular input pulses, $\langle I_{\mathrm{s}}\rangle$ is I_{max} when a pulse is present, while $\langle I_{\mathrm{s}}\rangle$ 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

$$\langle i_{\mathrm{n}}^2\rangle_1 = 2qI_{\mathrm{max}}M^2F(M)BI_2 + \langle i_{\mathrm{n}}^2\rangle_{\mathrm{T}} \tag{5.33}$$

and

$$\langle i_{\mathrm{n}}^2\rangle_0 = \langle i_{\mathrm{n}}^2\rangle_{\mathrm{T}} \tag{5.34}$$

where $\langle i_{\mathrm{n}}^2\rangle_{\mathrm{T}}$ 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 this is accounted for, and I_{min} is zero, Q will be given by

$$Q = \frac{I_{\mathrm{max}}}{\sqrt{\langle i_{\mathrm{n}}^2\rangle_1} + \sqrt{\langle i_{\mathrm{n}}^2\rangle_0}} \tag{5.35}$$

Thus the mean optical power required is

$$P = \frac{Q}{2MR_0}(\surd\langle i_n^2\rangle_1 + \surd\langle i_n^2\rangle_0) \tag{5.36}$$

If we substitute for $\langle i_n^2\rangle_1$ and $\langle i_n^2\rangle_0$, we get

$$\begin{aligned} P &= \frac{Q}{2MR_0}([2qI_{\max}M^2F(M)BI_2 + \langle i_n^2\rangle_T]^{\frac{1}{2}} + \langle i_n^2\rangle_T^{\frac{1}{2}}) \\ &= \frac{Q}{2MR_0}([4qPR_0M^2F(M)BI_2 + \langle i_n^2\rangle_T]^{\frac{1}{2}} + \langle i_n^2\rangle_T^{\frac{1}{2}}) \end{aligned} \tag{5.37}$$

After some lengthy rearranging, we can express the sensitivity as

$$P = \frac{Q}{R_0}\left(\frac{\surd\langle i_n^2\rangle_T}{M} + qBI_2Q\mathrm{F(M)}\right) \tag{5.38}$$

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

$$M_{\mathrm{opt}} = \frac{1}{\surd k}\left(\frac{\surd\langle i_n^2\rangle_T}{qBI_2Q} + k - 1\right)^{\frac{1}{2}} \tag{5.39}$$

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 the APD gain can be varied by altering the bias voltage, the optimum gain is often determined experimentally when the optical link is installed.

5.2.5 Timing extraction

In the sensitivity analysis just presented, we assumed that central decision detection was carried out. 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 first differentiated and then full-wave-rectified. These two operations

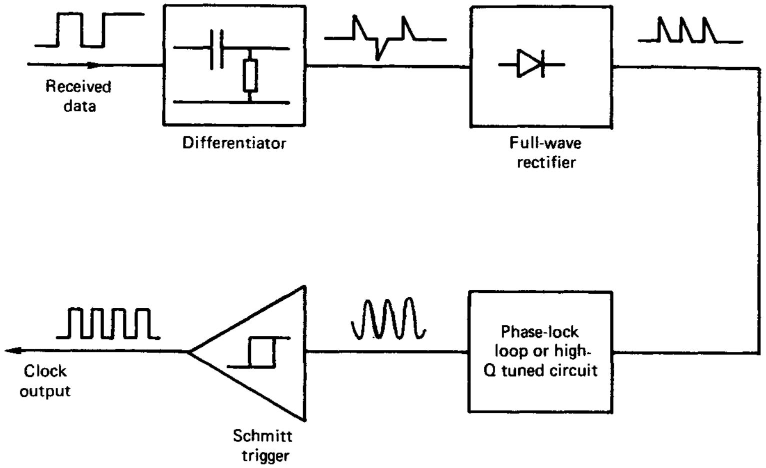


Figure 5.10 Schematic diagram of a timing extraction circuit

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 clock to the flip-flop will be maintained. 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.

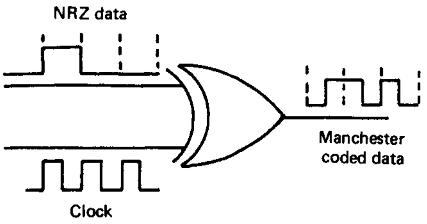


Figure 5.11 Generation of Manchester coded data using an exclusive OR gate

Although bi-phase coding is often used in low bit-rate links and local 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, the maximum number of consecutive like symbols should be limited, and the number of coded ones and zeros should be 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. Base-line wander is alleviated by the use of 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 has 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, base-line wander will be eliminated.

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	
00010	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, such as 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, and $s(t)$ is the modulating signal. (m is the modulation depth as defined in section 3.3.5.) For an APD, this signal produces a photodiode current, $i_s(t)$, given by

$$i_s(t) = R_0Mp(t) \tag{5.41}$$

and so the m.s. signal current, ignoring a constant d.c. term, is

$$<I_s^2> = \tfrac{1}{2}(R_0MmP_r)^2 \tag{5.42}$$

From our discussions about digital receiver noise, we can write the equivalent input m.s. noise current as

$$\begin{aligned} <i_n^2>_T = &\int_0^{B_{eq}} 2qR_0P_rM^2F(M)\,df \\ &+ \int_0^{B_{eq}} 2qI_{DB}M^2F(M)\,df \\ &+ <i_n^2>_c \end{aligned} \tag{5.43}$$

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

$$\begin{aligned} <i_n^2>_c = &\int_0^{B_{eq}} \left(S_I + \frac{S_E}{R_{in}^2}\right) df \\ &+ \int_0^{B_{eq}} S_E \times (\omega C_T)^2\,df \end{aligned}$$

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

$$\frac{S}{N} = \frac{<I_s^2>}{<i_n^2>_T} = \frac{1}{2} \times \frac{(R_0 M m P_r)^2}{2q(I_{DB} + R_0 P_r) M^2 F(M) B_{eq} + <i_n^2>_c} \tag{5.47}$$

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 the optimum gain can be found from,

$$M_{opt}^{2+x} = \frac{<i_n^2>_T}{q(I_{DB} + R_0 P_r) x B_{eq}} \tag{5.48}$$

where we have made use of $F(M) = M^x$, and $<i_n^2>_T$ 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, the APD bias can be altered 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). 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

$$P = \frac{6}{0.5}\sqrt{10^{-15}}\,\mathrm{W}$$
$$= 380 \text{ nW or } -34.21 \text{ dBm}$$

If we replace the PIN detector by the APD, we must first calculate the total signal independent shot noise. Thus

$$<i_n{}^2>_T = <i_n{}^2>_c + 2qI_{DS}I_2B + 2qI_{DB}M^2F(M)I_2B$$
$$= 10^{-15} + (1 + 4 \times 10^4) \times 6.14 \times 10^{-20}$$
$$= 3.45 \times 10^{-15}\ \mathrm{A}^2$$

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

$$P = \frac{6}{0.5}\left(\frac{(3.45 \times 10^{-15})^{\frac{1}{2}}}{100} + 7.4 \times 10^{-11}\right)$$
$$= 8 \text{ nW or } -51.00 \text{ dBm}$$

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 dB m (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

$$B_{eq} = \int_0^\infty \left(\frac{1}{1 + jf/f_0}\right)^2 df$$
$$= \int_0^\infty \frac{1}{1 + (f/f_0)^2}\,df$$
$$= \frac{\pi}{2} \times f_0$$

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

$$B_{eq} = \frac{17\pi}{2} \approx 27 \text{ MHz}$$

By using this result, together with parameters previously quoted, we find the receiver sensitivity from (5.47). Table 5.3 summarises the results.

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 relates 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)

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 which compares the detected PRBS with the transmitted sequence. This instrument counts the number of errors in a certain time interval, from which the probability of an error, P_e, can be calculated.

The mean optical power resulting in the measured error-rate can be found 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 5.9.)

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 is concerned with 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 upon 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 spectral 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 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.

As an example, let us consider the output noise voltage spectral density shown in figure 5.12b. When divided by the transimpedance, figure 5.12a, the equivalent input noise current spectral density takes the form of figure 5.12c. From this figure, the frequency invariant coefficient is $4 \times 10^{-24} A^2/$Hz, while the f^2 variant term is approximately $2.1 \times 10^{-40}\mathrm{A}^2/\mathrm{Hz}^3$. Hence the sensitivity of this receiver can be predicted from

$$P = \frac{Q}{R_0} \surd(4 \times 10^{-24} I_2 B + 2.1 \times 10^{-40} I_3 B^3) \tag{5.50}$$

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, and this can be obtained by connecting a sweep generator to the constant current source. An alternative method is to modulate a light source with the output from a sweep generator, and connect a spectrum analyser to the preamplifier output.

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].

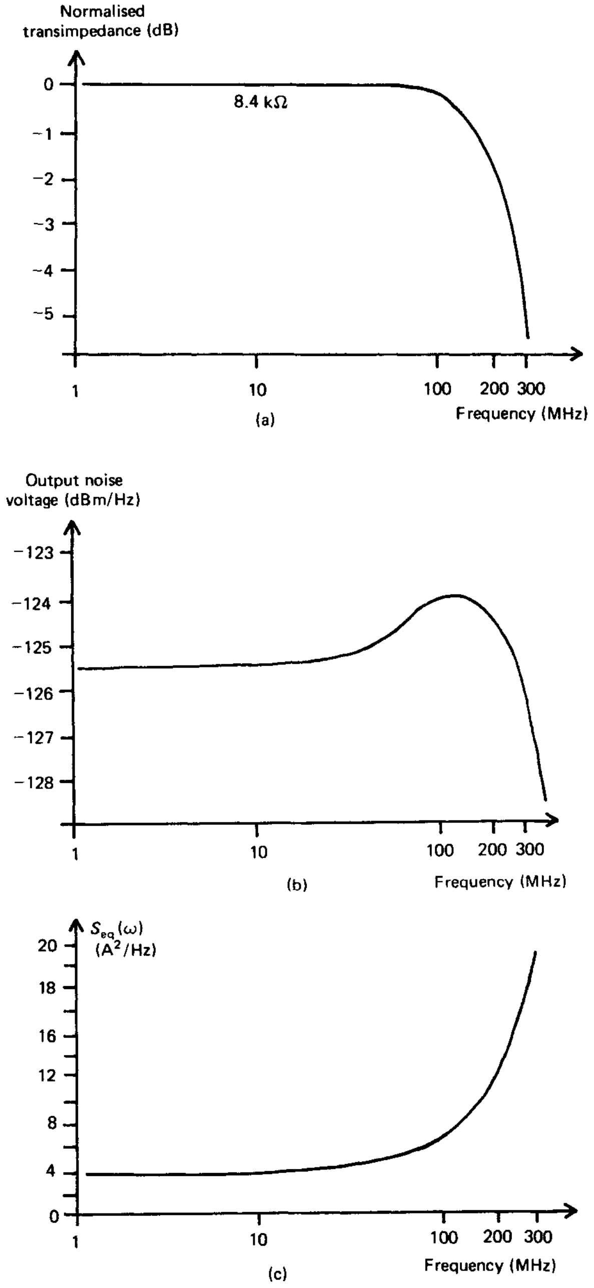


Figure 5.12 (a) Transimpedance relative to mid-band, (b) output noise spectral density, and (c) equivalent input noise current spectral density of a PINBJT transimpedance preamplifier

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 preamplifer — 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 500 MHz). When we examine the integrating front-end receiver, we shall consider a FET input design whereas, when we consider a transimpedance receiver, we shall use a BJT. These are the most common configurations in current use.

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.)

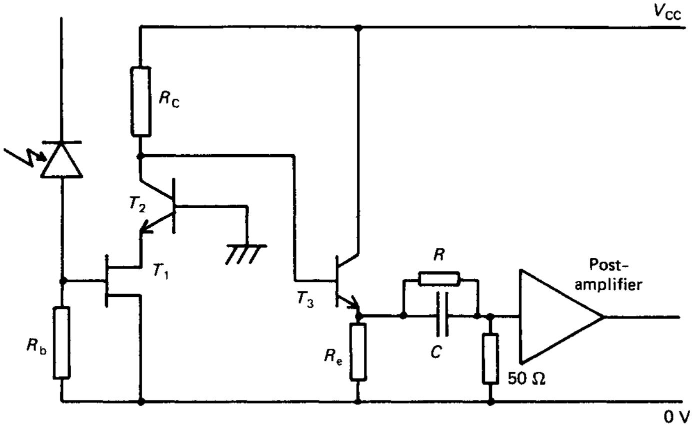


Figure 6.1 A basic PINFET optical receiver with equalisation network and post-amplifier

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 the 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.

6.1.1 Frequency response

The time constants associated 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 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 capcitance, 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 mV and g_{m1} is the FET *transconductance*. Thus C_{in} will be given by

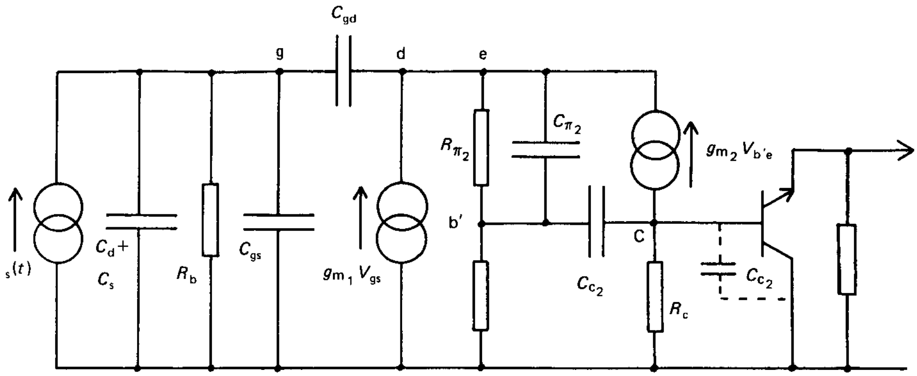


Figure 6.2 a.c. equivalent circuit of a PINFET receiver

$$C_{\mathrm{in}} = C_{\mathrm{d}} + C_{\mathrm{gs}} + C_{\mathrm{s}} + \left(1 + \frac{g_{\mathrm{m1}}V_{\mathrm{T}}}{I_{\mathrm{e2}}}\right) C_{\mathrm{gd}} \tag{6.4}$$

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 $g_{\mathrm{m2}}R_{\mathrm{c}}$, 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 for 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 compensate 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, $H_{\mathrm{eq}}(\omega)$, of the compensation network is

$$H_{\mathrm{eq}}(\omega) = \frac{50}{50 + R} \times \frac{(1 + \mathrm{j}\omega CR)}{1 + \mathrm{j}\omega CR50/(50 + R)} \tag{6.5}$$

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, this can be approximated to $1/C50$. As an example, if the front-end pole is at 30 kHz, and the required receiver bandwidth is 70 MHz, then the values of R and C are 118 kΩ and 45 pF, respectively.

At frequencies below s_1, the compensation network acts as a potential divider. If we take R equal to 118 kΩ, then the attenuation introduced by the compensation network is 4.2×10^{-4} and so, if the receiver input resistance is 1 MΩ and the voltage gain is 20, the transimpedance is 8.4 kΩ. This means that, when referred to the receiver input, the equivalent input noise voltage of the following amplifier is divided by 8.4 kΩ and this may result in a reduced 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 the front-end pole is compensated for, then 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

$$\tau_c = R_c 2C_c \tag{6.6}$$

and so

$$s_2 = \frac{1}{2R_c C_c} \tag{6.7}$$

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

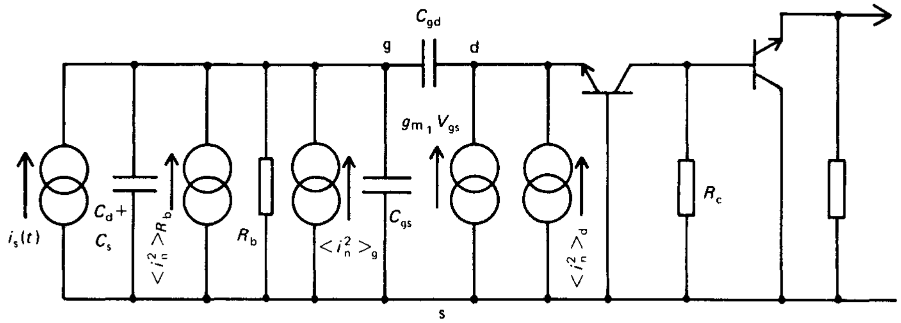


Figure 6.3 Noise equivalent circuit of a PINFET receiver

$$<i_n^2>_c = \frac{4kTI_2B}{R_b} + 2qI_gI_2B + \frac{4kT\Gamma}{g_{m1}}\left\{\frac{I_2B}{R_b^2} + (2\pi C_T)^2I_3B^3\right\} \quad (6.8)$$

where C_T is given by

$$C_T = C_d + C_{gs} + C_{gd} + C_s \quad (6.9)$$

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

$$<i_n^2>_{min} = \frac{4kT\Gamma}{g_{m1}}(2\pi C_T)^2I_3B^3 \quad (6.10)$$

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 previously discussed, then 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, keeps the bias conditions constant). PINFET receivers with this facility have dynamic ranges in excess of 20 dB.

6.1.4 Design example

If a PINFET receiver consists of a hybrid thick-film circuit using unpackaged devices, then we can take a *total* input capacitance of 0.5 pF. Thus, if the value of R_b is 10 MΩ, the highest practical value the input pole will be at 32 kHz.

If we take a cascode bias current of 15 mA, and a FET transconductance of 15 mS, then A_1 is 25×10^{-3}. If an overall voltage gain of 100 is required, A_2 needs to be 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, and so 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.

If we use microwave BJT transistors, then a typical value of C_c will be 0.3 pF, and so the compensated bandwidth will be 663 MHz. 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. If this is so, and we assume this stage has a voltage gain of 17, then the compensated bandwidth reduces to 70 MHz. (The Miller capacitance associated with the common-emitter stage causes this reduction in bandwidth.) As previously mentioned, some preamplifiers use only GaAs FETs. These devices have a very low capacitance, and so the compensated bandwidth can be very high.

If we ignore the photodiode leakage current, then the total equivalent input m.s. noise current will be given by (6.8). Assuming a gate leakage current of 15 nA then, using $g_m = 15$ ms and $C_T = 0.5$ pF, $<i_n{}^2>_c$ is 3.4×10^{-18} A^2 for a data-rate of 140M bit/s. This results in a sensitivity of −49.56 dB m, for R_0 equal to one and an error rate of 1 in 10^9. If we choose to neglect the noise from R_b and I_g, then $<i_n{}^2>_c$ is 2.8×10^{-18} A^2, resulting in a sensitivity of −50.0 dB m. Comparison with the previous result shows that the R_b and I_g noise is not very significant. At high data-rates, the channel conductance noise becomes dominant, and so (6.10) will accurately predict $<i_n{}^2>_c$.

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 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, that is, 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 high 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.

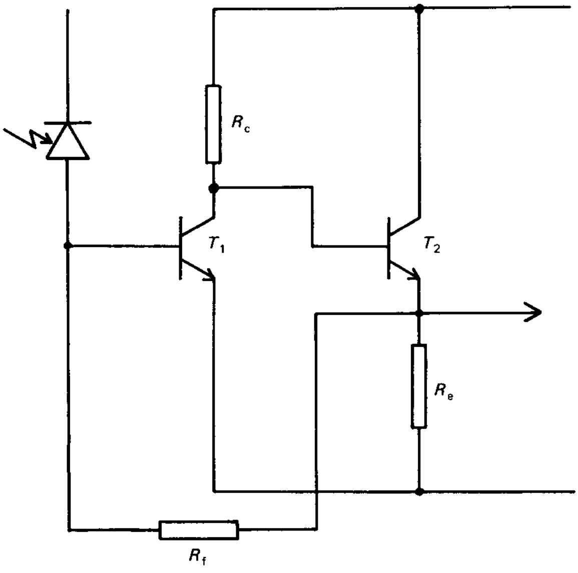


Figure 6.4 A simple common-emitter/common-collector, shunt feedback, transimpedance receiver

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, $Z_0(s)$, and the feedback network transfer function, $Z_f(s)$, by

$$\frac{1}{Z_c(s)} = \frac{1}{Z_0(s)} - \frac{1}{Z_f(s)} \tag{6.11}$$

It can be shown that $Z_0(s)$ is given by

$$Z_0(s) = A_0(s) \times \frac{R_{in}R_f}{R_{in} + R_f}$$

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

$$Z_0(s) = A_0(s)R_f \tag{6.12}$$

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 $Z_0(s)$ is

$$Z_0(s) = \frac{A_0R_f}{1 + s\tau_{in}} \tag{6.15}$$

Also, $Z_f(s)$ is given by

$$Z_f = \frac{R_f}{1 + s\tau_f} \tag{6.16}$$

where τ_f is the feedback circuit time constant, R_fC_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

$$Z_c(s) = \frac{A_0 R_{eff}}{1 + sR_{eff}(C_a + C_s + C_{\pi 1} + (1 - A_o)(C_{c1} + C_f))} \tag{6.18}$$

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, because of 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, that is

$$A_0(s) = \frac{A_1}{(1 + s\tau_{in})} \times \frac{A_2}{(1 + s\tau_c)} \tag{6.21}$$

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

$$<i_n^2>_c = \frac{4kTI_2B}{\mathrm{R_f}} + 2qI_bI_2B + \frac{2qI_{c1}}{g_{m1}^2}\left\{\frac{I_2B}{R_f^2} + (2\pi C_T)^2I_3B^3\right\}$$

$$+ 4kTr_{bb'}\frac{I_2B}{R_f^2} + (2\pi C_1)^2I_3B^3 \qquad (6.22)$$

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

$$I_{c,\,opt} = 2\pi V_TC_T\beta^{\frac{1}{2}}B_{eq}/\surd 3 \qquad (6.25)$$

(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 $I_{c,\,opt}$ by constant iteration.) If we substitute (6.24) back into (6.23), then the minimum noise from a bipolar front-end preamplifier will be

$$<i_n^2>_{c,\,min} = (8\pi kT)(C_T/\beta^{\frac{1}{2}})(I_2I_3)^{\frac{1}{2}}B^2$$

$$+ 4kTr_{bb'}(2\pi C_1)^2I_3B^3 \qquad (6.26)$$

Comparison with the minimum noise from a PINFET receiver, (6.10), shows that, *provided the $r_{bb'}$ 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 $r_{bb'}$ of the transistor will determine the exact cross-over 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 higher 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 $I_{max}R_{eff}A_0$. 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

If we consider a design constructed using state-of-the-art surface mount components on a p.c.b., then the package will tend to dominate the diode capacitance. In addition, special microwave packages need to be used to reduce the transistor parasitic capacitances. With these points in mind, we shall take a diode capacitance of 0.8 pF, and transistor collector–base capacitances of 0.3 pF.. Before we can find the receiver bandwidth, we must initially determine the location of the second pole.

If we assume that the front-end bias current is 2 mA, and we require a voltage gain of 20, then R_c is 250 Ω. So, taking a second-stage voltage gain of unity, and assuming C_{c2} is 0.3 pF, the open-loop pole due to the CE load is at approximately 1.0 GHz. If the f_T of this stage is 4 GHz, and C_f is 0.1 pF, then the open-loop input capacitance is 12 pF. With an R_f value of 4 kΩ, this gives an open-loop pole at 3.3 MHz, and so this is clearly the dominant pole. Thus from (6.20), the closed-loop receiver bandwidth is 70 MHz (just enough bandwidth to detect 140 Mbit/s). If the amplifier had a very high voltage gain, then the bandwidth calculated from (6.22) would be 400 MHz, which 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.

We can find the equivalent input noise current from (6.22). If the current gain is 120 and $r_{bb'}$ is 10 Ω, then $<i_n{}^2>_c$ is 7.6×10^{-16} A^2 — 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 4.8×10^{-16} 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 1Gbit/s, the difference in sensitivity reduces to 4.6 dB, and this 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. Thus if a wide bandwidth is required, then a cascode input, which has a very low input capacitance, is normally used. However, 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 these 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 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} will be

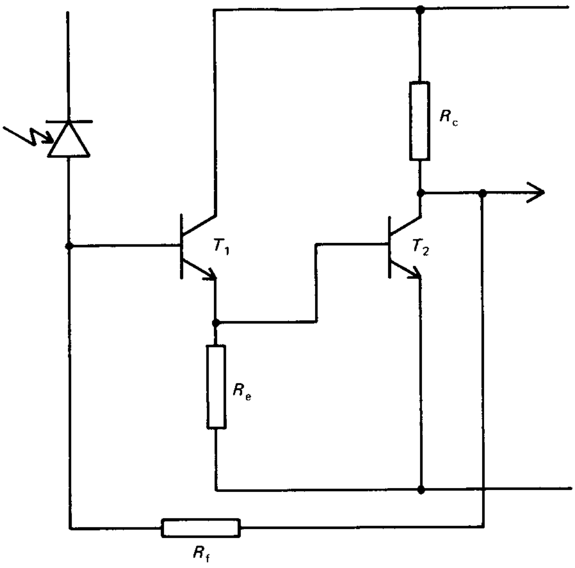


Figure 6.5 A simple common-collector/common-emitter, shunt feedback, transimpedance receiver

$$\tau_{in} = R_f(C_d + C_s + C_f + C_{cl}) \tag{6.28}$$

and τ_c will be given by

$$\tau_c = R_{c\pi2}(C_{\pi2} + (1 - A_2)C_{c2}) \tag{6.29}$$

Here $R_{c\pi2}$ is the resistance in parallel with $C_{\pi2}$, given by

$$R_{c\pi2} = \frac{(R_{01} + r_{bb'2})r_{\pi2}}{R_{01} + r_{bb'2} + r_{\pi2}} \tag{6.30}$$

where R_{01} is the open-loop, output resistance of the front-end. By following a similar analysis to that used previously, $Z_c(s)$ will be 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.

It should be noted that $r_{bb'}$ affects the location of the second stage pole — a high $r_{bb'}$ value results in a large $R_{c\pi 2}$, 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 undershoots. 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

$$<i_n{}^2>_2 = \frac{4kT_{rbb'2}}{A_1{}^2}\left\{\frac{I_2B}{R_f{}^2} + (2\pi C_1)^2 I_3 B^3\right\} \tag{6.33}$$

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

In order to make a fair comparison with the CE design examined previously, we will consider a CC input receiver with the same voltage gain and transistor parameters as before. With these conditions, the transfer function is two-pole in form and so, for a 70 MHz bandwidth, $\mathrm{R_f}$ is about 25 kΩ.

With this value of R_f, the sensitivity is −38.78 dBm (an increase over the CE design of 0.97 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.

At the end 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 a 850 nm wavelength, 10 Mbit/s link, operating over 500 m; and a long-haul, 1.3 μm wavelength, 650 Mbit/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 data-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 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

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 (650 Mbit/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	0.5 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

age. The headroom for the laser link is larger than for the LED link 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 the 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 17, with 16 fusion splices. Thus the maximum length between repeaters is 17 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, $f_{3\text{dB}}$, should be at least half the data-rate, that is

$$f_{3\text{dB}} \geqslant \frac{B}{2} \tag{7.1}$$

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 a general use of (7.2) is minimal.) If we combine (7.1) and (7.2), the minimum rise-time is given by

$$\tau \leqslant \frac{0.7}{B} \tag{7.3}$$

We can find the system rise-time by adding the rise-times of individual components on a mean square basis, that is

$$\tau^2 = \Sigma\tau^2{}_{\mathrm{n}} \tag{7.4}$$

(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, whereas receiver bandwidths are 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

$$\tau_{\mathrm{fibre}} \approx 2.3\sigma \tag{7.5}$$

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 7 ps/nm/km. A laser linewidth of 1 nm yields a dispersion of 70 ps for the 10 km link length. This results in a rise-time of approximately 160 ps which, together with a laser rise-time of 150 ps and a receiver bandwith of 400 MHz, yields a system rise-time of 900 ps. For transmission at 650 Mbit/s, the maximum rise-time should be 1 ns, 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.

7.2 Current systems

In this section we will examine briefly the first optical transatlantic cable, *TAT8*, and, by way of contrast, the computer communications link operating at the Joint European Torus at Culham, UK. Although the technology employed in the 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 responsiblity 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 a 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 results 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 optical links to transmit computer data around the site. Although the environment is electrically noisy, the main reason for the use of optical links is that of electrical isolation. Hence difficulties with varying earth potentials, which results in reduced noise margins, are avoided.

Some of the installed links are configured as local area networks, *LANs*, in which a host computer controls a number of remote stations. We should note however, that all of the links have a maximum distance between terminals of 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 attentuation 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 packaged Si PIN photodiode supplies a signal current to a transimpedance preamplifier. 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, leading to reduced costs. When commissioning the links, *margin testing* is carried out. This involves monitoring link performance when operating with a 3 dB reduction in transmitted power. If a link fails during testing, the cause is investigated and the fault corrected. (As the sources are LEDs, the output power is directly proportional to the drive current and so a 3 dB drop in power can be achieved by halving the drive current.)

As some of the links are LANs, 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 because of 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 with 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 2μm. 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 $\mathrm{ZrF_4}$–$\mathrm{BeF_2}$–$\mathrm{LaF_3}$–$\mathrm{AlF_3}$–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 12μm 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 Advanced lasers

An interesting advance in the area of laser design is the development of fibre lasers (Brierly and France [2]), which are made by doping a fibre core with rare earth elements such as neodymium, Nd^{3+}, and erbium, Er^{3+}. In order to achieve a population inversion, these lasers are optically 'pumped' with short-wavelength light. With such pumping, electrons are excited to a high energy level, from which they decay to an upper lasing level. This gives a population inversion, and so stimulated emission can occur. For operation as a laser, a Fabry-Perot cavity is formed by clamping the fibre between two reflecting mirrors, figure 7.1. In order to pump the laser, the left-hand mirror should be semi-transparent to the pump wavelength, but totally reflect the lasing wavelength, while the right-hand mirror should be semi-transparent to the lasing wavelength, and only slightly reflecting at the pump wavelength. Thus the pumping wavelength appears through the right-hand mirror, where it can be removed by optical filters.

The fibre laser is essentially a wavelength converter. In silica fibre lasers, doping with Nd^{3+} results in absorption at 800 and 900 nm, and emission at 900, 1060 and 1320 nm. Er^{3+}-doped silica fibres absorb at 800, 1000 and 1550 nm and emit at 1550 nm. Because both lasers absorb at 800 nm,

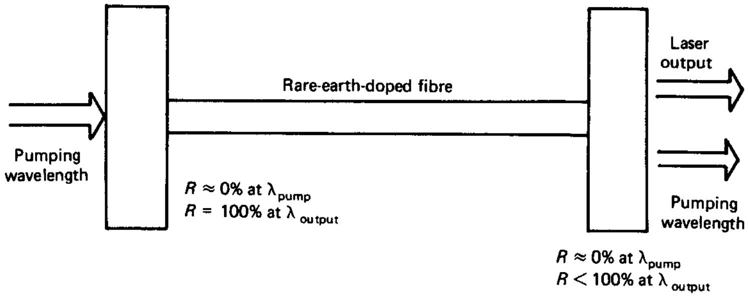

Laser output
Rare-earth-doped fibre
Pumping wavelength
$R \approx 0\%$ at λ_{pump}
$R = 100\%$ at λ_{output}
Pumping wavelength
$R \approx 0\%$ at λ_{pump}
$R < 100\%$ at λ_{output}

Figure 7.1 Schematic of a rare-earth-doped fibre laser

low-cost GaAlAs SLDs can be used to pump the fibre laser. As well as being a source of light, rare-earth doped lasers can also be used as amplifiers. If an Er^{3+}-doped silica fibre is operated at threshold by pumping with 800 nm light, then any incident 1550 nm light will cause the material to lase at the same wavelength, so effectively amplifying the received signal.

The highest wavelength transmission window exhibits the lowest loss, and so most semiconductor laser development has been concerned with 1.55 μm wavelength devices. As we saw in chapter 3, a laser amplifies light by stimulated emission. So, if we arrange for modulated light to enter one side of a laser, we can amplify the signal. There are two basic laser amplifier structures; *high-* and *low-reflectivity* resonators (O'Mahony *et al.* [3]). The high-reflectivity resonator is similar to the Fabry–Perot cavity we discussed in chapter 3. To ensure that light enters the laser, one of the reflecting facets should only reflect light from the cavity. As we have already seen, the cavity will only amplify certain wavelengths, and so the bandwidth of high-reflectivity laser amplifiers can be low (the linewidth of the laser). Because of this, such devices exhibit low noise, and may find applications as amplifiers prior to detection by an optical receiver.

In contrast, low-reflectivity laser amplifiers (also known as *travelling wave amplifiers*) have anti-reflection coatings applied to each facet. The coatings serve to remove the wavelength selectivity of the laser, and so such a device will amplify a wide range of wavelengths (typically 40 nm). With such a wide bandwidth, some spontaneous emission of light can occur, and so these devices are noisier than Fabry–Perot amplifiers. As we shall see later, wideband laser amplifiers are likely to be particularly useful in optical broadcast networks. Both types of laser amplifier have been developed for use in the 1.55 μm window, with typical gains of 20 dB, and output powers of 1 mW (0 dBm).

In chapter 3 we examined briefly single-mode lasers based on ridge waveguide structures. Unfortunately, because of manufacturing tolerances and temperature effects, the emission frequency of these lasers cannot be accurately controlled. In single-mode *long external cavity*, *LEC*, lasers, a diffraction grating provides one of the laser facets (Mellis *et al.* [4]). 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 7.3.4, LEC lasers are required for use in coherent detection receivers.

7.3.3. Integrated optics

When we examined optical couplers in chapter 2, we briefly considered a single-mode, waveguide coupler fabricated on a lithium niobate, *LiNbO$_3$*, substrate. In this material, the refractive index varies according to the strength of an externally applied electric field, the so-called *electro-optic* effect. By exploiting this property, a vast array of *integrated optics* components can be produced: phase and intensity modulators, frequency shifters, polarity controllers, optical switches, and even A–D and D–A converters. As the number of components is so great, we will only consider the phase modulator, and the *Mach–Zender* modulator (Nayar and Booth [5]).

The most basic integrated optic component is the phase modulator shown in figure 7.2a. 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. So, the optical signal experiences a change in phase, $\delta\phi$, given by

$$\delta\phi = \frac{2\pi n}{\lambda_0} \times \delta n \times L \tag{7.6}$$

where L is the length of the guide. The change in refractive index, δn, is related to the electric field by

$$\delta n = n^3 \times \frac{R}{2} \times E \tag{7.7}$$

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}. Phase modulators are very fast devices, with a typical maximum modulation speed in excess of 7 GHz. If these devices are used in a Mach–Zender interferometer, a very fast on–off modulator can be produced.

In the Mach–Zender interferometer shown in figure 7.2b, 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

$$P_{\text{out}} = P_{\text{in}}\cos^2 \frac{\Delta\phi}{2} \tag{7.8}$$

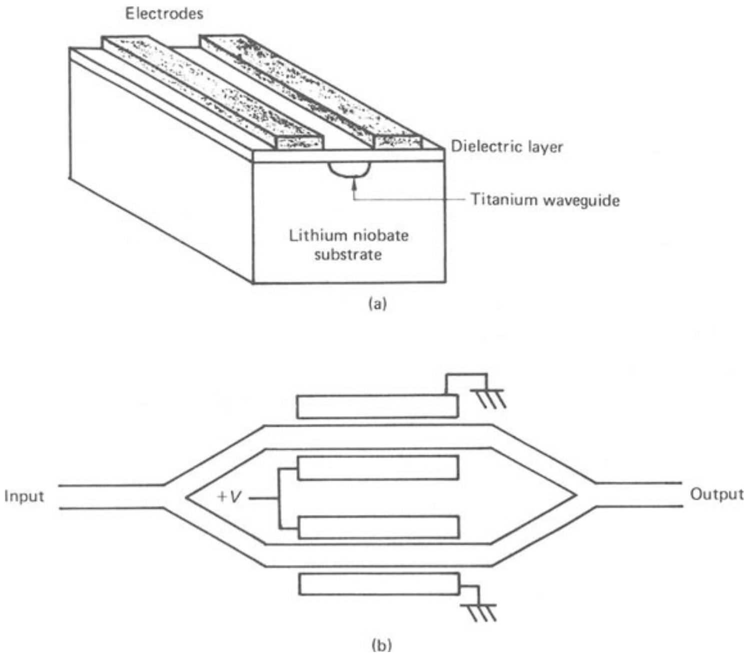


Figure 7.2 (a) A simple phase modulator, and (b) a Mach–Zender modulator

where $\Delta\phi$ is the phase difference between the two branches. Mach–Zender modulators are of great use when a laser has to be modulated at high speed. When the drive current to a laser is pulsed on and off, the wavelength of emission varies slightly. 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, the laser can be operated continuously, and a Mach–Zender modulator used to modulate the output.

Laser diodes cannot be fabricated out of $LiNbO_3$, and so light must be coupled into the device waveguides by some means. However, GaAs and InP are also electro-optic materials, and so lasers and photodiodes can be fabricated into electro-optic devices. The result is an opto-electronic integrated circuit, *OEIC*. By integrating optical processing circuits with lasers and detectors, very fast signal processing becomes possible, leading to applications such as optical radar and recognition systems.

7.3.4 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.* [6]). Heterodyning is used in most modern radio receivers, and so we will only examine this technique.

Figure 7.3 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 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

$$E_r = e_r\cos(\omega_r t + \phi) \tag{7.9}$$

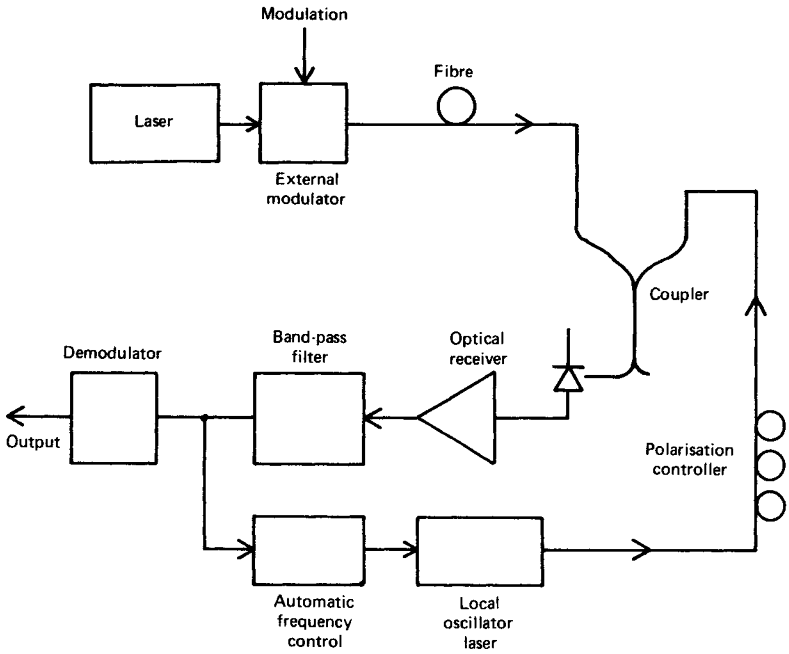


Figure 7.3 Schematic of an optical heterodyne detection system

and the local oscillator field, E_L, is

$$E_L = e_L \cos\omega_L t \qquad (7.10)$$

So, we can write the resultant field, E_i, as

$$\begin{aligned} E_i &= E_r + E_L \\ &= e_r\cos(\omega_r t + \phi) + e_L\cos\omega_L \mathrm{t} \qquad (7.11) \end{aligned}$$

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

$$\begin{aligned} E_i^2 &= e_r^2\cos^2(\omega_r t + \phi) + e_L^2\cos^2\omega_L t \\ &\quad + 2e_r e_L\cos(\omega_r t + \phi)\cos\omega_L t \end{aligned}$$

or, in terms of optical power

$$\begin{aligned} P_i &= P_r\cos^2(\omega_r t + \phi) + P_L\cos^2\omega_L t \\ &\quad + 2\sqrt{(P_r P_L)}\cos(\omega_r t + \phi)\cos\omega_L t \qquad (7.12) \end{aligned}$$

Expansion of the cosine terms in (7.12) 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

$$\omega_{if} = \omega_L - \omega_r \qquad (7.13)$$

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 (4.22)). 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

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	$\frac{1}{2}\mathrm{erfc}\left(\frac{P_L\lambda_0}{4hcB_{eq}}\right)^{\frac{1}{2}}$	72
	FSK	$\frac{1}{2}\mathrm{erfc}\left(\frac{P_L\lambda_0}{2hcB_{eq}}\right)^{\frac{1}{2}}$	36
	PSK	$\frac{1}{2}\mathrm{erfc}\left(\frac{P_L\lambda_0}{hcB_{eq}}\right)^{\frac{1}{2}}$	18
Homodyne	ASK	$\frac{1}{2}\mathrm{erfc}\left(\frac{P_L\lambda_0}{2hcB_{eq}}\right)^{\frac{1}{2}}$	36
	PSK	$\frac{1}{2}\mathrm{erfc}\left(\frac{P_L\lambda_0}{hcB_{eq}}\right)^{\frac{1}{2}}$	9
Direct detection	ASK	See chapter 4	21

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, tuneable lasers that have a narrow line-width. 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 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 [7]). 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.* [8]). 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 common-place.

7.3.5 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 about 2 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 common-place, 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 [9]). At the optical receiver, the use of coherent detection, with variable frequency lasers, means that the required channel can be selected from those available. At any intermediate repeaters, wideband travelling wave amplifiers, mentioned in section 7.3.2, can be used to boost the power of each indidual signal.

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 500 channels within the 40 nm bandwidth of a wideband laser 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. Such a scheme would allow the consumer to access a wide variety of services, and receive 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 specifically 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. 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 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, NJ.
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.

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, Part 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 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, pp. 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. Electrical 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 telecommunicatons', *British Telecom Technology Journal,* **5**, 28–44.
2. Brierley, M. C. and France, P. W. (1987). 'Neodynium doped fluorozirconate fibre laser', *Electronics Letters*, **23**, 815–817.
3. O'Mahony, M. *et al.* (1986). 'Wideband 1.5 μm optical receiver using travelling wave laser amplifier', *Electronics Letters,* **22,** 1238–1240.
4. Mellis, J. *et al.* (1988). 'Miniature packaged external cavity semiconductor laser with 50 GHz continuous electrical tuning range', *Electronics Letters* **24**, 988–989.
5. Nayar, B. K. and Booth, R. C. (1986). 'An introduction to integrated optics', *British Telecom Technology Journal,* **4**, 5–15.
6. Hodgkinson, T.G. *et al.* (1985). 'Coherent optical transmission systems', *British Telecom Technology Journal,* **3**, 5–18.
7. Walker, G. R. and Walker, N. G. (1988). 'A rugged all-fibre endless polarisation controller', *Electronics Letters*, **24**, 1353–1354.
8. Creaner *et al.* (1988). 'Field demonstration of 565 Mbit/s DPSK coherent transmission system over 176 km of installed fibre', *Electronics Letters*, **24**, 1354–56.
9. Brain, M. (1989). 'Coherent optical networks', *British Telecom Technology Journal*, **7**, 50–57.

Useful reference books

Kressel, H. (Ed.) (1980). *Semiconductor Devices for Optical Communications*, Vol. 39, Topics in Applied Physics, Springer-Verlag, New York.

Basch, E. E. (Ed.) (1987). *Optical-fiber Transmission*, Sams, New York.

Barnoski, M. K. (Ed.) (1981). *Fundamentals of Optical Fiber Communications*, Academic Press, New York.

Senior, J. (1985). *Optical Fiber Communications: Principles and Practices,* Prentice-Hall, Englewood Cliffs, NJ.

Index